NATIONAL INSTITUTES OF HEALTH

Applications for Enhancing Clinical Decision Making

8:15 a.m. to 5:01 p.m.
Tuesday, April 24, 2012

Lister Hill Auditorium
National Institutes of Health
Bethesda, Maryland

OLENDER REPORTING, INC.
1100 Connecticut Avenue, N.W., #810, Washington D.C. 20036
Washington: (202) 898-1108 / Baltimore: (410) 752-3376
Toll Free: (888) 445-3376

C O N T E N T S

Page

Welcome and Objectives
Co-chair Dr. Belinda Seto, NIBIB 4

Overview of CDS in Healthcare
Co-chair Dr. Blackford Middleton,
Partners Healthcare System and
Harvard Medical School 6

Panel 1: Clinical Perspective

Moderator: Dr. Richard Conroy, NIBIB 37
Keynote: Dr. Thomas Payne,
University of Washington 39
Dr. Frank Sonnenberg,
University of Medicine and Dentistry of
New Jersey 64
Dr. James Walker, Geisinger Health System 80
Dr. Eliot Siegel, University of Maryland 99
Panel Discussion

Panel 2:
Perspectives on Clinical Decision Support

Moderator: Dr. James Luo, NIBIB 139
Keynote: Dr. Robert Greenes,
Arizona State University 141
Dr. Li Zhou, Partners Healthcare System 166
Dr. Stephane Meystre, University of Utah 179
Dr. Mark Musen, Stanford University 199
Panel Discussion

Panel 3: Stakeholder Perspectives

Moderator: Steve Lohr, New York Times 247
Keynote: Jacob Reider, M.D. 249
Dr. Gregory Steinberg, Aetna 278
Dr. Marc Overhage, Siemens 293
Dr. David Gondek, IBM 306
Panel Discussion

C O N T E N T S (continued)

Page

Panel 4:
Future Challenge and Opportunities

Moderator: Dr. Vinay Pai, NIBIB 351
Keynote: Dr. Jonathan White, AHRQ 351
Panel Discussion

P R O C E E D I N G S

Welcome and Objectives

DR. SETO: Good morning, everyone. Welcome back to day two of Natural Language Processing Workshop. I am Belinda Seto. I am the Deputy Director of the National Institute of Biomedical Imaging and Bioengineering.

I have found this to be a fantastic experience, not only for me in terms of learning, but also in terms of the partnership with the National Library of Medicine, which has always been one of my favorite institutes, but this experience actually reinforce how great a partner they are to us.

Yesterday, we learned, at least I found it educational, it may be old news to many of you, I learned a lot about the statistical approach, the linguistic approach, and the very nicely drawn pendulum that maybe we would end up somewhere in the middle with a hybrid approach.

Today, we are going to hear talks about applying natural language processing to extract information from observed health data including electronic medical record, and apply that to clinical decision support.

Now, I recognize that in the latter area in clinical research support, it is still a very active area of research that debates whether we should do this based on evidence from literature, based on guidelines, how might we use natural language processing to help us decide the test results, and so on.

So, I look forward to vibrant, active, heated discussion even. So, at this point, it is my pleasure to introduce Dr. Blackford Middleton, who will give an overview this morning.

Dr. Middleton is the corporate director of Clinical Informatics Research and Development at the Partners Healthcare in Boston. He is assistant professor of medicine at Brigham and Woman, one of the best hospitals there. I have a bias, my daughter is there.

He is also a lecturer of health policy and management at the Harvard School of Public

Health. He has a very, very distinguished CV, and I am really abbreviating this. He studied biochemistry and molecular biology at the University of Colorado at Boulder. He received his master's degree in Public Health from Yale University of Public Health with a dual concentration in epidemiology, Health Services Administration.

He received his M.D. from SUNY in Buffalo, and was a resident in internal medicine at the University of Connecticut Health Center.

So, you can see he is preeminently qualified to give the overview for the morning.

Dr. Middleton.

DR. MIDDLETON: Thank you very much.

[Applause.]

Overview of CDS in Healthcare

DR. MIDDLETON: Thanks very much. Good morning and thank you kindly for that introduction, Dr. Seto. It is a pleasure to be here and to have worked with James Luo and the team from NIVIB to help put together this day, which I think will be really an outstanding overview of clinical decision support, current practice, research questions, and future directions and complements so nicely what we heard yesterday in the NLP day.

It really is an extraordinary opportunity I think to bring these disciplines even closer together. There certainly are many relations already between NLP and the CDS, and we will see examples of that throughout the day today in this sort of applications tract, if you will.

But my job at the beginning here is to talk about CDS from the highest level and give an orientation or overview to what is clinical decision support in clinical practice, what is it today, what is the evidence base suggesting the utility or impact of CDS, and both evidence for and against, because it is important to realize we are still evaluating the impact of CDS in many ways, and maybe give some pointers to what will CDS be in the future, and close with some of the research challenges and questions as I see them.

I think the first thing to remember, though, about decision support is simply accessing the data and visualizing the patient record is probably the most important thing we do with electronic health records or paper-based records.

This, of course, is the famous diagram from Minard, who plotted the progress of Napoleon and the Army in 1812 toward Moscow and the subsequent retreat, of course, when he got to Moscow and found it burned to the ground.

In this single diagram, Minard shows six different variables simultaneously. We can see latitude and longitude. We can see temporal progression. We can see size of the Army both in advance and in retreat, and there is one more I am forgetting, but the way stations along the way, geographic information, as well, and time.

So, you know, this information I think put together like this gives obviously as Tufte would say, you know, "A beautiful and elegant insight into the course of this march," and summarizes a lot of data in very interesting ways.

This is the first step we have in clinical decision support, is to simply gather, aggregate, and display information, and it is oftentimes forgot, I think when we talk about clinical decision support, but it is extremely important. I am going to come back to this at the very end.

In the U.S., why do we need clinical decision support in the current day, and in Washington, maybe this is all well-known and that the institute is well understood, but I think it is important just to drive home these issues that we still have patient data unavailable in 81 percent of cases, as Paul Tang found in his very famous study.

Lucian Leape found in the Harvard Medical Outcomes Practice Study 18 percent of medical errors are due to inadequate information availability, and Medicare beneficiaries in this country, those over 65, see on average 6.4

different clinicians. It is a fractured and unwired delivery system.

Many have written about the delayed translation of knowledge to practice. Marc Overhage and others, Andrew Balas estimated when you net it all out, it might take 17 years before a new innovation is actually brought into clinical practice routinely.

We know from a variety of careful studies in both urban and rural, and academic and community-based practitioners that information needs are often not being met at the bedside or in the clinical ambulatory care setting.

Bill Stead and others have written about the cognitive demands processing the information explosion that we are now confronted with. This chart shows the sort of well-known saw that we can handle, five items or seven perhaps on average on a good day. The human cognitive capacity mapped against all the other information bits that are coming at us, decisions on the clinical phenotype, that is, the patient's characteristics physically and what not, structural genetics, the SNPs and haplotypes that might be relevant, functional genetics, the gene expression profiles, and proteomics and any other omics that might apply to this decision we are needing to make at this point of time for this patient. Regrettably, our decisions, though, don't always result, in fact, only about half the time result in applying the best evidence to practice.

Elizabeth McGlynn did the famous study looking across the country at practice patterns, surveying physicians and examining the medical records of their patients to find these kinds of delivery patterns where compliance with idealized or best practices, evidence-based guidelines, and the like, went from 64.22 to 14.21 percent regrettably.

In the take away, the headline was for this article, "On average, patients receive 54.9 percent of recommended care," the so-called coin toss problem.

I think we still have the Flexner report to wrestle with 100 years after Abraham Flexner wrote it, "Society reaps at this moment but a small fraction of the advantage which current knowledge has the power to confer."

We have such a great knowledge base. We have excellent practitioners who are doing their best to deal with the exploding knowledge base and exploding dataset, but we do need to do more to bring them together.

If we did bring them together, a variety of studies we did at the CITL looked at using the evidence base, synthesizing best practices and expert opinion to estimate how much highly informed and interoperable clinical medicine would be worth, what is the value proposition if you will.

A variety of studies here looking at evidence for ambulatory computerized provider order entry, health information exchange, interoperability, chronic diabetes management, tele-healthcare, physician-physician, and personal health records.

We find across all these studies about 8 years worth of work of a team of 10, that about $150 billion could be saved if all this technology were adopted and used routinely. This might be an idealized estimate in some eyes, but we think actually it is fairly conservative and others have suggested it might actually be more like 30 percent including all of the administrative waste, as well.

So, where are? We are in the midst of this era with our fearless leader. No matter what you think about the politics of the issue, health care is in the midst of reform, which is just desperately needed to sort of correct the course, bend the curve, if you will, so that we don't have this inexorable rise in health care costs that simply is not sustainable.

It is too big to fail. I don't know how we need to get everyone behind this idea, but health care reform needs to be addressed, so that we can bend the curve of costs and try to do what is right for patients, and more of what is right and less of what is wrong, and get better outcomes at the same time, the triple A.

The health care reform effort, at least the ONC stimulus package from RO's high tech bill aims to both stimulate the adoption of health care IT, stimulate interoperability of that IT, and to then inform clinical practice with clinical decision support and achieve this transformation of health care.

A four-year time frame is pretty quick. Many have criticized this might take a generation or two, and the implementation I would suggest of health care IT is occurring without the essential ingredient of the knowledge base.

Knowledge is not resident in electronic medical records when implemented. It has to be added as a component either from the customer, the implementer, or from knowledge vendors, and the like.

So, I think actually we are in a perfect storm for CDS. Many saw this movie of the fishermen and lobstermen off the coast of Gloucester, Mass., the famous story about the perfect storm, I feel sometimes like a member of the crew on that boat in health care, but we have lots of clinical data going on line, increased standardization of data, increased interoperability gradually coming.

A tsunami of genetic data and personal and social data, increasingly geospatial data will be relevant to our clinical decision support, and I mentioned already this inexorable rise in the need for health care reform.

When we think about decision support, let's think first about how physicians reason. There is many, many different ways to look at the hypothetical deductive process that physicians employ to analyze a patient, assess the history and physical, assess laboratory findings, and then make a differential diagnosis or a list of probable causes of disease or less than wellness.

Hal Sox and David Eddy, among others, have written I think most persuasively about this. Physicians are extraordinary pattern recognizers, listening and generating hypotheses, cross-examining to validate or dispute, evaluating these hypotheses, and then taking action.

We know, though, that this decisionmaking process is subject to a number of important biases.

First, hypotheses are generated very, very early, sometimes in seconds upon entering a room or visiting a patient's bedside, and just a few active hypotheses are considered at any one time, but these biases are well understood now and well known, and I point you toward Daniel Kahneman's book, "Thinking Fast and Slow," which is really a very accessible treatise on some of his pioneering work with Tversky that really elucidated some of these findings.

I won't go into these in great detail now, I don't have time, but the representativeness heuristic, we can be biased by our mistaken assumptions around prior probabilities and inappropriately weighting dependent factors or independent factors both.

The availability heuristic, what did we see lately that kind of looks like this case, and then anchoring and adjustment heuristics. Physicians are well known to anchor sort of their perceptions and then have difficulty adjusting.

The bottom line from a decision theoretic point of view, what we aim to do is to make decisions that distinguish these three zones: the zones of don't treat, don't intervene perhaps, or test or intervene, and finally, the last zone of treat where the prior probability, test performance characteristics, and clinical intervention efficacy whether it's drug or procedure are all weighed in the balance of a nice decision tree that can distinguish these three zones, and they really help us determine which zone we are in and what actions to take or not.

We never get to do this at the bedside, of course. It is too complicated, the probabilities aren't known, and it is not applied in practice, so it is always an intuition, but this might be viewed as the goal.

So, Chuck Friedman wrote this nice paper on the Fundamental Theorem of Biomedical Informatices, and it is pretty simple. Brain plus computer is greater than brain. I hope that is true.

I am not sure always, but hopefully, that is true, and what we are really trying to do, I think, with decision support, is understand both the deficiencies of our clinical reasoning, understand the deficiencies of our health care processes and the delivery system, and then fill the gaps, fill the care gaps, fill the reasoning gaps, fill the process gaps. That partly can be done with the computer's assistance.

I like to recall Marston Stop Blois' diagram, though, from 1980, a long time ago, who wrote about the cognitive funnel and wherein is clinical reasoning, wherein is the physician's reasoning, most important, moving from left to right, if you will, in a cognitive funnel, we operate actually very well at the point of A, and not so well at the point of B.

A is where uncertainty is rampant, there aren't clear guidelines or the evidence may be conflicting, and machines can operate exquisitely well at B, where the problem is well defined and constrained, and can be really supported or is amenable to computation or symbolic reasoning of any kind.

I am humbled also, though, by where CDS is limited. In my own fellowship at Stanford, focusing on health services research and clinical informatics, I was a member of a team that built a Bayesian belief network for differential diagnosis, and my wife, bless her heart, halfway through the project found this cartoon and made sure I saw it, because we were getting results like this.

We would type in a large number of signs and symptoms and case findings into the QMR decision theoretic program, a Bayesian belief net of approximately 700 disease conditions and 40,000 findings, and all the probabilistic connections between diseases and findings were all characterized. We would get just sort of unintuitive, noninformative differentials all over the map.

So, creating decision support requires creation of this knowledge base. Duter and Shortliffe described a knowledge base as an AI program whose performance depends more on the explicit presence of a large body of knowledge than on the presence of genius computational procedures.

