PETER SZOLOVITS: OK. So today and next Tuesday, we're talking about the role of natural language processing in machine learning in health care. And this is going to be a heterogeneous kind of presentation. Mainly today, I'm going to talk about stuff that happened or that takes advantage of methods that are not based on neural network representations. And on Tuesday, I'm going to speak mostly about stuff that does depend on neural network representations, but I'm not sure where the boundary is going to fall.

I've also invited Dr. Katherine Liao over there, who will join me in a question and answer session and interview like we did a couple of weeks ago with David. Kat is a rheumatologist in the Partners HealthCare system. And you'll actually be hearing about some of the work that we've done together in the past before we go to the interview. So roughly, the outline of these two lectures is that I want to talk a little bit about why we care about clinical text. And then I'm going to talk about some conceptually very appealing, but practically not very feasible methods that involve analyzing these narrative texts as linguistic entities, as linguistic objects in the way that a linguist might approach them.

And then we're going to talk about what is very often done, which is a kind of term spotting approach that says, well, we may not be able to understand exactly everything that goes on in the narratives, but we can identify certain words and certain phrases that are very highly indicative that the patient has a certain disease, a certain symptom, that some particular thing was done to them. And so this is a lot of the bread and butter of how clinical research is done nowadays. And then I'll go on to some other techniques.

So here's an example. This is a discharge summary from MIMIC. When you played with MIMIC, you notice that it's de-identified. And so names and things are replaced with square brackets, star, star, star kinds of things.

And here I have replaced-- we replaced those with synthetic names. So Mr. Blind isn't really Mr. Blind, and November 15 probably really isn't November 15, et cetera. But I wanted something that read like real text.

So if you look at something like this, you see that Mr. Blind is a 79-year-old white white male-- so somebody repeated a word-- with a history of diabetes mellitus and inferior MI, who underwent open repair of his increased diverticulum on November 13 at some-- again, that's

not the name of the actual place-- medical center. And then he developed hematemesis, so he was spitting up blood, and was intubated for respiratory distress. So he wasn't breathing well.

So these are all really important things about what happened to Mr. Blind. And so we'd like to be able to take advantage of this. And in fact, to give you a slightly more quantitative version of this, Kat and I worked on a project back around 2010 where we were looking at trying to understand what are the genetic correlates of rheumatoid arthritis.

And so we went to the research patient data repository of Mass General and the Brigham Partners HealthCare, and we said, OK, who are the patients who have been billed for a rheumatoid arthritis visit? And there are many thousands of those people, OK? And then we selected a random set of I think 400 of those patients. We gave them to rheumatologists, and we said, which of these people actually have rheumatoid arthritis? So these were based on billing codes.

So what would you guess is the positive predictive value of having a billing code for rheumatoid arthritis in this data set? I mean, how many people think it's more than 50%? OK, that would be nice, but it's not. How many people think it's more than 25%? God, you guys are getting really pessimistic. Well, it also isn't. It turned out to be something like 19% in this cohort.

Now, before you start calling, you know, the fraud investigators, you have to ask yourself why is it that this data is so lousy, right? And there's a systematic reason, because those billing codes were not created in order to specify what's wrong with the patient. They were created in order to tell an insurance company or Medicare or somebody how much of a payment is deserved by the doctors taking care of them.

And so what this means is that, for example, if I clutch my chest and go, uh, and an ambulance rushes me over to Mass General and they do a whole bunch of tests and they decide that I'm not having a heart attack, the correct billing code for that visit is myocardial infarction. Because of course the work that they have to do in order to figure out that I'm not having a heart attack is the same as the work they would have had to do to figure out that I was having a heart attack.

And so the billing codes-- we've talked about this a little bit before-- but they are a very imperfect representation of reality. So we said, well, OK. What if we insisted that you have three billing codes for rheumatoid arthritis rather than just one. And that turned out to raise the

positive predictive value all the way up to 27%.

So we go, really? How could you get billed three times? Right? Well, the answer is that you get billed for, you know, every aspirin you take at the hospital.

And so for example, it's very easy to accumulate three billing codes for the same thing because you go see a doctor, the doctor bills you for a rheumatoid arthritis visit, he or she sends you to a radiologist to take an X-ray of your fingers and your joints. That bill is another billing code for RA. The doctor also sends you to the lab to have a blood draw so that they can check your anti-CCP titer. That's another billing code for rheumatoid arthritis. And it may be that all of this is negative and you don't actually have the disease. So this is something that's really important to think about and to remember when you're analyzing these data.

And so we started off in this project saying, well, we need to get a positive predictive value more on the order of 95%, because we wanted a very pure sample of people who really did have the disease because we were going to take blood samples from those patients, pay a bunch of money to the Broad to analyze them, and then hopefully come up with a better understanding of the relationship between their genetics and their disease. And of course, if you talk to a biostatistician, as we did, they told us that if we have more than about 5% corruption of that database, then we're going to get meaningless results from it. So that's the goal here.

So what we did is to say, well, if you train a data set that tries to tell you whether somebody really has rheumatoid arthritis or not based on just codified data. So codified data are things like lab values and prescriptions and demographics and stuff that is in tabular form. Then we were getting a positive predictive value of about 88%. We said, well, how well could we do by, instead of looking at that codified data, looking at the narrative text in nursing notes, doctor's notes, discharge summaries, various other sources. Could we do as well or better?

And the answer turned out that we were getting about 89% using only the natural language processing on these notes. And not surprisingly, when you put them together, the joint model gave us about 94%. So that was definitely an improvement.

