veradigm® WHITEPAPER

From Fragments to Clarity:

EMPOWERING RESEARCH WITH COMPLETE PATIENT DATA

Natural language processing can help tap into the missing element by transforming difficult-to-understand unstructured data into valuable insights.

Consider the following scenario: Pharmaceutical researchers are working with data that has been culled from electronic health records (EHRs). Structured coded data indicates that several patients suffer from psoriasis. But the researchers still wonder: Did these patients have little lesions on their hands—or did they have cracked, dark skin patches all over their body? Did they experience accompanying joint pain? What was the severity of their disease? Were these patients being treated with self-care, topical therapy, oral medications or injected medications?

Why all the questions? Coded data in electronic health records (EHRs) could indicate that patients have psoriasis, but this structured information typically does not really shed any light on the many other important details. Indeed, structured coded information can provide some insight, but it can fall short when pharma researchers are trying to get the complete picture, truly understand patient health and ultimately develop more effective drugs.

"It is important to understand how severe the disease is on a scale of 1 to 10. And, that information can only be captured in unstructured data or clinical notes," said Nital Patel, Vice President, Real World Data, Veradigm. Many other insights are also missing when working with structured data only. For example, while lab reports provide test results, accurate and complete lab values are often hard to come by in structured data.

"In many disease areas, clinicians use subjective 1–10 severity scales to document how severely a patient is affected. These scales may refer to pain, breathlessness, fatigue or functional impairment—but they are rarely captured in structured fields. Instead, they are embedded in free-text clinical notes. These insights are crucial for understanding disease progression, treatment response and patient-reported outcomes, but they're only accessible using natural language processing (NLP)," Patel added.

THE PROBLEM IS THAT **80%** OF EHR BUCKET FALLS INTO THE UNSTRUCTURED BUCKET—DIFFICULT TO BOTH ACCESS AND LEVERAGE.

WHY RESEARCHERS NEED TO KNOW IT ALL

The need for this comprehensive information, however, has grown recently as medical care has advanced.

Several years ago, diseases and treatments were seen in a linear manner. As the research advanced, the complexity arose in treating patients. Most treatments now have multi-indications and the need to understand these diseases have deepened, according to Patel.

"With the rise of personalized medicine and genomic profiling, much of this critical information isn't reflected at the ICD-10 code level in claims data," Patel said.

Indeed, structured, coded data can only do so much. For example, if healthcare professionals are examining simple cardiovascular diseases the structured information might suffice. "But when you work with diseases like myocarditis, when a patient's left ventricular ejection fraction (LVEF) falls outside the normal range and needs to be classified using NYHA Class I through IV, the critical clinical details aren't captured in structured data. This information exists only in unstructured clinical notes," Patel said. "As a result, researchers who are seeking to improve drug development and related commercial strategies need to access and work with a complete picture of the patient journey that includes real world evidence (RWE), when trying to advance clinical treatments."

THE STATE OF EHRS

A look at what is typically included in EHRs sheds light on some of the struggles that researchers could be dealing with as they strive to obtain this comprehensive picture of patients and their care.

EHRs typically contain structured information such as demographics, diagnostic codes, vital signs, lab results and prescription data as well as unstructured information such as physician notes, pathology reports, discharge summaries, patient narratives and other information. The problem: 80% of EHR data falls into the unstructured bucket [1]—and that information is extremely difficult to access and ultimately leverage.

"Accessing all this information, curating it, making it into uniform standard data structures has been the major challenge," Patel noted.

Unfortunately, there is no rhyme or reason for the unstructured data. "The data is highly unstructured and disorganized. A patient speaking continuously during a visit while the nurse practitioner tries to document everything in real time. The result is often a mix of relevant and irrelevant details, recorded without any consistent structure or clinical prioritization," Patel said.

In addition, the unstructured information could be taken out of context. "The patient could also mention anything about their family—something happening to the child, something happening to their parents. And this conversation could be irrelevant to the patient's health. The patient might just mention the word 'diabetes' and it gets picked up as the patient having the disease, but they are simply referring to someone else," said Patel.

Not surprisingly, making sense of this unstructured data is not an easy task. To tap into and analyze all the data buried in clinical notes traditionally has required a herculean human effort. Indeed, simply understanding symptoms and disease treatment could take a considerable amount of time, according to Patel.

THE PROMISE OF NATURAL LANGUAGE PROCESSING

NLP, however, can help. NLP makes it possible to understand human language and includes operations such as speech recognition, translation, text analysis and other language-related functions. Effective NLP technologies take both statistical and semantic approaches. Statistical NLP is based on machine learning and contributes to the increased accuracy of recognition. Semantic NLP analyzes unstructured datasets, such as physician notes and pathology reports. Therefore, NLP can uncover disease markers, patient-reported symptoms and treatment rationales from EHRs.

Integrating an NLP tool into an EHR dataset adds even more value. By accessing valuable insights from unstructured EHR data and combining this knowledge with insights culled from structured data, researchers can benefit from regulatory-grade, RWE that's purpose-built for drug development.

NLP MAKES IT POSSIBLE TO **UNDERSTAND HUMAN LANGUAGE** AND CAN UNCOVER DISEASE MARKERS, PATIENT-REPORTED SYMPTOMS AND TREATMENT RATIONALES FROM EHRS.