The point here is it is really about the knowledge. This came up several times yesterday. Small doses of knowledge in NLP can really inform the statistical processes and optimize them in ways that any amount of computation may never get to.

Creating these knowledge bases, though, is difficult, and I am going to come back to that. The knowledge base, though, can be subjected to any kind of computational process or symbolic reasoning, algorithmic, and you can run through the list, but the challenge with a good knowledge base is that we actually want to be able to use it in multiple different inference engines or different systems.

This idea of shareability of the knowledge base is central to a lot of the work over the years, but frankly, I don't think has ever really been achieved.

The last thing I will say about knowledge bases before diving further into CDS is even with all the best evidence we have, Brent James, among others, would say, "We only have evidence for about 25 percent of what we do, the rest of it, 75 percent, isn't supported by a particular guideline or even combination of guidelines, and it is really the clinician reasoning across uncertainty where there is not a solid guideline or solid evidence to pursue.

So, I would suggest we actually are on the verge of a dramatic need for not just knowledge base clinical decision support, but decision support which arises from analysis in real time of patients like mine, or decisions like mine.

The patients like me phenomenon has been extraordinary. Jamie Haywood and the team allow patients to provide information on line and get a lot of analysis and comparison and contrasts in other like patients, why don't we do the same thing for the provider.

So, at the point of care, if there is not a piece of evidence or a guideline which might apply, I might be able to say, you know, show me what the previous thousand patients at the Brigham have done in a similar situation, or the thousand physicians caring for those patients.

So, our knowledge translation specification research history over the years has moved from very early efforts at Stanford with Ted Shortliffe and Mark Musen and others through a variety of plan based and sort of nonplan based, if you will, representations of knowledge to try to get at this notion of interoperable knowledge base, knowledge which can be shared and communicated and then subsequently executed upon.

Our goal, of course, is to combine evidence and experience, make it into a guideline in a principled and unambiguous way, represents the knowledge in the guidelines in an unambiguous way, and then finally, share that knowledge and subsequently execute upon it in any different inference engine or receiving environment for the knowledge.

These knowledge bases, though, that we have built over the years, things like Mycin for the antimicrobial therapy in meningitis at Stanford, or QMR and Internist I for differential diagnosis in medicine, in a way, they have actually been very constrained to a particular decision type.

When you think about the actual knowledge stack that has to go into modern day clinical decision support, it has to include a wide array of components that oftentimes are not considered together, ranging from the terminology services, what is the controlled medical terminology, the underlying information model in otology, and then subsequently into progressive layers of abstraction that are used in creating these roles, and frankly, are used also to simplify the knowledge management process, problem lists, intermediate concepts whether it is drug classes or patient states, order catalogs and other information classes, and then Standard App templates, order sets, rules, and the like.

So, where is CDS today? The hospitals are riddled with CDS, our automobiles are riddled with CDS. The Smart Phone is riddled with CDS now. But Cooperman's review around 2000, I think, described these different classes, formatting, interpreting, consulting, monitoring, critiquing, and I would add, you know, this new class now of Smart Apps for the consumer where a lot of what were clinical decisions are being now made by consumers themselves, a real disruption, if you will, from a Clay Christensen point of view where the consumer is now being empowered to do what previously required professional expertise.

So, what is the evidence for CDS? Garg did a very famous review, revealing 97 studies finding practitioner performance improved in 64 percent of studies, 40 percent in 10 diagnostic systems, 76 percent in 21 reminder systems, and 66 percent in 29 drug dosing or prescribing systems.

Patient outcomes were only studied in 7 of 52 studies, and importantly, factors associated with success were automated prompts versus anything that required the user to do something else, and also the bias perhaps of studying one's own child when the authors were involved in evaluating these systems.

A lot of work we have done at the Brigham and elsewhere suggest, though, that CDS can increase guideline adherence at the point of care. We can improve order entry with CPOE-related prompts, we can decrease radiologic test ordering with showing the costs of test and procedures, and we can, like the Amazon checkout charge display, decrease the number of tests ordered when showing the cost, as well.

Problems with HIT have been more apparent, though, of late, as well. Ross Koppel, I think did the first interesting work in this space looking at how CPOE can facilitate prescription error, and he asked users of a new CPOE system in an academic medical center on the East Coast what were their perceptions of implementation of this HIT, and 22 categories of perceived increased risk were described, both the information errors and errors associated with a poor human computer interface or workflow.

This really lit the fuse, I think, though, under how we have to view HIT and CDS specifically, much more critically and establish a dose response curve, if you will, for HIT, how much of a dose of HIT is useful versus how much is problematic.

Alert fatigue is a great example that

usually comes to mind, how many alerts and the physician starts to turn off and ignore everything, not enough alerts, and the physician may not be paying enough attention, so what is the right amount of alerts is really of interest.

Perhaps some learnings can be borrowed from economics and behavioral theory, you know, thinking about the theory of subjective nullity, for example, do we entertain and do we inform enough to make the decision support interesting to the clinician.

Recently, at Duke, evidence-based report was completed by David Lobach and Ken Kawamoto and others, and just the nutshell here is that to underscore the idea of the automatic provision of decision support at the time and location of decisionmaking, making a recommendation that is actionable, and then integrating with the workflow and promoting action, no need for additional data entry, and make sure the user is involved in the local knowledge development process.

So, thinking about the decision support research, Chaudry found that 25 percent of all this decision support research that has been done has come from just four different institutions: Regenstrief, Brigham, VA, and Intermountain.

So, I would like to borrow from William F. Gibbs and his quote, "The future is already here, it is just not evenly distributed." We have to think about how to distribute this knowledge that CDS, wherever it is, can work well.

Let me take a small detour now and just talk about the other things I think impinging upon or encroaching upon decision support. Number one, of course, is just big data. How do we take advantage of the quantified self?

As our patients and ourselves are increasingly quantified with all types of ongoing measurements in the intricacies of care, how do we take advantage of that in the record and the body monitor, the Smart phones and Apps, behavioral monitors, the other types of biosignal monitors, BPs, et cetera, because the patient is becoming activated, and one of the studies we did suggested that if we have an activated patient and actually can activate the provider as well.

The next one, of course, is personalized medicine. We are now beginning to sequence the entire genome for 100 patients at the Brigham. Interestingly, in today's USA Today, how convenient it is when the USA Today provides a prop, but, you know, how much do you want to know? When you sequence your gene, there is things that you may not really want to know, and the provider really doesn't know what to do with the information anyway.

So, in addition to all the GWAS studies associating these polymorphisms and SNPS and whatnot with phenotypic disease, we need to assess the behavioral characteristics of what people want to know and how much they want to know.

Zak Kohane has written about the incident, the problem with sequencing the entire genome is it is simply the too many tests problem. If we have a sensitivity of 100 percent and a false positive rate of only 0.01 with 10,000 tests, 60 percent of the results will be false positive, so what should we tell the patient and what will patients want to know.

He has a lovely graphic in his Science and Translational Medicine article about the different dimensions and axes of patient preferences and whatnot that will have to be addressed before we can use, in a decisioned theoretic way, you know, provide answers that matter to the patient and matter to the provider, that have utility.

Let me close now just with a few words about some current research that we are doing at Partners with a large number of collaborators across the country that is called the CDS Consortium, funded by the AHRQ, and has the goal of accelerating the translation of knowledge into practice, disseminating knowledge, and then evaluating its impact.

We have a large number of partners. We have been very fortunate to have Regenstrief and UMDNJ, Frank Sonnenberg here, a number of technology partners and international partners, as well, collaborating in this work.

We had a few simple goals. One was to take the knowledge base developed at Partners over the 60-odd years that clinical information technology in the systems have been in use, and to try to generalize that and externalize it for the world to use.

So, we have created a prototype national knowledge repository, a knowledge management portal where one can go and browse the knowledge artifacts used in our clinical decision support systems and examine them and see if they are of interest for your own use for free.

So, we have tagged all of this knowledge, developed a model, and hosted it now, a sample of it I should say in this external knowledge portal.

The CDS consortium then serves up the

knowledge artifacts in the three domains we have had under study for the AHRQ: CAD, diabetes, and hypertension, and now serve up a knowledge transaction, if you will, with remote sites across the country - Regenstrief, Indianapolis, UMDNJ with GE, imminently with Dr. Sonnenberg and Dr. Gregg Frasier at a next-gen clinic in Mid Valley IPA, Oregon, but most importantly Regenstrief, which is now going housewide at the Wishard with the CDSC decision support services we have been providing.

At the same time, a related project PECARN which is a pediatric research collaborative, is using the CDSC services to provide traumatic brain injury decision support to over half a dozen sites now I think, all of which are using ethic EMR technology.

So, in closing, what are the grand challenges I think we have to address, summarizing patient level information, visualizing this information, prioritizing recommendations, merging guidelines with competing recommendations, disambiguating guideline content, improving the human computer interface, managing these large clinical knowledge bases, accessing and sharing knowledge in executable form, dissemination, and the list you can see here from my paper with Dean Sittig.

I put it together in another way for the bioinformaticians in the room to sort of draw upon a biological analogy. I think in many ways this is like the fundamental theorem of biology.

We need to translate or transcribe rather atomic knowledge objects built upon cognitive and behavioral foundations into patient-centered data abstractions and knowledge, representation or knowledge engineering.

We then need to translate that into decision proteins or essential codes and structures and architectures for supporting a reference architecture for inference.

Then, we need to implement and assess the effective use of personalized decision support at the point of care. Each one of these processes has a feedback loop, an evaluation loop, which has to be considered as well.

Hopefully, the NLM and IBIB, NSF, AHRQ, anyone else who is interested would like to fund this ongoing work.

The last slide then. I just want to return to Napoleon, who got to Moscow, found a city, was burned to the ground, and retreated all the way back to the greener fields of Paris and France, where he could replenish his army, but in many ways, I think, you know, the emperor has had no clothes in U.S. health care, and we need to think about how to reclothe the physician emperor.

I think CDS is the essential ingredient. We cannot process the explosion of knowledge and data without CDS, and knowledge sharing is the only way to scale this across our country and perhaps around the world.

I would like to see us not only share data, but share knowledge, as well, seamlessly across the land, and that's it.

Thank you very much.

[Applause.]

DR. MIDDLETON: Questions?

DR. HERSCHFELD: Does anyone have any quick questions for Blackford?

Jim, what do you think?

DR. WALKER: Jim Walker. I think your presentation was great, Blackford. I mean it raises all kinds of questions for me about process engineering, about the role of the care delivery organization whether it's a local sort of traditional organization or increasingly some kind of accountable care arrangement.

You know, it seems to us that a key part of this is getting commitment organizations and individuals, patients, and the rest of their care team, to 100 percent processes, and if you get that, then, it seems to me what you are doing becomes critical infrastructure.

Without that, it is not clear to me how it works out, but that is a comment, not a question. Is that any part of this, you know, part of the knowledge base, to understand what the levers are for patients, low locus of control patients, patients with limited educations, patients with modest interest in wellness, in some future state, how we engage those people in receiving this information.

DR. MIDDLETON: Excellent questions and a great point to bring this back to reality, how do we tune the organization or frame the problem for the organization, and how do we make this thing generalizable to those who even are in health care, the system, and those who are not.

One of the things I am encouraged about is just the plethora now of Smart Phone absent clinical acts arising in the consumer's hands where the very same knowledge will need to be applied, but as you say, tuned for the low activated patient or made appropriately interpretable, et cetera, and I think that is an excellent point and part of the implementation challenge.

DR. MIDDLETON: I guess what I would

suggest, you know, when you think about the EKG interpretation, the ABG interpretation, PFT in the hospital laboratories, everything comes with an interpretation, and that is now decision support.

If you look at the diagnostic differential, I agree with you that has been a challenge, because I think the problem is hard, but on the therapy side, the planning, radiation oncology, and other kinds of planning algorithm programs are used routinely.

So, we have made progress, but of course, Watson is going to solve this problem. We will hear about that later today.

Thank you very much.

[Applause.]

DR. CONROY: Thank you, Blackford.

Panel 1: Clinical Perspectives

DR. CONROY: My name is Richard Conroy. I am program director at NIBIB and we are going to start the first panel of the meeting today. It is on Clinical Perspectives.

When we were planning this meeting, what we were thinking about was we would like to hear kind of what the state of the art is for clinical decision support from the clinicians themselves, the people here using these systems, what the challenges of ACR, where they see opportunities for future research, but also where they would see systems helping them with their clinical practices.

We could have a whole day meeting just on this topic, but we have got four great speakers, and I am not going to give a long introduction for each of them, because I know they have lots of great things to say, but we have Dr. Thomas Payne from the University of Washington, Dr. Frank Sonnenberg from the University of Medicine and Dentistry of New Jersey, James Walker from Geisinger Health Care System, and Eliot Siegel from the University of Maryland.

I am going to ask each of them to come up. We will take questions after each of the talks, and then at the end we will a half-hour panel discussion, so if you have general questions, please save them for the end; if you have got specific questions for each of the speakers, please ask them at the end of each talk.

Thanks.

Keynote:

Dr. Thomas Payne, University of Washington

DR. PAYNE: Good morning. I am the first of the panel, and I am Tom Payne from the University of Washington at Seattle.

My topic today is an attempt to blink some of what we heard yesterday into the world in which I work, and which many American clinicians work in the clinical setting, and to kind of show how that relates to the topic that Blackford so ably reviewed, which is clinical decision support.