So this was published in 2010, and so this is not the latest hot off the bench results. But to me, it's a very compelling story that says there is real value in these clinical narratives. OK, so how did we do this? Well, we took about four million patients in the EMR. We selected about 29,000 of them by requiring that they have at least one ICD-9 code for rheumatoid arthritis, or that

they've had an anti-CCP titer done in the lab. And then we-- oh, it was 500, not 400.

So we looked at 500 cases, which we got gold standard readings on. And then we trained an algorithm that predicted whether this patient really had RA or not. And that predicted about 35-- well, 3,585 cases. We then sampled a validation set of 400 of those. We threatened our rheumatologists with bodily harm if they didn't read all those cases and give us a gold standard judgment. No, I'm kidding. They were actually really cooperative.

And there are some details here that you can look at in the slide, and I had a pointer to the original paper if you're interested in the details. But we were looking at ICD-9 codes for rheumatoid arthritis and related diseases. We excluded some ICD-9 codes that fall under the general category of rheumatoid diseases because they're not correct for the sample that we were interested in. We dealt with this multiple coding by ignoring codes that happened within a week of each other so that we didn't get this problem of multiple bills from the same visit.

And then we looked for electronic prescriptions of various sorts. We looked for lab tests, mainly RF, rheumatoid factor, and anti-cyclic citrullinated peptide, if I pronounced that correctly. And another thing we found, not only in this study but in a number of others, is it's very helpful just to count up how many facts are on the database about a particular patient.

That's not a bad proxy for how sick they are, right? If you're not very sick, you tend to have a little bit of data. And if you're sicker, you tend to have more data. So these were the cohort selection. And then for the narrative text, we used a system that was built by Qing Zeng and her colleagues at the time-- it was called HITex. It's definitely not state of the art today.

But this was a system that extracted entities from narrative text and did a capable job for its era. And we did this from health care provider notes, radiology and pathology reports, discharge summaries, operative reports. And we also extracted disease diagnosis notes, mentions from the same data, medications, lab data, radiology findings, et cetera.

And then we had augmented the list that came with that tool with the sort of hand-curated list of alternative ways of saying the same thing in order to expand our coverage. And we played with negation detection because, of course, if a note says the patient does not have x, then you don't want to say the patient had x because x was mentioned. And I'll say a few more words about that in a minute.

So if you look at the model we built using logistic regression, which is a very common method,

what you find is that there are positive and negative predictors, and the predictors actually are an interesting mix of ones based on natural language processing and ones that are codified. So for example, you have rheumatoid arthritis. If a note says the patient has rheumatoid arthritis, that's pretty good evidence that they do. If somebody is characterized as being seropositive, that's again good evidence. And then erosions and so on.

But they're also codified things, like if you see that the rheumatoid factor in a lab test was negative, then-- actually, I don't know why that's-- oh, no, that counts against-- OK. And then various exclusions. So these were the things selected by our regularized logistic regression algorithm. And I showed you the results before. So we were able to get a positive predictive value of about 0.94. Yeah?

AUDIENCE: In a the previous slide, you said standardized regression coefficients. So why did you standardize? Maybe I got the words wrong. Just on the previous slide, the--

PETER SZOLOVITS: I think-- so the regression coefficients in a logistic regression are typically just odds ratios, right? So they tell you whether something makes a diagnosis more or less likely. And where does it say standardized?

AUDIENCE: [INAUDIBLE].

PETER SZOLOVITS: Oh, regression standardized. I don't know why it says standardized. Do you know why it says standardized?

KATHERINE LIAO: Couple of things. One is, when you run an algorithm right on your data set, you can't port it using the same coefficients because it's going to be different for each one. So we didn't want people to feel like they can just add it on. The other thing, when you standardize it, is you can see the relative weight of each coefficient.

So it's kind of a measure. Not exactly of how important each coefficient was. That's our way of-- if you can see, we ranked it by the standardized regression coefficient. So NL PRA is up top at 1.11. So that has the highest weight. Whereas the other DMARDs lend it only a little bit more.

PETER SZOLOVITS: OK. Yes?

AUDIENCE: The variables where NL PRA, where it says rheumatoid arthritis in the test, were these

presence of or if they're count?

PETER SZOLOVITS: Yeah. Assuming it's present. So the negation algorithm hopefully would have picked up if it said it's absent and you wouldn't get that feature. All right? So here's an interesting thing. This group, I was not involved in this particular project, said, well, could we replicate the study at Vanderbilt and at Northwestern University?

So we have colleagues in those places. They also have electronic medical record systems. They also are interested in identifying people with rheumatoid arthritis. And so Partners had about 4 million patients, Northwestern had 2.2, Vanderbilt had 1.7. And we couldn't run exactly the same stuff because, of course, these are different systems.

And so the medications, for example, were extracted from their local EMR in very different ways. And the natural language queries were also extracted in different ways because Vanderbilt, for example, already had a tool in place where they would try to translate any text in their notes into UMLS less concepts, which we'll talk about again in a little while. So my expectation, when I heard about this study, is that this would be a disaster. That it would simply not work because there are local effects, local factors, local ways that people have of describing patients that I thought would be very different between Nashville, Chicago, and Boston.

And much to my surprise, what they found was that, in fact, it kind of worked. So the model performance, even taking into account that the way the data was extracted out of the notes and clinical systems was different, was fairly similar. Now, one thing that is worrisome is that the PPV of our algorithm on our data, the way we calculated PPV, they calculated PPV in this study, came in lower than the way we had done it when we found it. And so there is a technical reason for it, but it's still disturbing that we're getting a different result.

The technical reason is described here. Here, the PPV is estimated from a five-fold cross validation of the data, whereas in our study, we had a held out data set from which we were calculating the positive predictive value. So it's a different analysis. It's not that we made some arithmetic mistake.