It is important to work with a technology vendor that owns the data or has explicit permission to access the data to effectively apply an NLP tool that will create valuable insights, though. Some companies, such as Veradigm, started as EHR companies, making them data originators. As a result, these companies typically own all the data, including the clinical notes, and can freely perform the NLP, whereas other data providers might not be in a position to do so.

"It is really important to work with a data provider who has the appropriate data rights and the technical capability to be able to leverage NLP to mine the data," Patel said.

This access to NLP-enhanced data makes it possible for researchers to cull information from millions of unique patient records and gain insights on a wide range of geographically and demographically diverse patients. Veradigm datasets, for example, encompass more than five billion distinct patient notes* sourced from multiple EHR systems and a diverse range of specialties, collected over a rolling five-year period. As such, this standardized NLP-enriched EHR data can support advanced research efforts.

REDUCING THE BIAS USING MULTI-SOURCED EHR DATA REGISTRIES

Veradigm registries contain de-identified information about patients' health and care for a specific condition or disease. Veradigm's cardiometabolic registries, for instance, offer information from several cohort registries including Heart Failure, Atrial Fibrillation, Atherosclerotic Cardiovascular Disease (ASCVD), Hypertension, Type 1 Diabetes, Type 2 Diabetes and Chronic Kidney Disease. With access to this NLP-enriched data, researchers can focus on providing effective treatments for cardiovascular disease and metabolic disorders, which are a leading cause of death.

Data registries harmonize data from multiple different EHR systems into a research ready data model, leveraging both structured and unstructured data to provide precise insights into disease patterns, treatment outcomes and patient care. As such, the data is much broader than the data that emanates from just one EHR vendor.

"So, when researchers are using this data, it is much less biased," said Patel. "The data is being pulled from different EHRs with different patient demographics with a good representation from rural and urban settings. Therefore, the researchers can work with much richer datasetIn addition, it is important to seamlessly integrate EHR data with other third-party datasets, offering researchers the flexibility to link and expand their analyses."

CREATING A COMPLETE PICTURE WITH NLP

Gaining access to NLP-enriched, integrated data can offer valuable insights into therapy decisions, disease progression and patient outcomes. In addition, researchers can trust this information as accuracy increases when NLP-enriched data is pulled from various sources.

Indeed, RWE can support strategies for clinical research that include clinical trials, regulatory submissions and safety studies. In addition, researchers can leverage the data to support Health Economics and Outcomes Research (HEOR) and commercialization strategies.

"It is really important to work with a data provider who has the appropriate data rights and the technical capability to be able to leverage NLP to mine the data."

— NITAL PATEL |
VICE PRESIDENT, REAL WORLD DATA, VERADIGM

REAL SUPPORT FOR REAL PROFESSIONALS

MORE SPECIFICALLY, COMPREHENSIVE NLP-ENRICHED RWE CULLED FROM EHRS, AND OTHER SOURCES CAN SPECIFICALLY SUPPORT THE FOLLOWING:

Clinical development teams—insights from the data can help these teams answer specific questions. For example, clinical development teams could be looking to answer questions about market assessment and disease progress during the preclinical research stage. The researchers might also tap into the data to determine if there are any unmet needs that current treatments are not addressing.

HEOR professionals—comprehensive RWE empowers HEOR staff members to focus on cost effectiveness. For example, they could analyze the economic burden presented by certain diseases to patients. The analysis could even empower them to ascertain if the burden is greater at the beginning of the year when patients will have not met their deductible versus later in the year when out-of-pocket costs are diminished.

Regulatory affairs teams—these professionals can rely on comprehensive RWE to formulate and then submit credible clinical trial proposals to agencies such as the Food and Drug Administration.

Medical affairs professionals—with access to comprehensive data, medical affairs teams can track real-world outcomes, not just clinical trial results. For instance, RWE can uncover if there are side effects associated with continued use of a product, for instance. With this information, medial affairs teams can ensure that healthcare professionals, payors, policymakers and others make informed decisions that ensure the best use of products to benefit patients.

Market access teams—these commercial professionals can leverage comprehensive data to identify trends, evaluate competition and optimize brand strategies. They can also use insights to reach healthcare providers effectively, enhancing brand visibility and adoption.

VERADIGM'S COLLABORATIVE APPROACH IS RESULTING IN ACCURACY RATES OF **93%** AND HIGHER.

A MORE INFORMED FUTURE

In the final analysis, with access to comprehensive RWE, many drug development professionals can more effectively do their jobs. And, this will empower the industry, as a whole, to move forward.

"The future of healthcare research relies on access to real-time, comprehensive data and a significant portion of this data is unstructured and that has been a major stumbling block for quite some time. However, NLP is now providing a way for researchers to tap into all the valuable insights that can be culled from the various notes that were formerly so elusive," Patel concluded. "Fortunately, Veradigm is now taking a collaborative approach, which is resulting in accuracy rates of 93% and higher.[2] Ultimately, Veradigm has invested in developing solutions that bring together disparate structured and unstructured data to provide researchers a more comprehensive understanding of the patient journey."

REFERENCES

1. Kong, Hyoun-Joong, Managing Unstructured Bid Data in Health System. Healthcare Informatics Research. https://pmc.ncbi.nlm.nih.gov/articles/PMC6372467/
2. Veradigm Data on File, 2025.