I think a perspective that the group here can add to what we discussed yesterday is the workflow, the milieu into which all of this fits. That was something I wanted to add to the answers given to our questioner yesterday, who asked what does it take to take what we have learned about NLP and to impact the health of people in the United States.

There was a list of answers to that question. One more would be a recognition that the workflow in which all these things applied is a critical component.

So, I am going to give you a little story about my experience with that workflow in NLP. I want to tell you first how we got from paper to electronic notes, which is the grist for the NLP mill.

I want to show you how this relates to some potential and clinical decision support. I will give you three examples of how we are using natural language processing today in very early ways in the University of Washington in Seattle, and then, hopefully, we will have a few remaining minutes for discussion.

I work in a hospital system in Seattle. We have five hospitals. We have about 1 1/2 million outpatient visits a year and 60,000 discharges from a hospital, so it is a moderately sized academic medical center in the northwest.

This is where we started, which is the state of many American hospitals certainly 30 years ago and today in many hospitals, and our first step on this journey to electronics was you put that into an electronic medical record, which has some advantages, with a few disadvantages, as well.

So, the next step in this journey was to convert what clinicians documented in their care of patients, from that to this, which is an improvement, it is legible, it is available to many places at the same time, and it is useful for many other purposes beyond the immediate patient care purpose for which the notes were created.

It also is valuable to others who measure quality, who help us with compliance issues, and so on, but it has its own set of limitations.

We started our trip from paper to

electronics about eight years ago, and about six years ago we had largely finished this conversion from paper to electronic notes, so since then, we have thousands of physician notes written each day for the hospital and in the clinics, and that transition has included all other disciplines now from nursing notes, bedside nursing notes, to nutrition, physical therapy, and, interestingly, notes written by hospital chaplains, as well. It is all electronic, it is all in one place, a real advance for us.

What we found is that narrative text, the content of the note that I just showed you, we regarded to be very valuable, and some of the ways in which it is valuable I list here. First of all, the narrative contains the history and its details, the exam, and the thinking of the clinician, terse though it may be in the text.

I have to say that this is one area in which the needs of clinical care, the needs of our compliance reimbursement crew, and the teaching and research environments overlap, and that if we capture the thinking of the clinician, we can accomplish a lot for each of those groups, and each of these thousands, maybe 5,000 notes written by physicians each day contain some small, and maybe more than small, kernels of truth.

So, these findings, these episodes, or these components of the story that the vision tells are in that note. Now, it is not easy to find, it is obscured by templates, by the artifacts that we use to put the note into the record, such as copying and pasting, direct entry aids and so on, but it is there, and we have the ability to make it easier to find this, as well.

If you multiply that single note times the number of encounters that we have in our institution, and the number of hospitals in our community, the number of communities in our country, there is an awful lot of information there lurking within the narrative text.

There has been discussion over the last certainly the 25 years I have been in this field, the tension between use of narrative text and use of encoded or structured notes. By the way, Octo Barnett taught me years ago to call it narrative text, and not free test. He said, "Nothing is free about free text, somebody pays for it."

So, the narrative text contains information that we would like to get out. We have tools for structured note entry, as well, and this is a quick view slides of what they look like.

So, you can see that there are constrained choices from which the clinician chooses. I have illustrated there in a very small box that there is a choice that I can make to indicate whether there is cerumen in both ears, in the left ear or the right ear, and when I make that choice, it is stored as an encoded element. That is from one of our vendors.

Here is the same idea implemented by a different vendor, so these are alternatives to narrative text. You can use structured tools, and these are very powerful, developed with the idea that structured or encoded text is really the foundation on which we build clinical decision support systems.

However, what we have learned in our journey is that there are a lot of problems with structured note entry. The first is that it is harder to train people to use them. As a teaching institution, we have turnover every month with our trainees, our fellows, our junior faculty, and so on. We have five hospitals not including the childrens and the VA, so people move from one to another.

As they move, they have to learn how to use structured note tools, and that burden is considerable to us. In our experience, most physicians don't like to use these to write their notes, they prefer alternatives.

They will do it, and some ultimately choose this, but it is not the majority.

When you are done writing a note with these tools, most physicians aren't too happy about reading them, and they focus on the narrative text portion, the assessment or the very terse history at the beginning.

In creating these -- we heard a little about this yesterday -- that if you use these tools, you lose some very important detail that is contained in a richer narrative, and so for these reasons, we found that clinicians are moving more to narrative text.

So, that is why I view a topic such as natural language processing as so important, and so important for many reasons including clinical decision support.

Here are some, a quick walk through a single note, and that will help illustrate why I believe this is important, but before I show you that note, I would like to bring up the point that Gordon Schiff and David Bates made in The New England Journal about the problems that we face in clinical medicine today and how clinical documentation might help.

Their observation, which I find to be very accurate, is that the problem of having too much information is now surpassing that of having too little. When I was an intern, 20 percent of the time when I went to the clinic, there would be no chart. That isn't the problem anymore.

Now, I have a chart that is too large for me to absorb, and that happens in every encounter in which I enter the room unless I know the patient well. So, this is another need for help, and I view this an important element of clinical decision support.

So, let's see how that might come to bear on a typical note. For example, this narrative text note includes a list of medications that was entered by the intern. Is the reconciliation of the medications accurate? Can we help with that problem?

In the physical exam, vital signs and other findings are noted. Do these hide clues to an early sign of sepsis, which, if recognized, could help alter the course of that sepsis?

The imaging reports are listed in a summary that is useful for humans to read. What other recommendations lie in those imaging reports that are relevant to the care of that patient?

Here is the formulation of the thinking of that physician and their efforts to organize the problems. Should there be a broader differential? Is the care outlined there appropriate? The code status, is that accurate, is that up to date? All of these are areas in which we can help those clinicians who may have 15 such admissions in one night.

Lastly, for colleagues in CMS, are the compliance rules being followed? We are very attentive to this as are most institutions, and we spend a lot of time in trying to improve our compliance.

One way again to do this was to use structured notes. This is one of the advantages of using structured notes that you can quickly track the care given to people with, say, diabetes mellitus with encoded elements for foot exam, and so on, but as I mentioned, this is less commonly used in our experience than the narrative text report, and it is not just us.

When we looked at the core measures that are used to measure the quality of the care we deliver, that were gathered together by the UHC, what we found was that a lot of the information from the core measures is found in structured encoded information in our system, but most of it is not.

Most of the information that we use to measure quality lies in the narrative text of the note, and not just from one source, but from many sources, so there is where we need to direct our focus.

If we look at a broader perspective on the same issue, in this paper from Roth and her colleagues, we find that a lot of the quality measures that we seek to apply to the care of patients are hard to get to, and as you can see, on the righthand side of this graph, some of the things that are hard to get to are disease-specific history, physical exam, patient education, social history, and so on.

Those are very important to measuring quality, and yet they are largely not easy to get to, because they are in narrative text. So, I see a host of reasons that natural language processing can be a big help to us.

With that backdrop, the story of our transition from paper to electronic notes, our experiments using structured and unstructured, the advantages of both, I would like to show you some examples of what we have applied the field of natural language processing to in our production environment.

Now, a brief aside here to point out that as we take a great idea in computing, it has to pass through several hurdles including the rigorous testing, but the hurdles that I am familiar with in my day-to-day work as basically the CMI over our system, in order to get it into a production system it has to be, first of all, industrial strength, it has to have an extremely high performance or clinicians won't tolerate it, it has to compete with lots of other applications, and it has to fit into the workflow of the busy clinician.

If it doesn't meet those standards, it remains in the laboratory. Now, we have passed those hurdles in two areas, and we are working on the 3rd, and I will show you those examples today, and these added to my enthusiasm for the field of natural language processing.

First, complying with the law. Evaluation and management codes are difficult to assign, and for those of you who aren't familiar with this, I will give you a brief view of a few of the pages of the book that covers how you assign an E&M code to your note. By the way, this is how you are paid, so it is important that you master this.

So, you can see that this book, a few pages here only is not easy, and so it is not surprising that physicians have a challenge complying with these.

So, we thought what a great idea for the use of natural language processing, and just as we heard yesterday, computer system coding was appealing to us because of the complexity of these rules and because we also have the notes and machinery that will form.

So, our electronic medical record, which is provided to us by a vendor, contains a set of tools to process the narrative text, tag the documents with SNOMED codes, apply algorithms that are pertinent to assigning an E&M code, and you then have a very reasonable estimate of the E&M code supported by the document you have just signed.

The tools, handle negation qualifiers, we can add rules to increase its precision, and interestingly and importantly, this gives feedback to the provider on every note they sign within three seconds, so it fits into the workflow of the physician.

It is calibrated to meet the standards of our compliance officer, and after calibration, 93 out of 100 notes were given the same code as a team of professional coders assigned to that same note.

I will not tell you the precise figure that the clinicians had in assigning codes to the same notes, but it was substantially lower.

So, this is very encouraging to us that it gives physicians something they very much value, which is a safety net to complying with the law, and it also has corollary benefits.

This is how it actually fits into the work flow, which is sort of a model for how NLP and decision support applications might be used. First of all, we aren't particular as to how that note came to be, whether it was dictated, typed to template, or voice recognition software used.

When it is signed, then a fee sheet is completed. That signed note goes through the electronic medical record into the places that all notes reside, but a note is also sent to this NOP engine, which is used to derive the E&M code supported by the document, but also the codes that were pulled from that note. The SNOMED CT codes are stored and are also available for use for other purposes, and I will show you one of those other purposes shortly.

A quick screen print here shows that on the right side of the screen, you see the history of that note, and I have highlighted AML, AML is tagged, it corresponds to leukemia, you have a SNOMED CT code, and every phrase that is identified also has those. Again, this analysis is performed within three seconds and fed back to the physician.

The full note looks like this, and those notes are red for negation, green for some form of probability, and blue for positive. So, that is an example of a phrase identified, and the code that is assigned by the software is shown here, the one assigned by the physician is shown here, so they can understand where their estimation varies from what the system offers. This is a great educational experience to them and I think a good model for helping us do a better job.

So, that is one example, computer assisted coding. A second is that our notes look in many cases like the one I am showing you here, and this is a narrative text note that includes a detailed problem list written in terms that are familiar to the physician.

The problem is that none of the systems that Blackford referred to can do much with this until it is encoded, so we have taken that very physician friendly problem, which you can see in this note, and adapted the software system that I showed you earlier for computer assisted coding and help pull the problem list from that note.

So, you can see here a list of notes on the left. It analyzes the note, and it identifies a set of diagnoses that are mentioned by the physician in the note including some that are a little bit more challenging, for example, L, a Grade 3 renal laceration, T11, T12 anterior column compression fractures, are sacral fracture, follow those abbreviations that we heard so much about yesterday, and it does a very good job of pulling out SNOMED CT codes represented in that note for the physician to review and add to the problem list, which they can do with a single click.

So, this may seem like a small advance and I guess in a scale of things it is, but for those physicians finishing their note, doing their job for documenting that care and having the problem list automated and encoded as a byproduct of their work is very popular to them, so they don't have to go to another dialog box, interrupt their workflow to assign a problem list, very popular.

A third example, and this one has not yet achieved the standard that I mentioned for you of being available in production with performance standards that clinicians expect, and this is work I am doing with Meliha Yetisgen-Yildiz in our biomedical informatics group.

It tackles the unfortunate problem that we face, which is that imaging reports are so extensive, I am not referring to the images themselves, but the reports we get back, that within the reports are more information than we can easily process or remember.

This example shows that in the impression, this trauma victim has many serious problems including fractures, pneumothoraces, hematoma in the abdomen, a left hemothorax, and that is enough to get your attention.

So, you have lots to do to care for this patient, and you might overlook the fact that there is also a right adnexal cyst, which the radiologist recommends that we remember and follow up to avoid possible ovarian neoplasm developing.

So, all of this information is in narrative text. What can we do to help with that? Our project is looking not for the critical abnormalities, but for those that are subcritical, flagging them, so that they are available for the next person to review in the clinic, perhaps when they have recovered from their accident.

So, voice is catching on. Voice and speech technologies are now mainstream. This is noticed by physicians, they are using this extensively.

I think the things that we can use as we take the expressiveness of the physician and put it into text is that we have the ability to summarize, to search more easily, to extract key and coded information, such as problem list, we can focus our attention on things that might be overlooked, and so as the volume of narrative text grows, so does our need for the tools that we heard about described yesterday.

I think the trend toward narrative growth will continue and that NLP will help us use that to make better decisions. It fits into the workflow of electronic medical records, and as we implement medical records more extensively across our country, we need to make sure that we match those EMRs with human strength and with workflow.

I will stop there.

[Applause.]

DR. HIRSCHFELD: We have time for one or two questions for Tom.

DR. PAYNE: Steve.

DR. MYESTRE: University of Utah. I have a specific question about your implementation of NLP to encodes the narrative problem waste. Do you have some performance numbers or did you evaluate that? How well did it perform?

DR. PAYNE: For the clinician, the most important measure is how fast it is, and it is seconds, so it is very fast. The accuracy is not as high as a problem list might be, but again it isn't automatically adding it to the problem, it is giving it to the clinician to make a judgment about adding or not.

All I have is early -- it has been in production only about six weeks. I have just early data, but it is highly positive and very fast.

DR. McDONALD: Clem McDonald. Along the same lines, I think it is a great idea, the project. Do you deal with specificity, that is, the physician might say depression, and the problem lists says relenting, unremitting, blah-blah depression? I mean does it recognize that there is a more general or a little more specific one in the real problem list?