But this is interesting. And what you see is that if you plot the areas under-- or if you plot the ROC curves, what you see is that training on Northwestern data and testing on either Partners or Vanderbilt data was not so good. But training on either Partners or Vanderbilt data and testing on any of the others turned out to be quite decent. Right? So there is some generality

to the algorithm.

All right, I'm going to switch gears for a minute. So this was from an old paper by Barrows from 19 years ago. And he was reading nursing notes in an electronic medical records system. And he came up with a note which has exactly that text on the left hand side in the nursing note. Except it wasn't nicely separated into separate lines. It was all run together.

So what does that mean? Anybody have a clue? I didn't when I was looking at it. So here's the interpretation. So that's a date. IPN stands for intern progress note. SOB, that's not what you think it means. It's shortness of breath.

And DOE is dyspnea on exertion. So this is difficulty breathing when you're exerting yourself, but that has decreased, presumably from some previous assessment. And the patient's vital signs are stable, so VSS. And the patient is afebrile, AF. OK? Et cetera.

So this is harder than reading the Wall Street Journal because the Wall Street Journal is meant to be readable by anybody who speaks English. And this is probably not meant to be readable by anybody except the person who wrote it or maybe their immediate friends and colleagues. So this is a real issue and one that we don't have a very good solution for yet.

Now, what do you use NLP for? Well, I had mentioned that one of the things we want to do is to codify things that appear in a note. So if it says rheumatoid arthritis, we want to say, well, that's equivalent to a particular ICD-9 code.

We might want to use natural language processing for de-identification of data. I mentioned that before. You don't, MIMIC, the only way that Roger Mark's group got permission to release that data and make it available for people like you to use is by persuading the IRB that we had done a good enough job of getting rid of all the identifying information in all of those records so that it's probably not technically impossible, but it's very difficult to figure out who the patients actually were in that cohort, in that database.

And the reason we ask you to sign a data use agreement is to deal with that residual, you know, difficult but not necessarily impossible because of correlations with other data. And then you have little problems like Mr. Huntington suffers from Huntington's disease, in which the first Huntington is protected health information because it's a patient's name. The second Huntington is actually an important medical fact. And so you wouldn't want to get rid of that one.

You want to determine aspects of each entity. Its time, its location, its degree of certainty. You want to look for relationships between different entities that are identified in the text. For example, does one precede another, does it cause it, does it treat it, prevent it, indicate it, et cetera?

So there are a whole bunch of relationships like that that we're interested in. And then also, for certain kinds of applications, what you'd really like to do is to identify what part of a textual record addresses a certain question. So even if you can't tell what the answer is, you should able to point to a piece of the record and say, oh, this tells me about, in this case, the patient's exercise regimen.

And then summarization is a very real challenge as well, especially because of the cut and paste that has come about as a result of these electronic medical record systems where, when a nurse is writing a new note, it's tempting and supported by the system for him or her to just take the old note, copy it over to a new note, and then maybe make a few changes. But that means that it's very repetitive. The same stuff is recorded over and over again. And sometimes that's not even appropriate because they may not have changed everything that needed to be changed.

The other thing to keep in mind is that there are two very different tasks. So for example, if I'm doing de-identification, essentially I have to look at every word in a narrative in order to see whether it's protected health information. But there are often aggregate judgments that I need to make, where many of the words don't make any difference. And so for example, one of the first challenges that we ran back in 2006 was where we gave people medical records, narrative text records from a bunch of patients and said, is this person a smoker?

Well, you can imagine that there are certain words that are very helpful like smoker or tobacco user or something like that. But even those are sometimes misleading. So for example, we saw somebody who happened to be a researcher working on tobacco mosaic virus who was not a smoker.

And then you have interesting cases like the patient quit smoking two days ago. Really? Are they a smoker or not? And also, aggregate judgment is things like cohort selection, where it's not every single thing that you need to know about this patient. You just need to know if they fit a certain pattern.

So let me give you a little historical note. So this happened to be work that was done by my PhD thesis advisor, the gentleman whose picture is on the slide there. And he published this paper in 1966 called *English for the Computer* in the *Proceedings of the Fall Joint Computer Conference.* This was the big computer conference of the 1960s.

And his idea was that the way to do English, the way to process English is to assume that there is a grammar, and any English text that you run across, you parse according to this grammar. And that each parsing rule corresponds to some semantic function. And so the picture that emerges is one like this. Where if you have two phrases and they have some syntactic relationship between them, then you can map each phrase to its meaning.

And the semantic relationship between those two meanings is determined by the syntactic relationship in the language. So this seems like a fairly obvious idea, but apparently nobody had tried this on a computer before. And so Fred built, over the next 20 years, computer systems, some of which I worked on that tried to follow this method.

And he was, in fact, able to build systems that were used by researchers in areas like anthropology, where you don't have nice coded data and where a lot of stuff is in narrative text. And yet he was able to help one anthropologist that I worked with at Caltech to analyze a database of about 80,000 interviews that he had done with members of the Gwembe Tonga tribe, who lived in the valley that is now flooded by the Zambezi River Reservoir on the border of Zambia and Zimbabwe. That was fascinating. Again, he became very well known for some of that research.

In the 1980s I was amused to see that SRI-- which doesn't stand for anything, but used to stand for Stanford Research Institute-- built a system called Diamond Diagram, which was intended to help people interact with the computer system when they didn't know a command language for the computer. So they could express what they wanted to do in English and the English would be translated into some semantic representation. And from that, the right thing was triggered in the computer.

So these guys, Walker and Hobbs, said, well, why don't we apply this idea to natural language access to medical text? And so they built a system that didn't work very well, but it tried to do this by essentially translating the English that it was reading into some formal predicate calculus representation of what they saw, and then a process for that system. The original Diamond Diagram system that was built for people who were naive computer users and didn't

know command languages actually had a very rigid syntax.

