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, my name is Amie Moreno. I'm a Director in the Data, Advanced Analytics and Tools team in our Life Sciences area at Optum, and I am pleased to bring you a podcast which is going to discuss how we use electronic health records, specifically natural language processing, based on some of our experts on our team, and I'm happy to have with me John Seeger, who has done extensive work in this area. So John, I'll let you introduce yourself.

John Seeger: Hello everyone. I'm John Seeger. I'm the Chief Scientific Officer for Epidemiology here at Optum.

Amie: Great. Could you tell us a little bit more about your background in general?

John: Sure. I am trained as a clinical pharmacist and also as an epidemiologist. I have worked for about 18 years in drug safety research at Optum and have used Optum data resources to answer a wide range of drug safety and effectiveness questions.

Amie: Great. So what kinds of data assets do you typically use? I know today we're going to talk about electronic health records, so if you can talk a little bit about how you've used that and what advantages it has, maybe in comparison to other data you've worked with.

John: As a pharmacoepidemiologist, I see the claims data resource that Optum has as a platform for a lot of the work we have done historically in drug safety and effectiveness. It's a platform because it's very well-accepted within the discipline, but the claims data has sort of known limitations with respect to depth and breadth of the data. So, more recently, we've been using the electronic health record data either as a standalone alternative to the claims data or as a supplement to fill in some of the detail where needed.

Amie: Great. And so, I know you've done a lot of work specifically within the natural language processing capability. I know we worked together on a study for binge-eating disorder, which was a challenge as are a lot of therapeutic areas where there might not be structured information or

codes or things of that nature, so it's much more difficult to extract that type of information and understand that patient journey. So, would you talk a little bit about that or other studies where you specifically used natural language processing and EHR to resolve some research questions?

John: Sure, I'd be happy to. The binge-eating disorder study that you mentioned presented a number of challenges, particularly the sort that you describe that the disorder itself was not explicitly coded within the data that we used, the electronic health record data for this study, but it was latent within the notes of electronic medical record systems. What we were able to use was natural language processing to pull out concepts, but we didn't use natural language processing to pull out the specific binge-eating disorder, what we used was the various concepts that were each pulled out, we figured out how to put those concepts together, the combination of those concepts into a construct of binge-eating disorder that we then went through a process of validating through external experts who reviewed what we had done in comparison to the underlying notes. That led us to an algorithm for binge-eating disorder in an electronic health record setting where none of it was coded, it was all drawn from the free text itself. I have to say the natural language processing solved two problems for us. One is a problem of volume, the sheer volume of notes that we were reviewing and would have been impossible to review entirely one at a time, reading through the entire notes, but the natural language processing pulled out the essential components of those notes and did so and allowed us to apply coding to those notes that simplified the task of the, sort of do that train of thought.

The other challenge that the notes represented that was solved by the natural language processing is the privacy, that is the natural language processing did not pull out any protected health information and so that it's a tool for addressing both the volume of information and adding a privacy filter.

Amie: Yeah, and I think that's a really good point because a lot of the information that's in natural language processing are types of information you can get from chart pulls or primary research. But, to your point, that's a manual process that can take a lot of time, but you were able to do this in a grander scale, if you will, and I think it's also important that you mentioned the fact that you were able to check the quality and ensure that it was representative of what you were trying to research -- that's sometimes a question we get from clients or researchers that are using natural language processing. -- because we know there are limitations to any database, but what I think is important about the machine learning that we have and the fact that we have an entire team that works on that is that we do have those quality checks in place and researchers like yourself are able to, in a de-identified manner, account for that.

John: Yeah, the way we did it in this binge-eating disorder study was really quite efficient. We had this initial screen that went through thousands of patients and identified the subset that would potentially have binge-eating disorder and we only needed to validate several dozens of these cases against the classifier based on NLP to come up with a validated algorithm for the identification of binge-eating disorder and so we didn't have to manually review thousands of records to arrive at that place.

Amie: Right. So, is there any other example of how you could use the EHR data or the natural language processing in any other research studies that you'd want to share?

John: Sure. There's another place where we're using the NLP is in the capability that we have called Dynamic Assessments of Pregnancy and Infants, the DAPI, that we use claims as a platform for assessing the effect of drug exposures during pregnancy on pregnancy outcomes, either the outcome of the pregnancy as far as a live birth or a non-live birth, but also on malformations and other outcomes for the infant. What we are able to use is the natural language processed elements that are present in the integrated data, those integrated claims and EHR data, to fill in some covariate information on either the mother or the infant. We're able to pull out things like smoking status and Apgar scores that would not be present in the claims data.

Amie: Right. So, John, we know that there are several different types of data that exist. What do you think is the importance of being able to take those different data assets and integrate them in a way where you get a more holistic perspective of the patient?

John: In the Epidemiology Group here at Optum, we see methods as the primary foundation of our work, but then we apply those methods to a range of data resources and we always seek to match the method with the data and we have this range of data available to us and, on occasion, we'll find a good match with the claims data or with the electronic health record data or with the integrated claims and electronic health record data. In addition, we use other data resources, we use linkages to National Death Index as well as access to survey information or papered medical records from providers.

Amie: So, what else do you think there is in terms of the future for electronic health records and natural language processing, what would you like to see? As I mentioned, there are several experts that are able to extract this information and they’re working on a regular basis to enhance the way we look at machine learning. Is there something specific that you'd like to see or that you see as the future of natural language processing?