DR. PAYNE: It attempts to, and it does this by looking for existing problems with the same concept unique identifier as a clue that there might be overlap, and it will not propose them to be added. It actually shows the existent, and it doesn't propose to add them if it finds something close.

DR. RESNIK: Philip Resnik, University of Maryland.

Thank you for that. Given what I was talking about yesterday, it is incredibly exciting to hear this especially from the clinician perspective.

I guess my question is about where the state of things is right now. This is very exciting. Are you right now a lone voice in the wilderness? Are you starting to see what you are seeing here percolate out through the community?

In the computer-assisted coding world, you know, I have seen one perspective, but I am actually wondering whether right now you are leading edge, and nobody else is doing it, or you are part of a community of other people who are starting to explore this?

DR. PAYNE: I think the latter. I won't say it's widespread through the United States. I will say that with incentives for EMR use, the availability of electronic notes is rising quite rapidly.

There is, not a consensus, but a grumbling occurring across the country that we are losing some of the character of the notes, that point was made yesterday, and I am beginning to see the pendulum, since we are using that metaphor a lot today, swinging back to expressivity, but I can't quote you data on this. I am not a lone voice in the wilderness, but the group of people who are working on this is a relatively small band at the moment.

DR. RESNIK: We have to figure out how to get the rumbling louder.

DR. THOMPSON: Will Thompson, Northwestern University.

The application where you have the problem list being autogenerated, gives the clinician an opportunity to either accept or reject the proposals made by NLP, are you using that information to actually train the system and make it better?

DR. PAYNE: Good idea, but not yet, early days. I have to say there is an interesting anecdote. We are doing this at the same time we are implementing CPOE in three weeks and working on all the other meaningful use criteria, so it's an exciting time is one way to put it, but it also takes away your focus from all of the things that you would like to do. Our queue is getting bigger.

DR. SHANKAR: Shankar from Emory University.

I want do a structure note as well as a dictated note. The problem is if you go back after, say, five years, to do research, obviously, there are so many factors or elements missing to be a good research.

For example, the negative family histories are a very important thing if you are looking for any possible estimate, and you may not have it in the narrative text if you dictated it.

One way we addressed this was that if especially in an academic institution, you can have a structured note by some resident or someone, and then about the patient. I got an impression in a narrative, in a few sentences what they think about the patient, so in that way, you summarize a few sentences about the problem the patient is facing, so you have both a structured, as well as a free text note.

So, what do you think about that?

DR. PAYNE: I think we will have a mixture of narrative and encoded information.

The question is where does the boundary fall, and what I can report is there are clinicians that are pushing the boundary farther away from highly encoded and closer to more narrative text.

So, that box that we provide, we heard this yesterday, is being used a lot, and our requests are for templates that have less clicking and more typing of voice.

So, there will be a mixture. My sense is that people value both the creators and the readers, really value that narrative, so that pendulum is shifting more in that direction.

DR. HIRSCHFELD: Let's thank Tom again.

[Applause.]

DR. HIRSCHFELD: Our second speaker for this session, Dr. Frank Sonnenberg from the University of Medicine and Dentistry of New Jersey.

DR. SONNENBERG: Good morning. I want to thank Blackford for inviting me to this meeting. It has been very, very interesting so far.

I want to first explain the perspective

that I am speaking from. I am a general internist. I have been practicing for almost 30 years, more than 30 years if you include my residency.

I have an interest in applying this technology to my own patient care. I am Medical Director of Clinical Information Systems for a medical school and a large faculty, multispecialty group practice, so I have a responsibility for implementing clinical decision support, and I have come up against some very real practical concerns that I am going to illustrate for you in my talk.

I am also a co-investigator in the Clinical Decision Support Consortium, so I have taken an interest in the future of clinical decision support, and I am very well aware of the gap between where we are currently and where we hope it will be in the future, as Blackford presented this morning.

My conclusion in our current state of the art is that clinical decision support is still not readily available and it is not easy to implement with currently available tools.

I am just going to show you this is a fairly typical clinical decision support that is available. This is a commercial electronic medical record, the GE Centricity system which we use, and I would characterize the early decision support that is available as low-hanging fruit.

You can see here something called preventive care reminders. I have expanded that part of the screen, so you can see it in more detail. It is just a list of items that the EMR thinks are due for this patient, and they are characterized by things that can be, first of all, they are determined by very simple rules.

You know, the patient is a certain age, and they either have been done or they haven't been done, and the system can't distinguish between something that hasn't been done from something that maybe has been done and just hasn't been recorded in the right place.

When it comes to more complex decision

support, the system really has, at this point, relatively little to offer. So, I think that there are a number of problems that present challenges to clinical decision support currently.

First of all, the current generation of electronic health records are not designed to optimally organize patient information. They are still very heavily document based and episode based, so they don't capture very well some of the elements of the patient's data that are longitudinal and that carry over from one episode of care to the next.

Also, as Blackford pointed out this morning, patient tend to see a number of different providers, they are not always at the same institution, and so our total medical record is comprised of information that is derived from a number of different sources.

In particular, historical facts are not readily accessible. It is not easy to go in and see if a particular thing has occurred in the past, and certain kinds of data -- this is my own personal observation -- are not captured very well as discrete data in current EMRs.

One is symptoms. These tend to be in the narrative descriptions that we have been talking about.

Physical findings, they may be discretized, but in some cases they are discretized only in particular fields, so there may be a variable that says cardiovascular exam, and it may be a string of words, but it doesn't allow you to extract from that string what the actual findings were.

Previous procedures are not well documented as discrete data, and also measurements supporting diagnoses, I am going to illustrate some of these things in a moment.

In order to apply decision support of the type that we have been hearing about this morning where you are applying decision rules, trying to apply guidelines, you have to instantiate variables in order to decide how to execute those rules, and supplying variables automatically from the EMR requires capturing them during routine clinical care. They have to be represented somehow as discrete data. You have to identify what the pertinent variables are.

Another problem that we haven't discussed a lot at this meeting, but it is still a big problem, EMRs in general still do not employ standardized terminology.

That is beginning to change, but we are still not there yet. The system we use is a completely proprietary scheme, and, in fact, in our work as a demonstration site for the Clinical Decision Support Consortium, we have a whole subproject to translate our data into the standardized coding that the knowledge server requires. The bottom line is that much critical information is not captured in EMR variables.

I am going to walk you through a very detailed example that we analyzed as part of our work in implementing clinical guidelines. I am approaching this from the opposite perspective of what you heard yesterday. Yesterday, people were talking about looking at narrative text and asking what can we extract from it.

I am approaching it from the opposite point of view, which is what data do we need and where can we find it in the medical record.

I looked at two example guidelines. One is the JNC7, which is the standard and most highly accepted guideline for management of hypertension, and also the NCEP cholesterol treatment guidelines.

We selected these for two reasons. One is that they are very well accepted, they are not controversial at all, and they are both fairly complex. They are extraordinarily complex considering how commonly they need to be implemented.

So, we looked at the variables. We identified the following types of variables that are needed to implement these guidelines. First of all, they are simple variables, which I defined as single observations, something like "systolic blood pressure."

There are calculated variables, things that are not observed directly, but they are calculated from other observations. A simple example is "Age" that can be calculated from the birth date, or "Body mass index," which is calculated from the height and the weight.

Then, there are other complex variables that are defined in terms of other variables or observations, such as in these guidelines, terms such as "elevated coronary risk" or "metabolic syndrome" appear.

I have illustrated at the bottom, "metabolic syndrome" is defined in terms of several other terms, one of which, abdominal obesity is defined itself in terms of other variables, such as abdominal circumference. The rest of these are discrete observations.

In looking at these two guidelines together, which I think most primary physicians would be executing these guidelines many times a day, every day they see patients, we identified a total of 169 different distinct terms that had to be instantiated in order to apply these guidelines.

There were 42 variables that could be characterized as direct observation. There were 40 that constituted health issues, that is, diagnoses that would be on a patient's problem list, 4 were medication related.

Now, there were 15 variables that were categorized. We categorized these as needing to ask the clinician. That means there are things that were not routinely captured as part of a clinical assessment, but were needed in order to implement this guideline.

There were 32 variables that required calculation and 36 variables that were complex terms, that is, things that were defined in terms of other variables.

We found that in our EMR, only 51 percent of the variables were simple terms that were collected in the form needed for application of the guideline, 40 percent of them depended on other simple terms, and there were a lot of undefined terms, too, things such as children, adolescents, elderly, end-stage heart disease, that a clinician would need to understand in order to apply the guideline, but actually were not defined precisely.

Some terms were defined only outside of the guideline. There was one term "elevated coronary risk" that was referenced, but you actually had to look up another paper in order to find it.

I wanted to focus on just one of these complex variables that comes from the JNC7 guideline. It is called "Target Organ Damage."

Many clinicians may know what that means, but in order to implement this guideline, you have to know very precisely. If you look it up, there is a table in the guideline, and it consists of all of the following or any of the following: left ventricular hypertrophy, angina, prior myocardial infarction, chronic kidney disease, and so on.

In addition, not all of these things are simple terms either. For example, left ventricular hypertrophy, that is rarely going to be recorded in EMR as a term, that depends on some of these other variables, such as LV thickness or LVH by EKG criteria.

Another example is peripheral arterial disease, and it is defined in the guideline as consisting of one of these following things including, at the bottom, noted that one was just described as "Others," it was not specified.

These variables come from different parts of the record, so, for example, I have highlighted here the entities that would be retrieved from the patient's comprehensive history, whether they had a prior MI or not, or prior stroke, and so on, prior coronary revascularization.

Others depend on the physical exam - heart failure, for example, stroke could be something detected for the first time on physical exam, peripheral arterial disease can be both historical and a physical finding, and retinopathy depends on the physical assessment.

Some things depend on symptoms, such as angina, or symptomatic heart failure, and some things depend on diagnostic tests, things I have highlighted in red here, whether somebody has left ventricular hypertrophy or chronic kidney disease. There are diagnostic tests for heart failure, stroke, so some of these entities are captured by more than one modality.

In the case of peripheral arterial disease, carotid stenosis is detected by a diagnostic test.

The question is where are these things in the clinical record. Let's take left ventricular hypertrophy and congestive heart failure as two examples.

The data on LVH comes either from an echocardiogram, you can see there are two variables that I have highlighted, left ventricular posterior wall thickness and the interventricular septal thickness. In this case, they are normal, but that is where one would have to look.

The other place to look would be in the EKG report. Now, this represents the data as we received it. This is from an actual case. The echocardiogram was done in a private physician's office who is not part of our practice, and this report which we received, even though the LV wall thickness is a discrete variable, it wasn't captured that way by our system, because it is part of a text report.

The comment about the ejection fraction, which relates to the patient having heart failure, is not even recorded as data in the report. It is recorded as a narrative conclusion. The only way that we could possibly capture that is by finding this text and reviewing it.

Now, similarly, the EKG, this is a computer-interpreted EKG, but the way we received it in our electronic medical record is a scanned report, so the text is there, but the findings, if left ventricular hypertrophy were present, it could only be extracted from the text in that report.

Here is another example. Again, this is from the same case. This patient had coronary bypass surgery, and the only record of it, the only actual discrete recording of the prior surgery in the chart was in a letter from a consultant. This is from the physician who actually performed the surgery, because it was done in an outside hospital, the operative report was not available in our medical record.

One would not have this information without the ability to extract it from this letter, and it also provides the date that the test was done or that the surgery was done.

For evidence of peripheral vascular disease, we talked about doing a -- this comes from a carotid duplex report -- again, these are numbers, they could be discrete data elements, but that is not how we received it.

We received it as a text-based report.

It tells us exactly what the stenosis is, but again, this wouldn't be available to any of our decision support unless somebody extracted this information from the narrative text and put it into a variable.

This is one that was particularly vexing to me. This is an actual, this is scanned directly from the JNC7 guideline. They talk about patients with symptomatic ventricular dysfunction. So, how do you determine -- we know how to determine ventricular dysfunction, but how do you determine if it's symptomatic.

It comes again from narrative notes, in this case, a letter from the patient's cardiologist. She is doing well without any shortness of breath, so we can conclude that she is not symptomatic, but the only way to know would be to read this note. There is no variable in our EMR that says symptomatic congestive heart failure.

So, in conclusion, applying decision rules requires instantiation of a surprisingly large number of clinical variables, and many of these variables are not captured as discrete data even in electronic health records.

Many important data items are available only as textual entries in narrative reports.

One approach to this is increasing the discrete data capture by clinicians. I tend to agree with the consensus of people at this meeting that that is not totally realistic. I think there are limits as to how much we can capture discretely, but current decision support systems cannot make use of all the information that we have.

So, I think for the foreseeable future, natural language processing will be the only way of capturing these data from the electronic health record.

That concludes my talk.

[Applause.]

DR. HIRSCHFELD: We have time for one very quick question if anyone has one.

[No response.]

DR. HIRSCHFELD: We will move on to our third presenter, Dr. James Walker from Geisinger Health System.

Dr. James Walker, Geisinger Health System

DR. WALKER: Thank you. It is a pleasure to be here today. Thank you, Blackford, for the invitation. Thank you, Dr. Lindberg, for hosting us. It is always a pleasure to work with the highest functioning unit of the American Government.

Henry Adams said that the only obligation of a novel is to be interesting, and I think that is the only obligation of a presentation, so we will try.