And so what they discovered is that people are more adaptable than computers and that they could adapt to this rigid syntax. How many of you have Google Home or Amazon Echo or Apple something or other that you deal with? Well, so it's training you, right? Because it's not very good at letting you train it, but you're more adaptable.

And so you quickly learn that if you phrase things one way, it understands you, and if you phrase things a different way, it doesn't understand you. And you learn how to phrase it. So that's what these guys are relying on, is that they can get people to adopt the conventions that the computer is able to understand.

The most radical version of this was a guy named de Heaulme, who I met in 1983 in Paris. He was a doctor Le Pitie Salpetriere, which is one of these medieval hospitals in Paris. And it's wonderful place, although when they built it, it was just a place to die because they really couldn't do much for you.

So de Heaulme convinced the chief of cardiology at that hospital that he would develop an artificial language for taking notes about cardiac patients. He would teach this to all of the fellows and junior doctors in the cardiology department at the hospital. And they would be required by the chief, which is very powerful in France, to use this artificial language to write notes instead of using French to write notes. And they actually did this for a month.

And when I met de Heaulme, he was in the middle of analyzing the data that he had collected. And what he found was that the language was not expressive enough. There were things that people wanted to say that they couldn't say in this artificial language he had created. And so he went back to create version two, and then he went back to the cardiologist and said, well, let's do this again. And then they threatened to kill him. So the experiment was not repeated.

OK, so back to term spotting. Traditionally, if you were trying to do this, what you would do is you would sit down with a bunch of medical experts and you would say, all right, tell me all the words that you think might appear in a note that are indicative of some condition that I'm interested in. And they would give you a long list. And then you'd do grep, you'd search through the notes for those terms. OK? And if you want it to be really sophisticated, you would use an algorithm like NegEx, which is a negation expression detector that helps get rid of things that are not true.

And then, as people did this, they said, well, there must be more sophisticated ways of doing this. And so a whole industry developed of people saying that not only should we use the terms that we got originally from the doctors who were interested in doing these queries, but we can define a machine learning problem, which is how do we learn the set of terms that we should actually use that will give us better results than just the terms we started with? And so I'm going to talk about a little bit of that approach.

First of all, for negation, Wendy Chapman, now at Utah, but at the time at Pittsburgh, published this paper in 2001 called *A Simple Algorithm for Identifying the Gated Findings of Diseases in Discharge Summaries.* And it is indeed a very simple algorithm. And here's how it works. You find all the UMLS terms in each sentence of a discharge summary. So I'll talk a little bit about that. But basically, it's a dictionary look up. You look up in this very large database of medical terms and translate them into some kind of expression that represents what that term means.

And then you find two kinds of patterns. One pattern is a negation phrase followed within five words by one of these UMLS terms. And the other is a UMLS term followed within five words by a negation phrase, different set of negation phrases. So if you see no sign of something, that means it's not present. Or if you see ruled out, unlikely something, then it's not present. Absence of, not demonstrated, denies, et cetera.

And post modifiers if you say something declined or something unlikely, that also indicates that it's not present. And then they hacked up a bunch of exceptions where, for example, if you say gram negative, that doesn't mean that it's negative for whatever follows it or whatever precedes it, right? Et cetera. So there are a bunch of exceptions.

And what they found is that this actually, considering how incredibly simple it is, does reasonably well. So if you look at sentences that do not contain a negation phrase and looked at 500 of them, you find that you get a sensitivity and specificity of 88% and 52% for those that don't contain one of these phrases. Of course, the sensitivity is 0 and the specificity is 100% on the baseline.

And if you use NegEx, what you find is that you can significantly improve the specificity over the baseline. All right? And you wind up with a better result, although not in all schemes. So what this means is that very simplistic techniques can actually work reasonably well at times.

So how do we do this generalization? One way is to take advantage of related terms like hypo-

or hypernyms, things that are subcategories or super categories of a word. You might look for those other associated terms. For example, if you're looking to see whether a patient has a certain disease, then you can do a little bit of diagnostic reasoning and say, if I see a lot of symptoms of that disease mentioned, then maybe the disease is present as well.

So the recursive machine learning problem is how best to identify the things associated with the term. And this is generally known as phenotyping. Now, how many of you have used the UMLS? Just a few.

So in 1985 or '84, the newly appointed director of the National Library of Medicine, which is one of the NIH institutes, decided to make a big investment in creating this unified medical language system, which was an attempt to take all of the terminologies that various medical professional societies had developed and unify them into a single, what they called a meta-thesaurus. So it's not really a thesaurus because it's not completely well integrated, but it does include all of this terminology.

And then they spent a lot of both human and machine resources in order to identify cases in which two different expressions from different terminologies really meant the same thing. So for example, myocardial infarction and heart attack really mean exactly the same thing. And in some terminologies, it's called acute myocardial infarction or acute infarct or acute, you know, whatever. And they paid people and they paid machines to scour those entire databases and come up with the mapping that said, OK, we're going to have some concept, you know, see 398752-- I just made that up-- which corresponds to that particular concept. And then they mapped all those together.

So that's an enormous help in two ways. It helps you normalize databases that come from different places and that are described differently. It also tells you, for natural language processing, how it is-- it gives you a treasure trove of ways of expressing the same conceptual idea. And then you can use those in order to expand the kinds of phrases that you're looking for.

So there are, as of the current moment, there are about 3.7 million distinct concepts in this concept base. There are also hierarchies and relationships that are imported from all these different sources of terminology, but those are a pretty jumbled mess. And then over the whole thing, they created a semantic network that says there are 54 relations and 127 types, and every concept unique identifier is assigned at least one semantic type. So this is very useful for

looking through this stuff.