John: So, what I'd like to see is sort of expansion in the breadth and depth of what natural language processing pulls out. There's all the targeted elements that are brought out through natural language processing, all the laboratory tests, like the hemoglobin A1c's and other ones. But then there's this generalized NLP and we use this generalized NLP very extensively in this binge-eating disorder study that I was describing earlier, where the generalized natural language processing pulled out a range of elements that were able to be put together to identify the concept of binge-eating disorder even though it hadn't been specifically targeted.

Amie: So, John, in terms of data assets that you may have not had the opportunity to work with yet, what other data is out there that you could potentially use alone or linked to data assets you have access to now and why do you think that's important?

John: Sure. There's a couple of aspects of patient care that we currently don't have a good window on because the data doesn't yet exist in a usable, linkable format, but I would really like to see an expansion of consumer data, both over-the-counter prescription or non-prescription medication use data as well as data that might be derived from wearables and other connected devices that record measurements about people as they go through their daily activities. Those data sources will help fill in these black holes in what we know about patients, either medications that they're using over-the-counter or just things that happen to people throughout the day, information from accelerometers or other devices that will help assess the effect of medications on patients' health status.

Amie: Right. So, in addition to just the clinical drivers, what are those social and behavioral drivers that are causing a patient to behave in a certain way, like a Fitbit?

Amie: John, can you tell me a little bit more about how you use the data specifically in your role? I know it is different than our health economics and outcomes group and our commercial group, so specifically within the pharmacoepidemiology role, how do you and your team use the data?

John: Sure. Thanks, Amie, for the question. I would like to take a step back and think about the discipline of epidemiology, specifically pharmacoepidemiology. As an epidemiologist, I see data as one of three different variables. That is data can be an exposure variable, an outcome variable or a covariate and in each other those, there's three fundamental dimensions of epidemiology data, the person place and time. It can be lots of different people who have different measures of exposure, outcome, or covariate information and they have those same measures at different points in time and so that can have variation across time and then, of

course, variation across place. And so, what we do as pharmacoepidemiologists is use the data to inform those fundamental measures and do so in a structured format that becomes a study design. We use lots of established study designs and apply the structure of informing the various variables that are a part of that study design, so that's our approach to using the data and I use that structure for whichever data resource then I want to use to answer the scientific question.

The Bydureon study was a way to extend a randomized trial. And so, what we did, was we designed a study that mirrored, in a lot of ways, the randomized trial, but because the source data included a broader range of patients, both older age than had been included in the randomized trial and races that were not included or very poorly represented in the randomized trials, we were able to extend the results of this randomized trial to older patients, patients with renal dysfunction and also to African American patients.

Amie: So, you used the EHR to find the EGF or the renal function lab results, right?

John: Yeah, that's right.

Amie: For example.

John: Comparing Bydureon to basal insulin was able to show that there's a benefit with respect to weight and hemoglobin A1c with Bydureon relative to the basal insulins, that was very similar in magnitude to what the randomized trials showed, so that the real-world evidence was able to sort of support that effect of Bydureon in the real world is as the trial had predicted. But then beyond, just that overall effect, we were able to extend those results to subsets of the population where the trials had not been able to include patients with those characteristics, and that's the patients with low renal function and you're absolutely right that we used the estimated glomerular filtration rate that's available in the data to help show the effect of the Bydureon relative to basal insulin during follow-up and then we also did that for hemoglobin A1c, for weight, and then within those subsets of patients who are older and African American and with reduced renal function.

Amie: I think that's a good point that you bring up, because the EHR data is not necessarily replacing other types of research, other types of data, but yet either expanding upon things like clinical trials or with primary research, enabling a better understanding of what types of information you want to extract from those interviews and a better population to serve, so I think it's really important to note that it's additive to other research and other types of data that's out there.

John: That's right. I would say, you know, this was not a replacement for that original study, that is the original randomized trial was really valuable, but this was a way to extend that randomized trial without having to do another randomized trial, but to do it using the real-world data and extend the results to those sub-populations that weren't part of the original trial.

Amie: So, there are a lot of efficiencies that the EHR could be valuable for.

John: Absolutely. It was dramatically more efficient both as far as time and cost to extend the randomized trial through the use of the real-world data than it would have been to do it as another randomized trial.

Having at least the first randomized trial is really important because that can serve as a gauge of the real-world data, but then the real-world data can go beyond that original trial.

Amie: Even just understanding your population in a very clinically-specific way can help kind of better target the types of information you want to understand and the types of patients you want to identify.

John: Yeah, absolutely. One of the things, in particular, the patients with low EGFR, is that they would not have been eligible for that trial, and, you know, there sometimes are ethical considerations, that is if there's a suggestion that whatever treatment might be risky in those sorts of patients, it becomes unethical to include them in a randomized trial where they might receive a treatment that could put them at higher risk, but in real-world clinical practice, sometimes an individual prescriber will make that choice and balance its acceptable risk in this patient. And so, what the real-world data allows us to do is to use all of those individual clinician decisions and draw inferences from what those individual clinicians have to say.

Amie: Well John, thank you so much for sharing your insights about electronic health records and natural language processing. It was very informative and we appreciate your participation today.