I want to thank John Darer, who is my unindicted co-conspirator and the chief innovation officer at Geisinger. He is the person that makes all of this happen. I am the chief health information officer at Geisinger. I am an internist, a fugitive cognitive psychologist, and a student systems engineer.

At Geisinger, we operate a completely

integrated inpatient/outpatient EHR used by all clinicians for everything. We have a networked PHR that about 38 percent of the patients with whom we have some kind of ongoing relationship use and do things like receiving clinical decision support directly through that, and many of them, of course, activate it and get the responses back without any other human except the lab intervening.

We lead a health information exchange that links together several hundred facilities and I think 20 different companies, and lead a beacon community, which is engaged in trying to lead and execute communitywide evidence-based care processes across five counties of rural Pennsylvania.

So, I am going to talk about NLP-informed care-process improvement including clinical decision support. It is our belief we think we have learned that clinical decision support only is very useful when it is set in the larger context of care process improvement.

This is the way we think about that. The goal is not helping people make better decisions, whoever they are, patients or someone else on the patient's care team, but executing 100 percent processes.

The only number you need to know about American health care is the one that Blackford cited earlier, Beth McClynn's 55 percent. Right now -- and this was a very carefully done study, I have never heard any quibble with their methodology -- they identified 30 evidence-based interventions that apply to adults. They surveyed across several cities and communities in the United States and came up with this appalling number.

So, what our goal is, is to make sure that every patient gets every intervention offered to them, and if they elect to take it, execute it flawlessly 100 percent of the time. By the way, that would take care of health care disparities as a side benefit.

So, what does that look like for

Geisinger? What it looks like for Geisinger is we had our seven CT surgeons sat down and said let's identify everything that has been proven in a good clinical trial to decrease the risk of a patient having an adverse effect when they have elective open heart surgery.

They found 38 or 40, I can never remember the number, and they said, okay, we are going to commit ourselves as an enterprise to doing 100 percent, every patient gets 100 percent of those every surgery.

So, first, we looked at our existing performance, which, God help us, was better than national benchmark, and it was 59 percent. So, the team, and this is a team of surgeons and outpatient cardiologists and PCPs, post-op nurses and pre-op nurses, and everybody else involved in the process of cardiac rehab, everybody end to end, redesign the process, and then designed health IT to remind everybody what their part of the process was and let them know whether it had been done or not.

Within three months of implementation, we were at 100 percent, we have had glips since then, but this is not the most recent data, our run rate over the last 30 or 36 months is 98 percent. So, that is what we are after, and then clinical decision support is anything that helps us get there.

Alert fatigue, just so you have the right definition of it, because you probably haven't seen this before, alert fatigue is decision support provided to someone who is not committed to 100 percent process.

Our experience is that after everybody stacks hands and says we are going to do this, what we find is people actually in some instances scores of physicians sending us an e-mail and saying you need to put a soft stop on this particular process, because I keep forgetting, I keep closing the note because I remember to document medric conciliation, for instance, and then obviously, health IT in the absence of commitment to 100 percent process, it's just an expensive mess, and as the literature has abundantly demonstrated, it is unlikely to improve quality, efficiency, or satisfaction, and, of course, then, if you actually are trying to execute 100 percent processes, and that engagement runs across organization, then, it is indispensable and the argument goes away.

So, what does care-process improvement including clinical decision support look like to us?

First of all, the total goal is to improve health. As you probably know, it is estimated that health care accounts for about 10 percent of that, so we should be a little bit modest as we start out, but granted, that 10 percent is what we control, what is our take on it.

We need to support shared patient and clinician sense making, which as every clinician and all patients that have serious problems know, is an iterative approximate, incredibly high order intellectual task on both the part of the clinician and the patient.

Support knowledge acquisition, which may be part of clinical decision support, should support share decisionmaking, which is a little different animal than what we call clinical decision support often, and then translate those decisions into cost-effective processes.

One of the things that we do to try to achieve 100 percent process is what I am not sure if you would call a clinical decision support or not, when that cardiology team got together and did elective open heart surgery when we do congestive heart failure, when we do perinatal, what we do is embed lots of decisions into the process.

So, if we have decided that everyone with CKD, everyone with chronic kidney disease who has a glomerular filtration rate less than 60 should have one visit with a nephrologist, that is just built into the system, and no human has ever bothered with that.

If we decide that every woman over 50

should be offered a mammogram every year, that goes directly to the patient, and if the patient self-schedules in the electronic schedule that she is offered, gets the automatic reminders, gets it done, and it is normal, she gets a message back from her doctor says it is normal, no Geisinger clinician knows anything happened. So, a lot of what we call clinical decision support is embedded in all kinds of ways, so that clinicians are largely, I don't know about unaware, they are certainly not irritated by it, and then, of course we have got to execute the processes reliably.

So how do we use it currently?

What we do is what we call "closing care gaps," which we never see in public because we think that would sound appalling to patients, but what we do is work on care-process reliability.

So, for instance, we want to reduce the time from an abnormal mammogram to biopsy and from positive biopsy to treatment, and the actual performance standard is if you have an abnormal mammogram, the 100 percent process is you need to be offered a visit within 24 hours with the breast clinic, within 12 hours of the time that abnormal mammogram is reported. If not, somebody gets a nastygram, not the patient, of course.

So, we use NLP to find the positives. Why? Because the 19th century histology information system we use doesn't enable the histologist to flip an abnormal flag on it, so we have to do the NLP.

We have done the analytics. We have the process redesign underway, because obviously, there is no sense doing the NLP if we don't have somewhere to send the signal, and if those people don't have a performance expectation and an actionable usable, useful way of doing it, but once that is done, we will execute that one and then as you can see, we expect to do that with a lot of other problems.

You will notice the process real time. One of the things that is important for us is to understand how fast something has to happen for it to be useful, and it turns out that lots of things don't have to be in real time.

They have to be fast enough for whatever the process is, and so that is one of the ways we try to be efficient is to be smart about what the turnaround time for different kinds of NLP is.

That obviously is an advantage, because then we can process it off line, we don't have sub-second screen flips which our users rightly demand of us, and obviously, this is not fast enough to be point of care. We don't think it matters actually for reasons that I will talk about in a minute.

Other possible use cases, you know, you could imagine tracking all of these different process measures to make sure that we are taking care of patients appropriately.

This is what NLP informed care-process improvement looks like to us. Often, placing a clinical decision into business process management, you are all aware that everywhere in the world except health care there is a discipline of taking well-characterized processes, building them into software systems that then manage making sure that they get done or that the right person knows that it is their turn to do something.

Then, using NLP and BMP again, so that once the NLP is done, it still has to fit in a process and an execution system that is going to make it happen, and then get to 100 percent process.

So, why not point of care? We don't care that that thing I told you about isn't going to be point of care. The reason is because we don't want anything point of care that doesn't have to be there.

In our view, the patient's time with the physician ought to be focused on the very highest order, very most important, very most complex intellectual tasks, and remembering that it is time for the mammogram is not one of those.

So, our model is to take everything out of the point of care that could be done somewhere else as well or better, so that the patient and the doctor can look in each other's eyes and do that sense making, do that shared decisionmaking, that negotiation of what the care plan is going to look like, and so the patient wants out, saying gee, my doctor loves me and knows me and cares about what is happening to me.

One of the things is the networked patient health record is enormously valuable for this. Lots of patients in our underserved, undereducated, poor, old, immobile, underemployed population, about 30 percent of our patients are thrilled to take care of all this stuff themselves.

We send them an alert that says time for your diabetes bloodwork -- we don't call it that, of course -- and many of them go to the lab, get it done, get a report that goes back electronically to the doc, the doc sends an e-mail message to the patient, and that is done.

So, that is one of the things that helps us unload the point of care.

One thing we find is aggregation of appropriate information takes time. Our data warehouse has 13 databases feeding it, and not all of those work on the same time scale, and so not worrying about point of care helps us to aggregate that information, do the kind of heavy duty decision engine work that is often required to do a good job and then get it to the right person.

I want to talk very quickly about this, because I disagree fundamentally and almost completely with the idea we are going back to free text.

You remember that 55 percent? That was before EHRs destroyed the expressivity and richness and narrative art of the doctor's note.

The doctor's note is not the issue here. If doctors can write the American novel and execute 100 percent processes, we are all for it. If not, not, and it is not a measure, and so I think it is an important thing to get past that surrogate, so this is one reason why.

So, you wake up, and you have back pain so bad that you can barely get out of bed, and you can't stand straight up. You can barely get to the doctor, who miraculously is willing to see you today.

There are two kinds of back pain basically. There is benign, and it is over-treated, it is over-medicated, and enormous amounts of money are wasted on it, and then there is malignant, and it is under-treated and under-recognized.

So, this is about precision, this isn't about over-use, this is about precise care. If you are one of the 99.9 percent who have the benign kind, you would probably like to know that, and know that you don't have to be exposed to x-rays. It makes no difference in your outcome. The pain meds, if you need them, fine; if you don't, fine, that bed rest has been shown over a progressive set of randomized, controlled trials to be useless and actually bad for you, and so you can do whatever you feel like, and 90 percent chance you are going to be well in four weeks. You might want to know all that.

If, on the other hand, you are one of the people that has metastatic cancer, and if you aren't treated today or tomorrow, you are going to spend the rest of your life in bed and all kinds of miserable situations, you would probably like to know that, and you would probably like to be sent for an emergency MRI.

There are 16 questions that have been shown in a superb trial to differentiate those two states. We just ask all the doctors here, you just pigeonhole a doctor afterwards, have him tell you the 16. No one can do it.

If you do a chart review, you will find that there isn't a chart at your healthcare organization that has all 16 data elements in it, natural language processor and everything else.

So, what is the solution, and how can NLP help?

Well, first of all, NLP could help by, if it were fast enough, by identifying low back pain as the problem and teeing things up. It could look and see that there is no prior lower back pain anywhere in the record notes or otherwise.

It could identify that 4 of the 16 Deyo criteria are already known and pre-populate those, and then it could offer a template to somebody, the patient, the nurse, the doctor, whomever is appropriate, and that human or those humans could complete the 12, and then clinical decision support could calculate the likelihoods and the prognosis and the plan, and enable the patient and the doctor to work that through, because if it's benign, which it almost always is, the advice given isn't sort of the intuitive advice that you feel is appropriate when you can't straighten up because of the pain, and then the business process management system can put all of that into a after-visit summary that gets printed, because you are still a human being and like carrying paper around, and it also goes to your electronic file for you to look at later when the pain goes down and you can think, and then the BPM makes sure that you are followed up and really are one of the 90 percent at the appropriate time interval.

Are we out of time? That's enough.

Well, one thing just very quickly, let me suggest to you that level of automation is one of the things we need to think about. This is adapted from Parasuraman who actually outlines a very useful set of levels of automation which, as far as I can tell, we pay almost no attention to, but particularly when we are talking about NLP would help us a great deal, so if it's 90 percent recall and precision, then, maybe level 3 is appropriate; if it's 100 percent recall and precision, then, maybe it's appropriate to fully automate it.

Thanks.

[Applause.]

DR. HIRSCHFELD: We have time for one quick question for Jim.

DR. LINDBERG: I was just wondering, when you are contacting the patients, not all of them use the Internet --

DR. WALKER: Our database, the database knows every patient and their preferred communication channel, and it automatically routes it, so if they are using the network PHR goes there, if they still use snail mail, it goes there, where we are building out the ability to do appropriate things in text.

DR. LINDBERG: Can I call them up on a telephone and talk to them?

DR. WALKER: Yes, we are starting to use IBR also or both. There again, creating a call template that is usable for people takes work, but yes.

DR. WALLACE: I noticed in one of the slides in the middle there, you have the letters capitalized BPM, I assume that doesn't mean beats per minute, my background disambiguator I think is working. Could you tell us a little bit about business process management, how it relates to this?

DR. WALKER: Sure. Business process

management is a discipline. If you say, okay, Geisinger has 120 core processes that we need to characterize. We need to know all the steps, who is responsible, who can do them, what the triggers are what the time limits are.

Once you have characterized the process, then, there is a software system that you can put that process characterization into, and then that software system manages it automatically, so it says, oh, this woman, it has been 12 months since the last mammogram, send them a message, and by the way, they like it through Mike Geisinger, the network PHR, or, by the way, send it through snail mail.

So, it is a system that is just starting to be used in healthcare, that is critical infrastructure to making 100 percent process as a reality.

DR. WALLACE: And how does NLP fit into this?

DR. WALKER: Well, I think there would be a stack, and you would say look, you have got this process definition and you have all of these triggers, by the way, which tell you what information you need to collect one way or another, or go find an NLP, and say, you know, if we find this information, wherever it is, and if it is appropriate to use NLP for it, like abnormal mammograms, things like that, then, use the NLP in that layer, but then however it is collected, from the patient, from a clinician, from a physiologic monitor, from NLP, all of that feeds into the BPM layer and sort of runs the process.

DR. WALLACE: For example, you could use the process definition to decide whether to present a speech recognition module or a template based upon preference or other analytic studies.

DR. WALKER: Yeah, you could, absolutely.

DR. HIRSCHFELD: Our final speaker for this panel is Eliot Siegel from the University of Maryland.

Dr. Eliot Siegel, University of Maryland

DR. SIEGEL: I would like to thank the

National Library of Medicine and NIBIB for the invitation to present my perspective. It is really fascinating to hear the perspective of my colleagues in internal medicine, and what I wanted to do is give you a little bit of a perspective from my impression as a radiologist and diagnostic imaging where I have sort of had a career looking at decision support issues related to image analysis.