Here are the UMLS semantic concepts of various-- or the semantic types. So you see that the most common semantic type is this T061, which stands for therapeutic or preventive procedure. And there are 260,000 of those concepts in the meta-thesaurus. There are 233,000 findings, 172,000 drugs, organic chemicals, pharmacological substances, amino acid peptide or protein, invertebrate. So the data does not come only from human medicine but also from veterinary medicine and bioinformatics research and all over the place.

But you see that these are a useful listing of appropriate semantic types that you can then look for in such a database. And the types are hierarchically organized. So for example, the relations are organized so there's an effects relation which has sub-relations, manages, treats, disrupts, complicates, interacts with, or prevents. Something like biological function can be a physiologic function or a pathologic function. And again, each of these has subcategories.

So the idea is that each concept, each unique concept is labeled with at least one of these semantic types, and that helps to identify things when you're looking through the data. There are also some tools that deal with the typical linguistic problems, that if I want to say bleeds or bleed or bleeding, those are really all the same concept. And so there are these lexical variant generator that helps us normalize that.

And then there is the normalization function that takes some statement like Mr. Huntington was admitted, blah, blah, blah, and normalizes it into lowercase alphabetized versions of the text, where things are translated into other potential meanings, linguistic meanings of that text. So for example, notice this one says was, but one of its translations is be because was is just a form of be.

This can also get you in trouble. I ran into a problem where I was finding beryllium in everybody's medical records because it also knows that b-e is an abbreviation for beryllium. And so you have to be a little careful about how you use this stuff.

There is an online tool where you can type in something and it says weakness of the upper extremities. And it says, oh, you mean the concept proximal weakness, upper extremities. And then it has a relationship to various contexts and it has siblings and it has all kinds of other things that one can look up.

I built a tool a few years ago where if you populated with one of the short summaries, it tries to

color code the types of things that it found in that summary. And so this is using a tool called MetaMap, which again comes from the National Library of Medicine, and a locally built UMLS look up tool that in this particular case finds exactly the same mappings from the text. And so you can look through the text and say, ah, OK, so no indicates negation and urine output is a kind of one of these concepts. If you moused over it, it would show you.

OK, I think what I'm going to do is stop there today so that I can invite Kat to join us and talk about A, what's happened since 2010, and B, how is this stuff actually used by clinicians and clinician researchers. Kat? OK, well, welcome, Kat.

KATHERINE LIAO: Thank you.

PETER SZOLOVITS: Nice to see you again. So are the techniques that were represented in that paper from nine years ago still being used today in research settings?

KATHERINE LIAO: Yeah. So I'd say yes, the bare bones of platform-- that pipeline is being used. But now I'd say we're in version five. Actually, you were on that revision list. But we've done a lot of improvements to actually automate things a little more. So the rate limiting factor in phenotyping is always the clinician. Always getting that label, doing the chart review, coming up with that term list. So I don't know if you want me to go into some of the details on what we've been doing.

PETER SZOLOVITS: Yeah, if you would.

KATHERINE LIAO: Kind of plugs it in. So if you recall that diagram, there were several steps, where you started with the EMR. There was that filter with the ICD codes. Then you get this data mart, and then you start training. You had to select a random 500, which is a lot. It's a lot of chart review to do. It is a lot. So our goal was to reduce that amount of chart review.

And part of the way to reduce that is reducing the feature space. So one of the things that we didn't know when we first started out was how many gold standard labels did we need and how many features did we need and which of those features would be important. So by features, I mean ICD codes, a diagnosis code, medications, and all that list of NLP terms that might be related to the condition.

And so now we have ways to try to whittle down that list before we even use those gold standard labels. And so let me think about-- this is NLP. The focus here is on NLP. So there

are a couple of ways we're doing this. So one rate limiting step was getting the clinicians to come up with a list of terms that are important for a certain condition. You can imagine if you get five doctors in a room to try to agree on a list, it takes forever.

And so we tried to get that out of the way. So one thing we started doing was we took just common things that are freely available on the web. Wikipedia, Medline, the Merck Manual that have medical information. And we actually now process those articles, look for medical terms, pull those out, map them to concepts, and that becomes that term list.

Now, that goes into-- so now instead of, if you think about in the old days, we came up with the list, we had ICD lists and term lists, which got mapped to a concept. Now we go straight to the article. We kind of do majority voting with the articles.

We take five articles, if three out of five mention it more than x amount of time, we say that could potentially be important. So that's the term list. Get the clinicians out of that step. Well, actually, we don't train yet. So now instead of training right away in the gold standard labels, we train on a silver standard label.

Most of the time, we use the main ICD code, but sometimes we use the main NLP [INAUDIBLE] Because sometimes there is no code for the phenotype we're interested in. So that's kind of some of the steps that we've done to automate things a little bit more and formalize that pipeline. So in fact, the pipeline is now part of the Partners Biobank, which is a Partner's Healthcare. As Pete mentioned, it's Mass General and Brigham Women's Hospital.

They are recruiting patients to come in and get the blood sample, link it with their notes so people can do research on linked EHR data and blood sample. So this is the pipeline they used for phenotyping. Now I'm over at the Boston VA along with Tianxi. And this is the pipeline we're laying down for also the Million Veterans program, which is even bigger. It's a million vets and they have EHR data going back decades. So it's pretty exciting.

PETER SZOLOVITS: So what are the kinds of-- I mean, this study that we were talking about today was for rheumatoid arthritis. What other diseases are being targeted by this phenotyping approach?

