

Intro: When you take a magnifying glass and look at a picture, you see that it's made up of thousands of dots and when you pull back, those dots become the details that create the picture. There's real power in the details, because when you have more detail, the bigger picture becomes sharper and wider, and a story emerges.

Hi, this is Amie Moreno and you're listening to "Seeing the Big Picture: Conversations on how Data and Artificial Intelligence can Add the Details that Fuel Deeper Insights in the Life Sciences Industry."

Amie Moreno: Hi everyone, this is Amie Moreno, and today we are going to discuss

electronic health records and natural language processing. With me today is Mike Sanky, who is the Vice President of the Data and Tools team at Optum Life Sciences and who has extensive knowledge around EHR data, specifically natural language processing. Would you like to tell a little bit about your role at Optum and also some of your background, please?

Mike Sanky: Sure. I've been in the Data Focus Life Sciences space for over a decade now. In my current role, I lead the Client Solutions team with Optum Life Sciences, empowering clients with a variety of patient-level data sets, often enabled through our tools. Our team supports both clients as well as internal research teams using our data assets. In my role, I also lead the Commercial Analytics Consulting Practice, which delivers custom analyses to support product commercialization such as clinically-focused patient journeys. Prior to joining Humedica, which became part of Optum in 2013, I was a consultant with the Amundsen Group, now part of IMS Health, where I worked extensively with longitudinal patient data to better understand patient behavior, quality of access and drivers of product performance. And prior to Amundsen, I worked at the Center for Pricing and Reimbursement at United BioSource Corporation.

In all of these roles, I've used healthcare data to drive strategy for the life sciences industry.

Amie: Thanks. So, you've worked with several different types of data. I'm curious has the electronic health record or EHR data, specifically changed in terms of the way we look at healthcare research?

Mike: So, electronic health records have really been a game-changer in the industry. It's important that people use electronic health records in research because it contains actual clinical measures, not just claims-based proxies that are required to measure value in medicine. It's important to remember that claims systems are designed for reimbursement, not for treating patients. We need electronic health

records to have a true understanding of the impact of treatment, the right drug for the right patient at the right cost.

Let me go into a little more detail about how electronic health records have impacted different functional areas within pharma.

Amie: That'd be great.

Mike: With a commercial lens, we can map the patient journey and understand both disease progression and treatment choice. With the clinical development lens, we can optimize the design of trials by assessing true unmet need and prioritizing clinical development efforts. By understanding real-world treatment patterns and outcomes of representative target patients to inform endpoints, and by optimizing trial eligibility criteria based on real-world patient distributions. From a comparative effectiveness perspective, we can demonstrate real-world effectiveness and establish differential product value. We can create comparable cohorts based on clinical severity, risk factors and performance scores, such as New York Heart Association class. We can measure actual clinical endpoints as opposed to using claims-based proxies and algorithms. And with integrated, that is, linked electronic health record claims data, we can actually measure cost with clinical specificity. Most importantly, electronic health records enable all healthcare stakeholders, life sciences companies, payers, and providers to have a common understanding of value. We can quantify the impact of treatment on total cost of care for specific patient populations, and design value-based contracts accordingly.

Amie: So, one of the valuable and interesting aspects of EHR is the ability to understand physician and patient interactions and scales, measures, things that aren't necessarily available in traditional claims or even structured electronic health records data. We have seen in our conversations with others that NLP, or natural language processing, while it's very valuable, it can be challenging. So, from your perspective and some of the areas where you've used natural language processing, can you talk to us about what has been some of the challenges you face and how you are able to overcome them?

Mike: Amie, I'm so glad you asked that question. There certainly are a lot of challenges with NLP, and really, with unstructured data in general. First of all, the sheer volume of notes, the billions of notes that we capture through our contributing provider groups, and the fact that providers have inconsistencies in how they document things from misspellings to shorthand. But not only does NLP work, it's really the only way to analyze data at scale. Chart reviews and manual extraction have many limitations;

subject to error, subject to bias, subject to time and cost. Let me give you a recent example of where we put natural language processing to work.

We extracted variables using NLP for a prostate cancer cohort. These variables range from Gleason score, which is the severity measure at diagnosis, to hormone response, which is one of the key metrics for disease progression. We actually found that the natural language processing pipeline produced superior results in terms of accuracy for many concepts when compared to the average of two human annotators. So, this was really powerful work and one example of how natural language processing gives us novel insights and the foundational variables that we need to really measure and quantify disease.

Amie: So, just in general, what types data assets do you typically use in your work and how is the integration of those important, whether it be the electronic health records data, claims data, other novel data assets?