As time goes on, I am increasingly becoming convinced that really the future of our specialty is going to be completely dependent on our ability to integrate with the electronic medical record and to take advantage of natural language processing and enhanced clinical decisionmaking.

So, I am Professor and Vice Chair at the University of Maryland, and I am also Chief of Imaging for the VA Maryland Health Care System. I have some responsibility some of the other VA hospitals in the area.

I also work on personalized medicine at

the National Cancer Institute. I am looking at cross-correlating imaging with clinical and genomic and proteomic and other factors.

So, I have lots of interests in these areas.

Radiology is a specialty that has a long history of research and natural language processing and enhanced decisionmaking, but I think that these are now going to be absolutely critical to the success in our specialty.

I had to miss yesterday because this is also the week of our annual American College of Radiology meeting, and the theme of the ACR and the theme of this meeting this week is on quality is our image, but as I talk to my radiology colleagues, and as I hear some of the presentations at the ACR meeting, what I am really hearing is yeah, quality is really important, and we absolutely need to maintain it, but they are really getting an incredible amount of pressure to increase efficiency, and as time goes on, it is kind of like Lucy and Ethel at the candy factory as far as the increased volume of studies that are coming.

Blackford mentioned radiology, and we are trying to keep up with that incredible volume, and so are there opportunities and tools that might allow us to do that, that we can get from natural language processing and intense decisionmaking.

SPIE, the Society of Photo-optical and Industrial Engineers, just celebrated its 30th anniversary of PACS and researching computer-aided detection and quantification. I had the privilege to actually be in charge of that meeting for a few years, and in that meeting, we did a tremendous amount of research related to image segmentation and image processing, kind of like the natural language processing and decision support associated with the medical images themselves.

For example, I am looking at an unknown image and trying to find images that are similar or like that using many different types of characteristics, such as texture, color, morphology, and many others.

That research I think has been really important. We have been doing it for many years, but at this point, I think it is critical that we start looking at integrating with the electronic medical record to a greater extent.

I have been asked to give a talk at the Society of Imaging Informatics annual meeting this year in June about where we are going with the next generation of radiology systems, picture archival and communication systems.

I believe that our future advancements are really going to rely very heavily on the work that is done in natural language processing and enhanced decision support. The kinds of things that I see for the future of diagnostic imaging include providing relevant clinical information.

As a radiologist, the information that I get as far as the indications for studies go is really very minimal and very primitive, and I don't have the ability or time to go through the electronic medical record even much less so arguably then some of our other colleagues.

I am heading in the direction of personalized medicine where there is going to be increased information overload including genomic and other omic information. I want to correlate my findings in order to optimize quality with pathology, and I want to be able to instantly know what has happened with the patient and a patient chart.

For example, this is a graphical sketch of a patient's radiology history where you can see on the left there is a brain mass and lung nodule in January of 2000. In May of 2001, the brain mass is smaller, the nodule is stable. In March of 2002, the brain mass is gone, nodule is stable, but now the patient has developed new cardiac symptoms and a new cardiac problem with an acute myocardial infarction.

Well, if I had a computer system that had the capability of being able to present these three images for me as I am doing image interpretation, I could consume this in 5 to 10 seconds or so, and then drill down to detailed information, and, of course, not just me as a radiologist, but this is the type of thing that I would like to have in an automated fashion distilled, so that I can be as productive as possible.

When Blackford called and talked with me about what I might talk about, I gave him three suggestions. One was talking about natural language processing and enhanced clinical decisionmaking in diagnostic imaging, which is a three-hour topic.

Then, I said, well, I could also talk about a really cool huge natural language processing project called VINCI within the VA, and I said I could also talk about the work that I am doing with IBM and Watson with BQA related to natural language processing and the electronic medical record.

His answer was, "Yeah, they sound like great topics." So, what I am going to try and do in the next 10 minutes is cover all three, so kind of hold onto your seat because I am going to be moving really quickly.

First of all, natural language processing and radiology has a very long history, and because of the fact that we have a relatively constrained vocabulary, and a limited number of concepts for each modality, we have been an ideal specialty radiology reports to study as far as natural language processing, and hundreds of articles have been written over the past 30 years on NLP and radiology.

If you look at recent ones, there have been some great ones looking at many different topics. Dr. Payne wrote a really excellent one with colleagues on automatic identification of critical followup recommendations, and we are going to talk about that.

Natural language processing for devices, discerning tumor status from unstructured MR reports, being able to look at recommendations, natural language processing in chest, neuroradiology, pneumonia in infants, all sorts of different topics, and the science is getting better and better as time goes on.

In order to do what I want to in the future with regard to changing the way that we practice as radiologists, there is four fundamental things that are really important from my perspective, that idea that you would be able to abstract from me as far as natural language processing.

One, I want to know whether old studies were positive or negative. Two, I want to know whether recommendations have been made in the past. Three, if there are unexpected findings, I want the system to figure out that it is unexpected and have an algorithm to do that, so that it can alert clinicians who don't always read our radiology reports or pay attention to all the elements. Number four, I would like some help in automatically generating a protocol, and we will talk about some of those.

There has been great work done in many

different institutions, and just one that I want to highlight is some work that has been done by Keith Dryer and colleagues at Mass. General. It is something that they call LEXIMER, the LEXIcon Mediated Entropy Reduction.

Essentially, what they have done in a simple way is to write software that tries to extract for their database that they have of millions of reports whether or not the report was positive or negative, whether there are recommendations that were made or not essentially.

Even the simple things allow some really fascinating questions to be asked, and also allow us to be able to, in real time, guide clinicians in requesting radiology studies.

So, for example, if I can mine my millions of studies, I can collect data in addition to the American College of Radiology appropriateness criteria that would help to guide an ordering physician, let's say in this case ordering a head CT study for dizziness.

With the head CT for dizziness, mining the data that I have and looking at appropriateness criteria, I can make the determination that even though a head CT is being requested for the indication of dizziness, it looks like an MR would be a significantly higher yield.

You can again get that from expert advice or from the literature, or you can mine your own data as far as percentage of positive studies and as far as previous recommendations.

This can essentially be the same thing could be used whether it is a head CT. Here is an example of an extremity MR where there is a request for an MRI for a patient for arthritis. Here is another one that is relatively low utility, essentially a patient with back pain. Even an MRI in that particular case is relatively low yield as far as having a significant impact on the patient's care.

So, in this case, we can provide that feedback that for that indication, MR might be a

3 out of 10 as far as indication, but if we add abnormal extremity reflexes to the history, then, go back, that significantly increases the utility and the added value that the MR presents, and now it is a 9 out of 10.

This capability of being able to mine that sort of data adds a tremendous amount. As was mentioned, you know, the radiology reports can be fairly complex and being able to mine the important concepts is really, really critical.

Once I can mine those concepts, I can start looking at ordering physicians, for example, and see how they cluster as far as different studies. For MR of the knee, here is a clinician that is an outlier, that has a lower incidence of positive findings, but also a higher incidence as far as recommendation for followup studies.

Here is another one where we compare radiologists. So, we have two radiologists that kind of cluster for chest CT studies of the lower incidence of positive findings, but a higher recommendation rate. I don't know if these two are better or worse than the other radiologists, but it is interesting to see how they cluster.

Here is some clustering of positive findings versus recommendation rate for MR, CT, and X-ray. It is really interesting to see which ones had the highest yield of the positive findings. Presumably, as far as recommendation rate, what was happening was the radiologists would look at the X-ray and recommend a CT or an MR study. This is the difference for females.

One of the things I would like to see with EDM and natural language processing would be some help in evaluating the indication as far as automated protocoling, so once an exam has been accepted, such as an MR, then, it would be helpful for me to have assistance in automatically protocoling which MR sequences might be best to use.

Also, I mention for unexpected findings. We have a lot of patients who have, for example, a lung nodule that is incidentally noted on a trauma CT series, just like the ovarian cyst that was mentioned this morning, is an incidental finding. There findings frequently, quote, unquote, "fall through the cracks."

So, having a system that could reliably extract that information would allow us in a much better way to close the loop. We have manual methods right now, but the radiologists vary in their identification of these, quote, unquote "unexpected findings."

Of course, there is great work that has been done in natural language processing and decision support at many facilities. Here is a great one at Indiana University, and there are some really wonderful ones that have been done.

There have been many radiology clinical decision support tools. I am really happy to see Dr. Greenes in the audience, because he is an expert on these, and actually, I can remember as a radiology resident back in the early 1980s, Dr. Greenes coming by as Visiting Professor and talking about the wonderful tools that were available for decision support.

Between many that are either rule based or ones that use case base reasoning, there are a wide variety of these. One of the challenges, though, in being able to create these rules-based systems is collecting large amounts of data, and so it was great to hear that, you know, Partners is doing that, that Geisinger is doing that also, but I can't think of a larger database or a larger healthcare system than the Department of Veterans Affairs which has collected and now consolidated data within VINCI, so we have data now from 163 hospitals, 800 clinics.

The VA has been electronic and digital essentially for 20 years, and VINCI has data that goes back for over 12 years on over 28 million patients, and so there are huge numbers of amounts of volume, and so the VA has taken essentially a layered approach with its electronic medical record.

This is CPRS, which is familiar to many of you who practiced within the VA. It really looks a lot like a paper-based record, which was kind of something that they thought would be helpful in the transition.

All the information within the system is there, but you can't mine it. You can't ask questions like show me all of the incidents where this particular patient had a rash, or show me all rashes, for example.

So, the VA has created a team of dozens and dozens of folks who support HSR&D, essentially, a health services research, who are taking advantage of consolidating these data, and who have a number of natural language processing experts who are creating pipelines, that are doing processing and reprocessing of the data that is available to make it easier and easier to be able to do different types of statistical analysis.

So, I can't imagine any better sandbox that exists right now than the VINCI database. The team has focused predominantly on research applications. What I am hoping to do is to be able to utilize it to a greater extent for day-to-day types of decision support.

We have had to deal with issues related to security. There are many different data types, and I won't go through those, but pretty much everything in the electronic medical record, structured and unstructured, is available within VINCI.

Some of the biggest challenges that we have in natural language processing include the fact that templates are widely used, and templates can be very confusing when you try and do natural language processing, incomplete sentences and jargon, of course.

So, as far as next generation, I am not going to go into detail. I have been here actually in this auditorium presenting some of the work that we have done along with IBM and Watson utilizing the Jeopardy software.

I had the good fortune to get involved even before they played Jeopardy with medical applications and being able to use the deep QA software. As I am sure you all are, I am looking forward to David's talk in detail about how their deep QA technology works, but just suffice it to say that we are really interested and we have been working with IBM on trying to utilize what they have, which I think is a fundamentally different approach using very, very high speed.

If you could imagine that you have had an infinite amount of time, an infinite amount of processing power to form a hypothesis with every question, and go in search of dynamic database every time you formulate a question, what would be the potential associated with that.

I think the technology that they have offers that, so, in conclusion, I think radiology and diagnostic imaging has historically and will continue to be a rich sub-specialty for image processing, not only that, but also computer-aided detection, and also for natural language processing and enhanced clinical decisionmaking.

I think using these two techniques on

very large databases, such as the VA's VINCI database for research and clinical support purposes, have the potential to have a fundamental major impact on research, as well as our day-to-day decisionmaking in medicine and certainly in radiology.

So, thanks again.

[Applause.]

DR. HIRSCHFELD: Can I ask our three previous speakers to come up, and if anyone has any specific questions for Dr. Siegel. Otherwise, we can spend 10 or 15 minutes.

DR. McDONALD: I like your first slide that said you would like to always see positive or negative.

DR. SIEGEL: Yes.

DR. McDONALD: Is that really feasible given that sort of the national shrub of the radiologists is the hedge?

DR. SIEGEL: I think that is a great question and we kind of joke amongst ourselves as radiologists, as hedgers, and it brings up a couple fundamental questions.

Fundamental Question No. 1, of course, is what is a positive or negative radiology report especially when you may be commenting on multiple different findings.

The other one is how does one determine level of certainty, and so in radiology, our pendulum has also swung back and forth, and as time has gone on, we have gone from kind of structured reports back to unstructured, and now we are heading in the direction of structured reports, which increasingly require a level of certainty.

The best example, of course, is BIRADs, which is the classification system for mammograms, which essentially forces a radiologist to categorize every single mammogram, and that has had major positive impact.

So, I think to your point, it is difficult with free text to constrain an interpreter to being able to indicate level of certainty of a particular finding, and I think it is really important to do that.

So, a lot of the work that I have done in addition to natural language processing has been in support of structured reporting. In fact, one thing that we published on was actually complete graphical reporting where all you do is essentially take a pen on an image, delineate the area of pathology, and then essentially just put markers on it, and that becomes the entire report including your level of certainty.

But a lot of radiologists have objected because that does not allow them the free expression that they enjoy having, and what value that free expression has is something that one could debate.

There is value in the nuances associated with the English language, but the structure I think adds a tremendous amount, and I think, as the previous speakers have said, we are going to be in a situation where we are using a combination of structured and structure data as time goes on.

DR. SONNENBERG: I would like to know if for most radiographic procedures, if standard sets of findings or features have been defined, so that you could go through a checklist and say this feature is present or not.

DR. SIEGEL: That's a great question. The Radiologic Society of North America has recently released about 100 templated, kind of best practice reports, and so associated with those, there are certain findings within that template that a series of experts have gotten together, reviewed the literature, and said if you are reporting on an MR of the knee, these are the things that you need to include.