KATHERINE LIAO: So all kinds of diseases. There's a lot of things we learn, though. The phenotyping approach is best suited, the pipeline that we-- the base pipeline is best suited for conditions that have a prevalence of 1% or higher. So rheumatoid arthritis is kind of at that lower bound. Rheumatoid arthritis is a chronic inflammatory joint disease. It affects 1% of the population. But it is the

most common autoimmune joint disease.

Once you go to rare diseases that are episodic that don't happen-- you know, not only is it below 1%, but only happens once in a while-- this type of approach is not as robust. But most diseases are above 1%. So at the VA, we've kind of laid down this pipeline for a phonemic score. And they're running through acute stroke, myocardial infarction, all kinds of these-- diabetes-- just really a lot of all the common diseases that we want to study.

PETER SZOLOVITS: Now, you were mentioning that when you identify such a patient, you then try to get a blood sample so that you can do genotyping on them. Is that also common across all these diseases or are there different approaches?

KATHERINE LIAO: Yeah, so it's interesting. 10 years ago, it was very different. It was very expensive to genotype a patient. It was anywhere between $500 to $700 per patient.

PETER SZOLOVITS: And that was just for single nucleotide polymorphism.

KATHERINE LIAO: Yes, just for a snip. So we had to be very careful about who we selected. So 10 years ago, what we did is we said, OK, we have 4 million patients and partners. Who has already with good certainty?

Then we select those patients and we genotype them. Because it costs so much, you didn't want to genotype someone who didn't have RA. Not only would it alter the-- it would reduce the power of our association study, it would just be like wasted dollars. The interesting thing is that the change has happened.

And we can completely think of a different way of approaching things. Now you have these biobanks. You have something like the VA MVP or UK Biobank. They are being systematically recruited, blood samples are taken, they're genotyped with no study in mind. Linked with the EHR.

So now I walk into the VA, it's a completely different story. 10 years later, I'm at the VA and I'm interested in identifying rheumatoid arthritis. Interesting enough, this algorithm ports well over there, too. But now we tested our new method on there. But now, instead of saying, I need to identify these patients and get the genotype, all the genotypes are already there. So it's a completely different approach to research now.

PETER SZOLOVITS: Interesting. So the other question that I wanted to ask you before we turn it over to questions from the audience is, so this is all focused on research uses of the data. Are there clinical uses that people have adopted that use this kind of approach to trying to read the note? We had fantasized decades ago that, you know, when you get a report from a pathologist, that somehow or other, a machine learning algorithm using natural language processing would grovel over it, identify the important things that came out, and then either incorporate that in decision support or in some kind of warning systems that drew people's attention to the important results as opposed to the unimportant ones. Has any of that happened?

KATHERINE LIAO: I think we're not there yet, but I feel like we're so much closer than we were before. That's probably how you felt a few decades ago. One of the challenges is, as you know, EHR weren't really widely adopted until the HITEx Act in 2010.

So a lot of systems are actually now just getting their EHR. And the reason that we've had the luxury of playing around with the data is because Partners was ahead of the curve and had developed an EHR. The VA happened to have an EHR.

But I think first-- because research and clinical medicine is very different. Research, if you mess up and you misclassify someone with a disease, it's OK, right? You just lose power in your study. But in the clinical setting, if you mess up, it's a really big deal. So I think the bar is much higher.

And so one of our goals with all this phenotyping is to get it to that point where we feel pretty confident. We're not going to say someone has or hasn't a disease, but we are, you know, Tianxi and I have been planning this grant where, what's outputted from this algorithm is a probability of disease. And some of our phenotype algorithms are pretty good.

And so what we want to test is what threshold is that probability that you would want to tell a clinician that, hey, if you're not thinking about rheumatoid arthritis in this patient-- this is particularly helpful in places where they're in remote locations where there aren't rheumatologist available-- you should be thinking about it and maybe, you know, considering referring them or speaking to a rheumatologist through telehealth, which is also something. There's a lot of things that are changing that are making something like this fit much more into the workflow.

PETER SZOLOVITS: Yeah. So you're as optimistic as I was in the 1990s.

KATHERINE LIAO: Yes. I think we're getting-- we'll see.

PETER SZOLOVITS: Well, you know, it will surely happen at some point. Did any of you go to the festivities around the opening of the Schwarzman College of Computing? So they've had a lot of discussions. And health care does keep coming up over and over again as one of the great opportunities. I profoundly believe that.

But on the other hand, I've learned over many decades not to be quite as optimistic as my natural proclivities are. And I think some of the speakers here have not yet learned that same lesson. So things may take a little bit longer. So let me open up the floor to questions.

KATHERINE LIAO: Yes?

AUDIENCE: So the mapping that you did to concepts, is that within the Partners system or is that something like publicly available? And can you just transfer that to the VA? Or like, when you do work like, how much is proprietary and how much gets expanded up?

KATHERINE LIAO: Yeah. So you're speaking about when we were trying to create that term list and we mapped the terms to the concepts?

AUDIENCE: And you were using Wikipedia and three other sources.

KATHERINE LIAO: Yeah. Yeah. So that's all out there. So as an academic group, we try to publish everything we do. We put our codes up on GitHub or CRAN for other people to play out and tests and break. So yeah, the terms are really similar in UMLS. I don't know if you had a chance to look through it. They have a lot of keywords. So there is a general way to map keywords to terms to concepts.

So that's the basis of what we do. There may maybe a little bit more there, but there's nothing fancy behind it. And as you can imagine, because we're trying to go across many phenotypes, when we think about mapping, it always has to be automated. Our first round was very manual, incredibly manual. But now we try to use systems that are available such as UMLS and other mapping methods.