Mike: So, I think you've hit on a lot of the key assets. We use all these classes of data on a daily basis in our work, from payment and access data, which includes insurance claims and pharmacy claims, to healthcare delivery data in the form of our very robust electronic health record repository as well as patient powered data, which is a range of social, environmental, and behavioral data. For example, continuous blood glucose monitoring data. All of these data classes come together to create a 360-degree view of the patient and the patient journey through the healthcare system. And we believe that we need to bring these disparate views together to really understand what's happening in the patient journey.

Amie: Right, and to understand that patient holistically.

Mike: Exactly.

Amie: So, in your opinion, what's the most exciting thing that you've been able to use with the EHR data or NLP in the research that you've done?

Mike: So, the most exciting thing to me with natural language processing with unstructured notes in general, and we talked about some of the challenges, it enables us to step into the provider/patient interaction in a way that we just haven't been able to historically. The majority of medical information is contained in unstructured text; unlocking these data give us a better understanding of the why including drivers of treatment choice at critical care milestones. Relatively few data points are captured in structured fields compared to the wealth of insights in the unstructured provider narrative. That's really where we're understanding how the patient is presenting with disease and how the provider is evaluating the patient. These include patient signed diseases and symptoms, social and

family history, treatment rationale, disease-specific severity scores and measures. EMR data in and of itself is relatively unique for health services research even now. Electronic health record data complemented with unstructured provider notes adds a whole other dimension to our clinical insights and our understanding of the patient journey.

Amie: So, you mentioned treatment rationale, can you talk a little more about exactly why that's important and what types of insights you can find that you wouldn't be able in other data assets?

Mike: Sure. So, treatment rationale includes all of the reasons why providers are starting a patient on new medication, stopping a patient, titrating a patient up or down on a medication, perhaps because of tolerability. And with claims data alone, we'll see those changes in medications, but we won't have any insight into why it's happening. With the electronic health record data, we actually see explicit reference to why providers made one medication choice over another and, again, that's really powerful in understanding which medicines are working for which patient populations.

Amie: So, we talked about understanding the motivation for a physician or clinician to make changes in drug treatment. I would imagine that would be something very powerful for, let's say a brand person, a brand team to understand. Can you talk a little bit about why that might be?

Mike: Sure. So, these types of insights would enable a brand team to tailor all of their educational and promotional material for providers. They are able to meet providers where they are in their prescribing paradigm and able to have a conversation with them knowing how they're actually treating their patients in the real-world.

Amie: And that's affecting patient health outcomes, so really all of the different stakeholders that are involved could be impacted by some sort of educational messaging or some sort of marketing campaign, for example? Is that correct?

Mike: Exactly.

Amie: Can you give me an example -- have you experienced that working with a more commercially-focused team -- what insights they were potentially able to garner that allowed them to approach the marketplace differently?

Mike: Sure. So, we had a client that we were working with promoting a portfolio of products competing across multiple indications in the immunology space. Let me give you an example of how we use electronic health record data and, more specifically, insights extracted through natural

language processing to help a commercially-focused team assess the market and implement their strategy differently.

We were helping a team promoting a portfolio of products in the immunology space with diseases including rheumatoid arthritis, ankylosing spondylitis, psoriatic arthritis, plaque psoriasis, Crohn's Disease, and ulcerative colitis. This is a really complex space where there are different specialists such as rheumatologists, dermatologists, gastroenterologists, and it's grown more complex over the years due to new product launches that have different mechanisms of action, different routes of administration, and also existing products that are expanding into new indications. Our client had a need to reevaluate the competitive landscape and define its appropriate market basket across indications and disease specialty, again, to inform and tailor educational and promotional materials.

Amie: In terms of other ways or what other types of data could potentially be used outside of EHR, why do you think there was a value in EHR? I know you're going to get to the details, but as opposed to what other kind of data would they potentially use to attempt to potentially, incorrectly define what their competitive landscape looks like?

Mike: So, I wouldn't say incorrectly, I would say incompletely. The really salient point I want to get across with electronic health record data is that it gives us a more complete view, and with claims data alone, you might be looking at a line of therapy and be using a line of therapy as a proxy for how severe patients are in their disease progression. With the electronic health record data, we can actually clinically measure how far along in the disease progression patients are and what products prescribers are using to treat a similar clinical need. So, rather than take a product attribute approach to define the market, in this case, we used our electronic health record data and we assessed quantitative labs and scores like pain scores, rapid three scores, body surface area for plaque psoriasis, as well as qualitative signs and symptoms like fatigues, stiffness, and swelling, which really are indicative of how rheumatologists in particular are treating disease.