If you are reporting these particular other studies, this is what you need to include. We have taken it a step farther at the National Cancer Institute where we have looked at reporting out glioblastoma multiforme brain tumors;

As it turns out, if you look at free text reports, maybe you get 5, 6, 7 parameters in an average report, but when we looked at the literature of what correlated with genomic and proteomic data, it turns out that there were somewhere between about 20 and 30 that we extracted.

So, we actually created a templated work station where the work stations walks the radiologist through making measurements, asks specific questions that correlate with those particular findings, and now what you have is an automatic work station interaction that automatically generates a report that is tailored to a specific disease entity using a new form of annotation that has been created at NCI that we are hoping will be universal, called AIM, Annotation Image Markup.

So, you can see varying degrees of added structure as time goes on. One of the things that we are starting to see in radiology, and we are seeing for carotid studies, is pay for performance reimbursement, so they are saying you can charge for the report, but we will pay you a little bit extra if you make sure you include these particular elements, and this is only the start in diagnostic imaging.

Five, 10 years from now, in order to get paid, we are going to have to be able to document that we have answered certain questions, and the only way to do that is either with structure or with natural language processing that will allow coders to know whether or not there has been a report that meets certain quality criteria, and this quality criteria will be including these elements.

So, I think this radiology best agreed templates is a good first start, but from an informatics approach, there are multiple things that we could do to make it even better. So, thanks for the question.

DR. CARRELL: A question for Tom Payne. This is David Carrell from Group Health.

Really interesting stuff you are doing there with NLP-assisted composition of the problem list, and you mentioned that you just rolled it out six weeks ago.

About how long from the time that that project was a glimmer in your eye until six weeks ago elapsed?

DR. PAYNE: Well, we had an early effort to do this on our own, and we actually had Immay Schulte, who is not in this audience, but was here yesterday, and is over at NLP, that was three years ago we had this idea and had it in a prototype.

What we weren't able to was to convert that idea into a production system that worked in the real world. When we started doing the coding, it is an interesting metaphor here, a lot of our health IT in our hospitals began with billing, you know, 40 years ago, 50 years ago, and then we broadened it to a broader portfolio. The same thing is true here. We started with billing, and as a corollary of that tagging that I showed you, we realized we could do the problem with it.

So, that was very quick. That was after the agreements were signed, you know, less than a year that it was in production.

DR. GREENE: So, what I am hearing from this session is a number of hybrid solutions, it is not a debate of whether NLP instructor, but I think we need both, and somebody in each place is sort of drawing the line. You know, so much of an area is probably opportunistically or where your urgent problems are.

Can we create any general principles from this that can emerge as standard approaches? I think Eliot mentioned the ACR templates, for example, can we do that across our specialties and begin to have templates for the common or the high risk or the high utilized conditions, but we do have to capture these kinds of things.

DR. WALKER: One generalization that is important that this discussion has brought up is that the creator of the information is always more interested in expressivity than the recipient, so that our doctors over and over again say with the radiology interpretations, for instance, say can you please just put the impression first and all the rest of it later, and the radiologists actually get angry and say, no, they have to read the whole thing.

Doctors act the same way, so one of the rules is that if you think in terms of the information consumer, you will have a different approach to lots of this.

DR. SONNENBERG: Just a comment on that. I think one thing that would help a lot is if some of the reports that we get, you know, for example the echocardiogram reports that have measurements in them, were delivered to us as discrete data in addition to the narrative report.

For example, if you want to know left atrial size can be a determinant in somebody's risk if they have atrial fibrillation, why not just make that a variable and important, the same as a lab test. I would like to see a lot more of that done.

DR. WALKER: And that would be another principle. If you start with 100 percent process that you measure, then, it becomes easier to identify what you want to capture, and people who are responsible for capturing it, physicians and others, in our experience, have minimal resistance to capturing that.

What they hate like fire is their experience of, what, 100 years, of being asked to record all kinds of things that they know for a fact nobody is ever going to do anything with.

DR. PAYNE: I will just say this discussion we are having here is not new, it has been going on in my career. It has always been a tension between the way we put notes in.

What is different and new today is the adoption of these tools as they exist in the commercial world and their impact on the workflow and the time that providers have to spend with a patient. That is new and what I am reflecting here is their reaction to this enormous change.

The other thing that is new is the power of NLP, which 30 years ago wasn't to the degree that it is today. So, I think there will still be a spectrum where that line falls is going to change, and we have to respect the fact that really busy people are pushing back on some of the things asked of them.

DR. SIEGEL: But I think you ask a great question, and one of the things it gets at is what really is quality, and as time goes on, I think that we are going to have an increased amount of attention paid and meaningful use to what represents a quality note.

I mean I talked about what is a quality radiology report. How about a quality progress note? I remember when I was a third year medical student writing a really detailed, three-page note on a patient as an admission, and we physicians maybe can relate to this sort of thing, and got a B on it essentially, because it was graded. I asked, you know, why did I get a B, and the answer was, well, the other students who were in your group wrote 20- and 25-page admission notes.

I am thinking, well, what really is a

quality note and what should be in it, and how do we define it, and who is going to end up defining it, and I think one of the problems is we don't define quality at this point, which makes it difficult for us to really know what are the elements that should be in a template, a good quality note, and what things do we need for this vision of the future that we all have.

DR. WALKER: Just real quick on that one. Enrico Coiera had a great article in JAMIA 2000 when communication is better than computation. I think he addresses this very usefully.

There are some things that are so highly characterized, some data that are so highly characterized and so computable, you know, did I give the patient an aspirin when they had an acute heart attack, that they should be captured and structured, standardized form, and computed on, and there are other things. The patient orderly walks into the room and says the cone builds up at the top of my head and it really hurts and then it explodes, and I feel fine.

It would be foolish to try to put into computational form, and then the question is at least an askable and answerable question, is this datum, does it fit into a clinical prediction rule, is it used in some way to help inform patient care, or is it something that we are better off just enhancing communication, and not trying to standardize, compute, template, do all the things that we need to do to the really highly characterized information.

DR. McDONALD: So, this whole question about, you know, narrative text versus entering specific questions I have suffered with, too, as a clinical caregiver, as well as infomatitian, but the thing I think that is missing in our field and in the medical field in general, so if you look at things that are structured, like echocardiograph reports, and curiously, albeit echoes and cardiograms are all highly structured and radiology is narrative, someday we will figure that out, but they have done research for years, and they have figured out this ejection fraction of 40 percent means this, this, and this. Now everybody wants that number, because they have figured out which is the important of all these 50 things they provide and when it is important.

We haven't done that with clinical data, and everybody believes intensely that their physical exam finding is the most important, there is no data whatsoever, so this 20-page physical thing, you know, no human sane person would ever do that, but you are supposed to do it because the cardiologists are going to check that you have got this part of this murmur, that part of that murmur, which you never need anymore anyway, because you get an echo.

But anyway, so I think we need to invest some money in research and the conical data elements that are not measurements, that figure out which ones mean anything and which ones -- and a lot of stuff we do as physicians, in our notes, is just to help us get to the next visit, you know, a little memory jogging, has no use to anyone except ourselves, but we have enshrined this now, do all this work and getting into the database.

So, the last thing is that now that we have the computers in our workflow, that means the administration can make us answer any dumb question they think of. That is going to be an issue.

DR. PAYNE: A good example. Taking a family history, creating a pedigree, how is that done? We have a genetics clinic. They spend hours doing that in very great detail, which they should, and when they are done it is gorgeous, but I see these people, you know, day after day, I get a little snippet with a family history each time.

If it were possible for me to pull this together without spending three hours taking it, but each time asking a little bit more, put that story together, I mean this ultimately does have an impact on that person's health, but the tools to capture that right now are not facile enough for me to do what they do in the medical genetics clinic, so I do a little piece of it at a time, and a lot of it is no longer a sketch on a piece of paper, it is now in text. We need to be able to do better with that, and get that text into something that can help me counsel this person about what further testing might be warranted.

DR. SIEGEL: I just had a question for you about the problem list. In my experience, it has been really interesting, there is no sheriff of the problem list, so anybody can write into the problem list, and the problems that we have -- and I am wondering how you deal with it -- is we have things, lots of stuff goes on the problem list, but it is like a roach motel, you know, stuff checks in and it never checks out again.

So, what happens is you have one episode where somebody measured a blood pressure that was elevated, someone writes in the problem list "hypertension," but whose responsibility is it to take that off, or somebody has an elevated blood glucose or somebody has back pain, you know, transiently, they come into the ER. It is on the problem list, but whose responsibility is it to take it off, and have you guys looked at actually taking things off the problem list?

DR. PAYNE: We are doing it. I have to say my first job as clinical fellow at Mass. General was helping with the problem with CoStar, and they had over years created a working tool. We are not there yet, and many organizations aren't there yet. I would guess that is one of the reasons it's on the meaningful use trajectory.

So, we are changing the culture from putting it as a list of my thoughts about this patient, and making it encoded, and that is growing, and we want all disciplines involved. We thought that far ahead.

We want everyone who has a view on that patient to contribute, and then we have to decide as a community how we are going to tailor and trim it, but I can't give you experience on that yet.

DR. SIEGEL: I just wanted to make one comment about the problem list. You know, in one other meaningful use requirement now is that we make electronic information available to the patients, and that includes the problem list.

Some of our physicians were really upset with the idea that patients would be seeing a problem list. Well, it is outdated, and inaccurate information, and we said well, then, fix it.

So, some of them who never cared about it before are now going to be motivated by the fact that they know the patients are going to see that.

DR. WALKER: We are making a discipline of managing the problem lists, so at some levels, it is an enterprise.

DR. SONNENBERG: So, who is responsible?

DR. WALKER: At one level, the enterprise is, so if you have got a GFR of less than 60, you have CKD on your problem list automatically now, the docs all agreed to that, but that is at an enterprise level automatic, nobody thinks about it.

If it looks like a patient might have diabetes, we send a message to the doc and say it looks like they might have diabetes not on the problem list, pick one of these three. If they have diabetes, but the neurological adverse effects aren't documented, we capture that, so there is that level. There is the level that the PCP in some respects on this problem list, and then there is some which is just a food fight.

DR. RIPPEN: Helga Rippen. It is more a philosophical question. It is really about where we are right now, the direction that we want to go, so if we think about information and how do we actually provide tools to improve the quality of care that we are actually delivering, support the caregivers and the consumers with care as opposed to measuring them, again kind of an interesting thing.

If you reflect back on the days of HETAS, where the intent was really to do preventive care, it became a well, let's go report and focus on actually getting the measures.

So, again, as we start leveraging these tools, let's think about what is it that we are trying to actually accomplish and then perhaps how do you most effectively do that without the adverse consequences.

DR. PAYNE: I am a big fan of doing what we know to be the right thing to do. I also practice, as well, and I have. The other requirement is that my thoughts, my confusion about a person's symptoms are helpful to me and to my colleagues and to sorting things out, to knowing what the heck is this causing, and sometimes I can figure it out myself, sometimes I can't.

So, in addition to making sure that I, to the best of my ability, prevent problems that we know how to prevent, I also want to summarize what I have heard and found, so that I can eventually figure out or ask a colleague to help me, because ultimately, that can also save lives.

DR. SIEGEL: You mentioned preventive medicine. One thing that most of us have emphasized is the electronic medical record, but there are so many sources of information, as time goes on, I think we are going to see more and more patients entering information whether or not it is their weight or their blood pressure or glucose monitoring, et cetera, into the record, or even tweeting, I mean we have been monitoring tweets at the University of Maryland across the country, and so as far as preventative medicine and surveillance, I think that we should look at natural language processing and enhanced decisionmaking beyond just the electronic medical record, and look at the entire continuum, and I think it is an excellent point that you make.

DR. HIRSCHFELD: Just to wrap up this session, I just wanted to ask everyone on the panel if they had one thing on top of their wish list, what would it be. You know, in clinical practice, if you could have it, what would change your life?

DR. PAYNE: I guess the one thing I am just reflecting on the safety problems that we work on, but we don't fully have. The one thing I would ask for would be a way to protect us from not following our oath to do no harm, and to save the patients, but also the practitioners who suffer when harm occurs, so a better way to do that.

DR. S0NNENBERG: I think for me, what would make the biggest difference is to have truly seamless health information exchanged, so that everything that was available, everything that was known about that patient would be available to me when I see them.

DR. WALKER: Particularly in the context of today, I would love to have an NLP processing engine that ran against all of the data including our community data warehouse, and was able to provide appropriate information with the best estimate of its reliability that we could deploy wherever and whatever process we needed to.

DR. SONNENBERG: And I agree with Frank,

what I was going to ask for may sound like a contradiction in terms. What I would like to do is be able to have a mechanism to monitor all health information related to all patients without any compromise of privacy or security, and figure out a way to do that, and I think it is something for us to strive for.

DR. HIRSCHFELD: Great. Let's thank all our speakers again for this session.

[Applause.]

DR. HIRSCHFELD: I think we will take a 10-minute break and reconvene at 10:50. Thanks.

[Break.]

Panel 2:

Perspectives on Clinical Decision Support

Moderator: Dr. James Luo, NIBIB

DR. LUO: This session, we are going to hear the perspective from the natural language processing expert, to see how the natural language processing and the critical decision will be able to address some of those challenges and help that meet this need.

We have an expert panel of speakers for this session. We have Dr. Robert Greenes from Arizona State University and Dr. Li Zhou from Partners Healthcare, Dr. Stephane Meystre from the University of Utah, and Dr. Mark Musen from Stanford University.

