Verana Health

How AI is Unlocking Unstructured Healthcare Data to Transform Prostate Cancer Treatment

Data quality and completeness have become hot button issues as the world has grown increasingly reliant on algorithms and artificial intelligence (AI) to assess risks, spot anomalies, and forecast events. The concern, of course, is that models based on inaccurate, incomplete or incompatible datasets will fail to represent the entire population and produce skewed results. While we've heard a lot about AI perpetuating inaccuracies, less has been written about how AI is being used to extract a more complete picture from underlying datasets. Yet, that's exactly what's happening today in the field of prostate cancer research using real-world evidence (RWE) and AI-powered analytics to receive a more accurate and comprehensive view of the patient population.

Verana Health is at the center of this effort through the work it's performing to tap into unstructured data in the clinical notes of urologists' electronic health records (EHRs). This valuable information can help provide a better understanding and more accurately identify the key signals of prostate cancer disease progression.

Despite having one of the highest five-year survival rates of any form of cancer, some 10-20% of prostate cancer patients develop more severe, castration-resistant forms of the disease and roughly 6% progress to metastatic prostate cancer. Historically, efforts to identify key signals and patterns of treatment consistent with those cases of increased severity, using traditional approaches to RWE, have understated the total patient population. That's because the coding taxonomy is inadequate to follow disease progression, such as identifying metastasis, castrate resistance, or rises in prostate-specific antigen (PSA) levels.

In the real-world, prostate cancer is more often staged by a urologist at the point of diagnosis and any further progression is documented not as a new TNM stage, but as information using other descriptors in clinical notes. That means in many cases when the cancer metastasizes, the clinical note may not reference "TNM," but instead may state: "growing sites of metastasis on scan" or "positive bone scan." These alternative mentions of disease progression are the best clues available for signaling metastatic risk, but until the use of AI and large language models, they were nearly impossible to identify at scale.

In the pages that follow, we will outline the progress Verana Health has made in applying AI-driven large language models to RWE for prostate cancer and codifying signals for disease progression on a nationwide scale.

Applying Large Language Models to Urology Clinical Notes

Two key developments have made it possible to unlock critical insights from unstructured clinical notes. The first is the evolution of AI-powered large language models and cloud computing capabilities that can ingest massive volumes of data, identify patterns in that data, and generate predictive outcomes based on pattern recognition. The second is the ability to access granular, patient-specific data that captures many aspects of the patient journey—including the clinical notes keyed into the free text section of EHRs.

Verana Health has harnessed both of these to develop the most advanced analytics of unstructured urological patient data available today. As the exclusive data partner of the American Urological Association (AUA) Quality (AQUA) Registry, Verana Health is able to tap into a 10-year longitudinal database that includes outpatient data for more than 11.9 million de-identified patients from more than 2,200 active clinicians. The AQUA Registry is the largest urology patient registry of its kind, and the ability to capture this data in near-real-time allows Verana Health to understand all aspects of individual patient experiences throughout their healthcare journeys.

In order to extract meaningful insights from this data, Verana Health has developed a unique approach to data curation, whereby we apply AI-powered large language and machine learning (ML) models to analyze patterns of language in unstructured clinical notes that signal key milestones and clinical insights that occur during the patient journey. Most importantly, our team of clinical experts, which includes experienced urologists with deep expertise in data-driven research, is continually training and establishing rules for how this unstructured data is cataloged and categorized to make it useful in the real world.

Verana Health is unique among healthcare data and analytics providers in its ability to analyze this depth and breadth of RWE at scale. While some companies have set out to manually parse clinical notes for insights, and others have tried to automate the entire process, Verana Health is the only company of its kind to model patterns of language in this manner using the robust amounts of EHR data captured in the AQUA Registry.

Modeling Prostate Cancer Progression

To better understand how Verana Health's approach to unstructured data curation works in the real world, let's take a closer look at prostate cancer. As previously noted, when it comes to structured, standardized disease staging, the formal industry standard codes used to identify critical markers of disease progression are often not captured in structured EHR or claims data. Typically, in the free text notes in an EHR, a patient is diagnosed and staged once, and any subsequent disease progression is captured in a variety of ways.