PETER SZOLOVITS: So what map-- presumably, you don't use HITex today.

KATHERINE LIAO: No.

PETER SZOLOVITS: So which tools do you use?

KATHERINE LIAO: Just thinking I had a two hour conversation with Oakridge about this. We're using a system that Cheng developed called NIAL. And it had to do with the fact that cTAKES, which is a really robust system, was just too computationally intensive.

And for the purposes of phenotyping, we didn't need that level of detail. What we really needed was, was it mentioned, what's the concept, and the negation. And so NIAL is something that we've been using and have kind of validated over time with the different methods we've been testing.

PETER SZOLOVITS: So Tuesday, I'll talk a little bit about that system and some of its successors. So you'll get a sense of how that works. I should mention also that one of the papers that was on your reading list is a paper out of David Sontag's group, which uses this anchorous concept. And that's very much along the same lines.

That it's a way of trying to automate, just as Kat was saying, you know, if the doctor's mention some term and you discover that that term is very often used with certain other terms by looking at Wikipedia or at the Mayo Clinic data or wherever your sources are, then that's a good clue that that other term might also be useful. So this is a formalization of that idea as a machine learning problem.

So basically, that paper talks about how to take some very certain terms that are highly indicative of a disease and then use those as anchors in order to train a machine learning model that identifies more terms that are also likely to be useful. So this notion of-- and David talked about a similar idea in a previous lecture, where you get a silver standard instead of a gold standard. And the silver standard can be derived from a smaller gold standard using some machine learning algorithm. And then you can use that in your further computations.

AUDIENCE: So what was the process like for partnering with academics and machine learning? So did you seek them out, did they seek you out? Did you run into each other at the bus stop? How does that work?

KATHERINE LIAO: Well, I was really lucky. There was a big study called *The Informatics for Integrating Biology and the Bedside Project* called i2B2 led by Zak Kohane And so that was already in place. And

Pete had already been pulled in and Tianxi. So what they basically did was locked all us in a room for three hours every Friday. And it was like, what's the problem, what's the question, and how do we get there.

And so I think that infrastructure was so helpful in bringing everyone to the table, because it's not easy because you're not rotating in the same space. And the way you think is very different. So that's how we did it.

Now it's more mainstream. I think when we first started, everyone was-- my colleagues joked with me. They're like, what are you doing? R2D2? What's going on? Are you going off the deep end over there? Because you know, the type of research we do was more along the ways of clinical trials and clin-epi projects.

But now, you know, we have-- I run a core at Brigham. So it's run out of the rheumatology division. And so we kind of try to connect people together. I did post to our core the consulting session here. But you know, if there is interest, there's probably more groups that are doing this, where we can kind of more formally have joint talks or connect people together. Yeah. But it's not easy.

I have to say, it takes a lot of time. Because when Pete put up that thing in what looked like a different language, I mean, it didn't even occur to me that it was hard to read, right? So it's like, you know, you're into these two different worlds. And so you have to work to meet in the middle, and it takes time.

PETER SZOLOVITS: It also takes the right people. So I have to say that Zak was probably very clever in bringing the right people to the table and locking those into that room for three hours at a time because, for example, our biostatistician, Tianxi Cai just, you know, she speaks AI or she has learned to speak AI. And there are still plenty of statisticians who just have allergic reactions to the kinds just things that we do, and it would be very difficult to work with them. So having the right combination of people is also really I think critical.

KATHERINE LIAO: As one of my mentors said, you have to kiss a lot of frogs.

AUDIENCE: I wondering if you could say a bit more about how you approached the alarm fatigue with how you balance [INAUDIBLE] question around how certain you are versus clinical questions of how important this is versus even psychological questions of, I said is too often to a certain amount of people. They're going to start [INAUDIBLE]?

KATHERINE LIAO: Yeah, you've definitely hit the nail on the head of one of the major barriers, or several things. The alarm fatigue is one of them. So EMRs became more prominent in 2010. But now, along with EMRs came a lot of regulations on physicians. And then came getting rid of our old systems for these new systems that are now government compliant.

So Epic is this big monster system that's being rolled out across the country, where you literally have-- it's so complicated in places like Mayo. They hire scribes. The physicians sits in the office and there's another person who actually listens in and types and then clicks all the buttons that you need to get the information there. So alarm fatigue is definitely one of the barriers.

But the other barrier is the fact that the EMRs are so user-unfriendly now. They're not built for clinical care. They're built for billing. We have to be careful about how we roll this out. And that's one reason why I think things have been held up, actually. Not necessarily the science. It's the implementation part is going to be very hard.

PETER SZOLOVITS: So that isn't new, by the way. I remember a class I taught in biomedical computing about 15 years ago. David Bates, who's the chief of general internal medicine or something at the Brigham, came in and gave a guest lecture. And he was describing their experience with a drug-drug interaction system that they had implemented. And they purchased a data set from a vendor called First Databank that had scoured the literature and found all the instances where people had reported cases where a patient taking both this medication and that medication had an apparent adverse event. So there was some interaction between them.

And they bought this thing, they implemented it, and they discovered that, on the majority of drug orders that they were making through their pharmacy system, a big red alert would pop up saying, you know, are you aware of the fact that there is a potential interaction between this drug and some other drug that this patient is taking. And the problem is that the incentives for the company that curated this database were to make sure they didn't miss anything, because they didn't want to be responsible for failing to alarm.

But of course, there's no pushback saying that if you warn on every second order, then no one's going to pay any attention to any of them. And so David's solution was to get a bunch of the senior doctors together and they did some study of what actual adverse events had they experienced at the hospital. And they cut this list of thousands of drug interactions down to 20. And they said, OK, those are the only ones we're going to alarm on.