Amie: Do you find that clinicians in this space are using more of those type of symptoms outside of maybe the traditional measures and scores? Is that what you're seeing in real-world practice?

Mike: That's right. So, certainly in clinical trials, we'll see a lot of the severity scores and measures done on a recurring basis because they're part of the endpoints of the trials. Once patients get into the real world, by and large a lot of specialists are relying more on the patient presenting the disease and expressing the symptoms like fatigue and like swelling, like

range of motion in the case of MS, to actually treat those patients and understand whether their medication is working. So, ultimately, in this example for the client, statistically comparing products across the clinical profile, across all of these disease-specific metrics that we talked about, illuminated the competitive set of products that providers were using to address similar clinical needs, which ultimately informed our clients go-to-market strategy for this portfolio of products.

Amie: So, outside of just the commercial space or brand teams, for example, can you give us an example of why this would be important, something like this would be important, whether it's in the, you know, space or any other space for a researcher? So, for a health economics outcome researcher, for example.

Mike: So, certainly in the comparative effectiveness space, these data have been widely adopted to measure product performance and establish differential value. We've also seen a lot of uptake, particularly in recent years in the clinical development space and we've done a lot of work using the notes to assess true unmet need. For example, we had a client investigating in early clinical development around the condition that is under-diagnosed and often misdiagnosed and also comorbid with other conditions like oncology. Using the notes, we built a classifier, incorporating machine learning approaches on both structured electronic health record data as well as the unstructured notes, the patient presentation and signs and symptoms that we've been talking about, to improve patient identification. We ultimately found that the patient population was underrepresented using diagnostic algorithms alone, confirming the unmet need, which ultimately accelerated investment and refined the commercial forecast for this company.

Amie: Interesting, so I noticed you mentioned oncology and that's something I wanted to ask you about in thinking about what other data sets out there do you think could be incorporated to what you're currently using in terms of we're always thinking about the, like you mentioned earlier, the importance of integrating all these different data assets. So, given that things like rare disease, oncology, genomics, there's a lot of information needed that's hard to extract. Have you had an experience with that? Or, is that something that you can see being of value today or going forward?

Mike: So, genomic data is certainly a class of data that I would love to explore. We've worked a lot with biomarker expressions, for example, HER2-positive as extracted using natural language processing from the notes. Complete genomic data would give us a lens into really personalized medicine to better pinpoint the right medicines for the right patients, and it's something that is rapidly growing and we're looking to integrate into our data ecosystem.

Amie: So, with natural language processing, I'm wondering, are there any regulatory agencies in terms of the use of the notes, knowing that they are unstructured in nature and there would obviously be concerns with potential identifiers in the data?

Mike: So, that's a really hot topic of debate in our industry right now. Where are regulatory agencies on the use of electronic health record data and, more importantly, natural language processing? The FDA, actually, just approved PASS, that's Post Authorization Safety Study, based exclusively on the use of electronic health record data including natural language processing derived outputs. The FDA's willingness to embrace NLP to monitor the safety of an approved drug, really is a major breakthrough and, undoubtedly, will lead to wider adoption and use of EHR and NLP among other stakeholders. We're also seeing the FDA begin to consider the use of EHR and NLP not only in this post-authorization world, but also in the approval of new drugs and indications for existing drugs.

Amie: That's really exciting, knowing that the FDA would be willing to use EHR and NLP to do these types of studies, and like you said, it just leads to broader usage and broader uptake of this kind of novel data.

Mike: Absolutely, it's one of the most conservative agencies for good reason and for them to embrace these types of data, I really think we will see a ripple-through effect in the entire industry.

Amie: So, Mike, a lot of what we're talking about here is the differentiation between what we call structured and unstructured data, unstructured being sourced from the notes. Can you define in a little more detail about what the difference is in case not everyone's familiar with that terminology?

Mike: Sure. So, think of it in terms of responding to a survey. You might respond to 10 different questions on a survey that ask to evaluate your preference on a scale of 1 to 5, that would be an example of structured data. And then you might get to the end of the survey and there's a free text box that says, please enter any additional comments that you have -- that would be unstructured data. So, that's the richest component of medical data and that's what we need natural language processing in order to extract those insights.

Amie: Can you also find quantitative data in the physician notes?

Mike: Yes, so both qualitative and quantitative data. We'll see reference to all manner of acronyms and measurements, really anything that the provider

deems relevant to his or her evaluation of the patient on that given day is contained within those free text, unstructured sections of the notes.

Amie: Thanks. Well, Mike, it was a real pleasure talking with you today. I learned a lot. I think our listeners are going to get a lot out of this, so thank you very much for your time today.

Mike: I'm happy to talk to anyone about electronic health record data and natural language processing.