We know this because we've tracked it. In fact, when we examined our total universe of over 364,000 patients with prostate cancer and conducted a basic screen for patients who experienced metastasis, using the documentation of "M1," we discovered 6,000 total patients. However, when we expanded that analysis to include other phrases that also indicated metastasis in clinical notes, we discovered 29,000 patients, a five-fold increase.

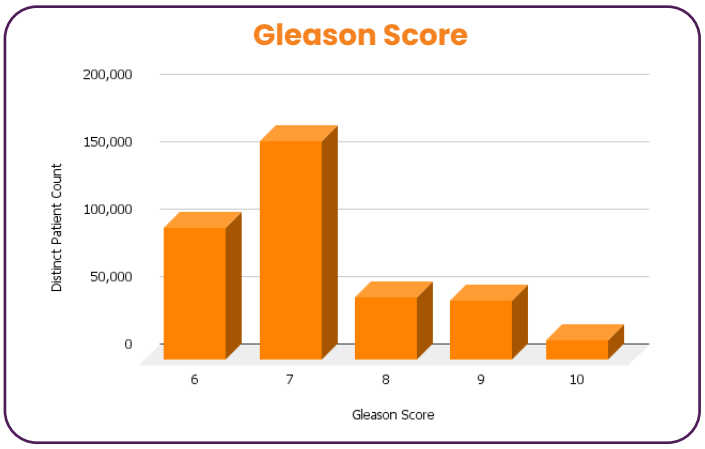
The key to finding all of those missing patients was fine-tuning AI-powered large language models to flag keywords and patterns of language consistent with certain clinical cues. For example, we have been able to develop ML models that identify phrases in clinical notes, such as "growing sites of metastasis on scan," or "positive bone scan," which provide critical clues that signal things like metastatic risk or development of castration-resistant forms of the disease.

Metastasis (at Any Point in Time)	
Metastasis Stage	**Estimated Unique Patients in Qdata Prostate Cancer**
M0* (from TNM staging)*	53,000
Metastatic (from TNM staging, M1)*	6,000
Metastatic (from clinical documentation)*	29,000
Metastatic Total	30,000

When we expanded the analysis for patients who experienced metastasis from "M1" to alternative phrases in clinical notes, we discovered

5x more instances of metastasis.

*These categories are not mutually exclusive. Data as of November 2023.



Data as of November 2023.

Other key variables involved in the diagnosis and staging of prostate cancer are Gleason scores, which are based on biopsy samples and describe how aggressive cancer cells are, and PSA levels, which is a lab result used to track disease progression. These measures are not captured in standard medical claims databases, and since they are not EHR structured fields, they are not recorded the same way by every clinician. As a result, attempts to extract meaningful insights from these unstructured datasets historically required labor-intensive, manual searches that were not particularly efficient and not at all scalable. With our AI-driven models, we capture Gleason scores for 100% of patients in our prostate cancer dataset. Similarly, by analyzing patterns of diagnosis and patient PSA levels over time, it is also possible to identify patients with localized cancer and evaluate treatment patterns and outcomes over time.

Unlocking A More Complete View of the Patient Population

When it comes to using RWE to transform healthcare, conversations about the completeness of a dataset often center on numbers of covered lives or total database size. Make no mistake, those are important variables. A certain critical mass of clinical encounters is essential to produce meaningful, representative results. But, as the level of sophistication and depth of analysis conducted using RWE continues to expand, it is important for those using this technology to recognize that there is more to data quality and completeness than just raw patient counts.

The next frontier of RWE is using AI and large language models to unlock deeper insights, than ever before possible, from both structured and unstructured datasets. The only way to be certain that we're viewing the full picture is to scour every aspect of the patient journey, from the diagnostic codes to the staging systems to the clinician's insights entered into EHR clinical notes.

By training our algorithms based on rules and nuanced interpretations developed and continually refined as well as validated by practicing clinicians, we are able to not only deliver the most data, or the biggest universe of patients—we are able to deliver the best insights into what's really happening at each step in the healthcare journey of real-world patients with prostate cancer.

To learn more about Qdata Prostate Cancer, and how our RWE solutions for prostate cancer can be used to unlock critical signals and spotlight important trends, click here.