KATHERINE LIAO: And then they threw that out when Epic came in. So now I put in an order, I get like a list of 10 and I just click them all. So that's the problem. And the threshold is going to be-- so there's going to be an entire-- I think there's going to be entire methods development that's going to have to happen between figuring out where that threshold is and the fatigue from the alarms.

AUDIENCE: I have two questions. One is about [INAUDIBLE]. Like how did you approach that because we talk about this in other contexts in class? And the other one is, like, how can you inform other countries [INAUDIBLE] done here?

Because, I mean, at the end of the day, it's a global health issue. And also drug systems are different even between the US and the UK. So all the mapping we're doing here, how could that inform EHR or elsewhere?

KATHERINE LIAO: Yeah. So let me answer the first one. The second one is a work in progress. So ICD-10 came to the US on October 1, 2015. I remember. It hurt us all. So we actually don't have that much information on ICD-10 yet.

But it's definitely impacted our work. So if you think about when Pete was pointing to the number of ICD counts for ICD-9, for those of you who don't know, ICD-9 was developed decades ago. ICD-10 maybe two decades ago. But what ICD-10 did was it added more granularity.

So for rheumatoid arthritis, I mentioned it's a systemic chronic inflammatory joint disease. We used to have a code that said rheumatoid arthritis. In ICD-10, it now says rheumatoid arthritis, rheumatoid factor positive, rheumatoid arthritis, rheumatoid factor negative. And under each category is RA of the right wrist, RA of the left wrist, RA of the right knee, left knee. Can you imagine? So we're clicking off all of these.

And so, as it turns out, surprisingly-- we're about to publish a small study now, is RA any more accurate now they have all these granular-- it turns out, I think we got annoyed because it's actually less accurate now than the ICD-9. So that's one thing. But that's, you know, only two or three years of data. I think it's going to become pretty equivalent.

The other thing is, you'll see an explosion in the number of ICD codes. So you have to think about how do you deal with back October 1, 2015 when you had one RA code, but after 2015, it depends on when the patient comes in. They may have RA of the right wrist on one day, then on the left knee the other day. That looks like a different code.

So right now, we have to think of systematic systems to roll up. I think the biggest challenge right now is the mapping. So ICD-9, you know, doesn't map directly to ICD-10 or back because there were diseases that we didn't know when they developed ICD-9 that exist in ICD-10. In ICD-10, they talk about diseases in ways that weren't described in ICD-9.

So when you're trying to harmonize the data, and this is actively something we're dealing with right now at the VA, how do you now count the ICD codes? How do you consider that someone has an ICD code for RA? So those are all things that are being developed now. CMS, Center for Medicaid and Medicare, again, this is for billing purposes, has come up with a mapping system that many of us are using now, given what we have.

PETER SZOLOVITS: And by the way, the committee that is designing ICD-11 has been very active for years. And so there is another one coming down the pike. Although, from what I understand--

KATHERINE LIAO: Are you involved with that?

PETER SZOLOVITS: No. But Chris Chute is or was.

KATHERINE LIAO: Yes, I saw. I said, don't do it.

PETER SZOLOVITS: Well, but actually, I'm a little bit optimistic because unlike the traditional ICD system, this one is based on SNOMED, which has a much more logical structure. So you know, my favorite ICD-10 code is closed fracture of the left femur due to spacecraft accident.

KATHERINE LIAO: I didn't even know that existed.

PETER SZOLOVITS: As far as I know, that code has never been applied to anybody. But it's there just in case. Yeah.

AUDIENCE: So wait, for the ICD-11 you don't think take that long to exist because it's a more logical system?

PETER SZOLOVITS: So ICD-11-- well, I don't know what it's going to be because they haven't defined it yet. But the idea behind SNOMED is that it's more a combinatorial system. So it's more like a grammar of descriptions that you can assemble according to certain rules of what assemblies make sense. And so that means that you don't have to explicitly mention something like the spacecraft accident one. But if that ever arises, then there is a way to construct something that would

describe that situation.

KATHERINE LIAO: I ran into Chris at a meeting and he said something along the lines that he thinks it's going to be more NLP-based, even. I don't know. Is it going to be more like a language?

PETER SZOLOVITS: Well, you need to ask him.

KATHERINE LIAO: Yeah, I don't know. He hints at it [INAUDIBLE]. I was like, OK, this will be interesting.

PETER SZOLOVITS: I think it's definitely more like a language, but it'll be more like the old Fred Thompson or the Diamond Diagram kind of language. It's a designed language that you're going to have to learn in order to figure out how to describe things appropriately. Or at least your billing clerk will have to learn it. Yeah?

AUDIENCE: I know we're towards the end. But I had a question about when a clinician is trying to label data, for example, training data, are there any ambiguities ever, where sometimes this is definitely-- this person has RA. This person, I'm not really sure. How do you take that into account when you're actually training a [INAUDIBLE]?

KATHERINE LIAO: Yeah. So we actually have three categories-- definite, possible, and no. So there is always ambiguity. And then you always want to have more than one reviewer. So in clinical trials when you have outcomes, you have what we call adjudication.

So you have some kind of system where you have the first sit down, you have to define the phenotype. Because not everybody is going to agree, even for a really clear disease, how do you define the disease. What are the components that has to happen. For that, they're usually for societies or classification criteria for research.

So there actually is one for RA, you know, for coronary artery disease. And then it is having those different categories in a very structured system for adjudicating. You know, blindly having two reviewers review 20, you know, let's say 20 of the same notes and look at the integrated reliability. Yeah. That's a big issue.

PETER SZOLOVITS: All right. I think we have expired. So Kat, thank you very much.

KATHERINE LIAO: Yes, thank you, everybody.