I would like to call everybody up to the front row, not here. It is hard for people to watch the screen from upfront, but once we get into the discussion, a portion of the session, then, everybody will sit in the front.

With that, I would like to quickly introduce the keynote of this session. Dr. Robert Greenes is from Arizona State University, and he is Ira Fulton Chair, Professor of Biomedical Informatics, and he is a also Professor of Biomedical Informatics at the Mayo Clinic.

His research is focused on clinical informatics and particularly on clinical decision support and health care quality improvement and application usability and interoperability to optimize clinical care process.

This is a very, very short summary. With that, I would like to invite Dr. Greenes to give the presentation.

[Applause.]

DR. GREENES: Thanks, James.

I am not a natural language expert, and I think what I was asked to do is provide kind of a perspective on clinical decision support more broadly. You will see how NLP fits into this. It is kind of a research agenda talk about some of the remaining -- you have heard a lot of this already -- but I am going to kind of provide a perspective on the grand challenges for clinical decision support from the current perspective.

This is actually not a new thing. Actually, there was a nice article by Dean Sittig and others in 2008, indicating 10 grand challenges. They are not the same as mine. In fact, in four years or so, a lot has changed already. There have been some other studies recently. There was David Lobach and colleagues had a review of effectiveness of CDS, I don't know if it has been actually released yet or not.

Jonathan Teich at the last AMIA meeting had a summary of some of the state of the art of clinical decision support. So, these were all kind of beginning to set the stage for where we are now, but I think there is a lot new in the current environment.

First of all, we are seeing finally some convergence on data models and data interoperability, and many large-scale projects are beginning to work together to try to get to that aim, so I think that is very exciting. Obviously, we need that for decision support.

There is also increasingly the availability of large databases. We heard about VINCI. The Mayo Clinic has a large database, Partners has a database. There are many sources now where we can begin to do population-based kind of evaluations.

Another key thing that is driving decision support new perspectives is the idea of the continuity of care across the continuum, and patient-centered medical homes, the kind of care organization movement is beginning to change that focus to looking at that whole issue.

This is also being driven by meaningful use initiatives, incentive payments, and quality foci, and are emphasizing the need for continuity and for quality and for value and for efficiency.

Another thing that is not often realized is part of this issue is this emergence of an APP culture, because I think what happens with that is it now frees our thinking about how to present and utilize decision support in other kinds of ways to interact with our patient care delivery tasks more efficiently and effectively, so it creates an innovation environment.

I think with all of these and with the concept of continuity of care, we are beginning to see glimmers of something that some of us have talked about for years, the integrative longitudinal permanent patient record.

I always like this slide. I have shown

it for 20 years, and it is always true. My top 10 list of challenges actually fits in three categories. One is the framework for care, and I will talk about each of these briefly.

The knowledge sources where we are going to get our knowledge, how we derive the knowledge, how we assemble it, and then how we use it. So, we will talk about each of these, I am not going to spend time reading through this slide.

But the scope of the care process I have touched on already is really now increasingly able to be considered to be the whole care continuum including health and wellness. We are beginning to see, at least talk about, but there is no emergence of a longitudinal patient record information model.

We do focus on how to get data across EHRs in continuity of care documents. We talk about PHRs sometimes tethered to a health system, but independent sometimes there is no standard for that yet. Eventually, I think we need to come to grips with the fact that we are talking about a patient and a longitudinal record for a patient, so we need a single integrated data view, and I submit that is a key challenge.

We also need a single source of truth, and so as have multiple providers, we have to do medication reconciliation, problem list management and reconciliation, care plan management, reliable timelines for when events occurred and being able to look across those, and also managing the roles and responsibilities of who can update, the problem lister who can update the medication list, and so on, and do you trust other people and their lists before you update them.

There is a lot of research in various forms of CDS delivery, and I won't go into these in detail either except to touch on the fact that these are all areas where there is work to be done, where we thought we knew how to do decision support, and yet there is much to be done in each of these areas. So, we have new modes as well as refinement of old modes.

One of the oldest ones is rules, and we are still having hot debates. ONC sponsored a meeting, what was it, just a week ago, trying to converge on what we mean by rules and what we mean by how to adapt them to local factors, so that they are actually used in the workflow in setting of an individual site, guidelines and protocols and capturing the workflow, not only of the individual and the patient going through the process, but the team and the care process that has to support that. We don't really have a good way, and I will come back to that one actually.

Calculations and algorithms increasingly now have warfarin dose than someone taking account of genetics, and complicated radiation dosing and other kinds of things. These are all areas of increased activity, and I should mention image processing methods indeed that fit into this.

Then, we have prediction models, and we are not using them at all well, but there is all this potential for prediction models, machine learning, fuzzy models, and so on.

Anyway, I said I wasn't going to read through the list, I am starting to read through it, but in addition to the advice tools, there are information tools, and I consider these basically presentation and viewing tools, enhanced visualization and summarization trend viewing dashboards, and things like that.

Ways to retrieve information at point of need, question and answering systems, the Watson and other, the use of info buttons increasingly trying to actually anticipate context, so that they get down to the paragraph level or the very specific level of information you need for that context.

Feedback and quality reporting are information tools, and they drive performance improvement, possibly social networking in terms of figuring out what works best and what kinds of information are most useful in various settings.

Then, we have all the kind of subliminal types of ways of improving decision support by making the right thing, the easy thing to do, having order sets or structured documents or templates for particular situations.

Now, we don't at all deal well with patient-centered and shared decisionmaking, and this was supposed to be a link, I guess, there is no Internet connection here, but to the Lyrica ad that you may have seen on TV.

More than half the ad, I timed it, is all the contraindications and risks, and it says, at the end, what? Ask your doctor. Well, what is the doctor going to say, and is the doctor equipped to answer all the relative risks?

Then, we have sites for cardiac diseases, risk assessment, cancer risk assessment, and so on, and then the genome profiling, following sites patient like me, and so on, and again ask your doctor how to interpret this.

Then, we have personal sensors and home health care, and all the potential for patients entering other data and updating their data in their personal health record. We talked about gathering family history. Well, the patient is a great source for that. Why does the doctor have to spend all the time entering it? They can look it over and kind of disambiguate things, but we can get a lot of this information from our patients.

What kind of decisions board can we deliver directly to patients, what kinds of shared decisionmaking processes need to be there, and when do you escalate decisionmaking to the provider.

Now, we also have lots of challenges in knowledge representation and formalization that we have touched on where you have this whole process of going from an evidence-based medicine recommendation, or a population-derived presumed best practice to codifying it, so we have lots of approaches I have listed here that are beginning to try to take the information from narrative to executable.

We have standards, some of them, order

sets, info button manager. There is a data model called the VMR. There is a decision support SOA-based standard, but we have no executable guideline standard after many years of trying to do that.

We have no health care workflow model although there are business process workflow models. There is a need for context setting factors as a kind of set of standards.

We use the customerizer decision support. We need a better understanding, and I think Marc will probably talk about this, about where guidelines fit into the care process, where patients, they don't come in labeled to be on guidelines, they come with multiple problems, and you need to figure out which guideline fits them, or guidelines fit them at this point in time, where are they on that guideline, and you need to take parts of the guideline that are not executable and separate those from those that are executable, and where do you embed them into the care process. So, how do you decompose them?

One thing I like to think about, you know, and it has been very intriguing, maybe as we can start setting goals for patients, we can get there, is the GPS analogy. You know, it is a beautiful piece of engineering. We can use it in passive mode or in goal-directed mode.

We can say where we want to head, and it kind of tells us the road to get there, tells us points of interest or points of knowledge along the way that we may need to know, and if we get off course, it can tell us how to get back on course. It basically tracks what we are doing and keeping track of that can help us to figure out where we want to go.

I like that a lot, and we haven't figured out how to kind of use this for health care, but it might be the kind of both passive and active guidance system that we all want to strive for.

Now, when we amass our knowledge, we have to worry about where are those authoritative knowledge bases, which ones do we trust, what kinds of governance or oversight peer review is going to manage them, who does that, how and when, how do you disseminate them, and how do we overcome the intellectual property, what is the role of knowledge vendors in this versus public and professional subspecialty provided data.

Then, separate from the knowledge content resources are the knowledge management tools and resources, and we have heard about CDSC consortium, for example, building tools and some other groups are doing that, as well, but many places, there are really no marketplaces for good knowledge management tools for enterprises to manage their knowledge resources.

Can we create a set of public domain or open-source tools to build on to be able to allow both the national knowledge management and also local adaptation of it.

Then, I mentioned how do you customize, can we create a level of ability for, let's say a practice, to be able to identify its workflow and its settings, not at the level of Arden syntax code or jobber code or jewels code, but at the level of saying how I want things triggered.

I have a nurse assistant that can gather information before I see the patient, or whatever the workload is, and be able to add that in as factors, and then have the decision support be customized or semi-automatically adapted to the vendor platform to be able to be delivered.

Now, I think another area that we have touched on, and this obviously is one area where NLP and probability fit in, is can we use our massive databases that are being collected.

First of all, temporal modeling of longitudinal databases is a big issue. You can get data arriving at episodes of time, and actions are started at episodes in time, but what can you say about the data or the values at an arbitrary point in time when you are trying to look at a patient's status.

So we don't have good models for that, and I think we can build them, and there are ways that people are trying to do that. We also want to build prediction models from those databases, and we want to be able to retrieve cohorts of patients for building those models, but also to structure the encounter.

I think we can anticipate, for example, a diabetic coming in -- and I will talk about this in a minute -- for a particular setting, and you know exactly what kind of diabetic this is, and you have a cohort of patients that are just like that, and you can tell what worked and what didn't work. So, this is that "patients like mine" concept.

Then, there is the related issue of when you have population-based cohorts, what do you do when some new finding comes along, and how do you update this new information that is not in the database, and kind of weight it appropriately, so that is able to not get overwhelmed by the years and years of old data that you have, so there is research that has to be done there.

This is just an example of population management from Mayo Clinic strategy that is being developed where you take all this data and you organize it in tables, be able to find outliers in terms of findings, actions, and outcomes that are not anticipated, and also be able to manage both your high risk and your high utilizing patients, but also to begin to build these cohorts, so that you can pull out and build the on-the-spot decision support, much like Aramis did years and years ago in a structured rheumatology environment that Jim Fries developed. We can't do that with our regular patient databases yet. I am hopeful.

I think a major challenge and opportunity is to be able to integrate and create interoperability for widespread use by taking advantage of APPS and approaches to building visualization metaphors and interaction metaphors.

A number of groups, including our own, are starting to work on building this APP layer with our middleware set of specifications that take advantage of existing middleware and also new middleware that could be developed, so that you can really build the best or have competition for the best visualization or the best reconciliation or best problem list management tool.

A lot of works are related to APPS, is on usability, can you capture context, more specifically, as I mentioned, can you create visualization and other kinds of metaphors, can you deal with team cognition and status evaluations and handoffs, and things like that.

I think you can't read these, but these are two examples, and on the left, if it's in video, is this medication reconciliation package Catherine Plaisant and Ben Shneiderman's group put together to match the patient's description of medications and the hospital's or health care organization description of what they think the patient is taking, figure out how they reconcile and which ones don't reconcile that need further attention.

Then, there is some work on Life Lines, which is basically -- and I don't particularly like this particular model -- but the idea being to look at trend plots of problems and diagnoses in medications, lab values that you can track over time and various time windows to rearrange and things like that.

One of the key opportunities of health care I think is what I call CCC, care, continuity, and coordination, and it is sort of back to the future, because it sort of relates to the problem-oriented medical record that Larry Wei championed back in the '60s and '70s, and especially for chronic disease.

We have patients with multiple problems, they have multi-specialty care, we have critical care, and any kind of team-based situation, we have handoffs, we have management across episodes of care, and in and out of the hospital, and our current EHRs are really poorly suited to this, as somebody mentioned. They are basically a document paradigm and what we need is kind of a transaction dashboard manipulation paradigm.

I like to think about patients coming in sort of states, and gets back to my cohort idea. If we could figure out the patient's state by some descriptors, and certainly if they come back for a return visit, we already know what state they were in by the data that we already have on them.

We can anticipate what data we need on this new encounter, and we can also identify what the prototypical assessments might be, and for each of those, what the prototypical plans might be.

So, you can basically organize the encounter around that, and this is just an example of diabetes that may evolve from initially controlled by diet to insulin dependent, to complicated, and so on.

So, just kind of schematically, you can think of a patient coming in for a return visit perhaps with multiple problems, and for each of those problems, there should be a goal or several goals associated with them, like lose weight, you know, get your blood pressure below a certain level, or whatever.

Then, some other problems may have the same goals, overlapping goals, like obesity, for example, might be a problem. The goals, of course, multiple problems, as well, so there is a backward link. For each of those goals, there is an assessment of what the gap is between the current patient's status and where your goal is. So, that is the assessment part, and then your care plan should be tied to that. So, I am just expanding one of these rules at a time.

Actually, when the patient goes home and sort of being monitored, you can be monitoring their status, and this all, of course, feeds back, so that when you have return visits or return encounters, by phone or however, e-mail, that you have this up-to-date information.

So, what the challenge is, is to figure out how can we create dashboards so that problems be manipulated, their status be updated, disregard the ones that are inactive, and focus on the ones -- and having your own specialty view or care domain highlighted, so that you focus on the ones that you are involved in, but the others are available, and be able to kind of tie the problems and their update to the goals in the care plans.

So, I think we have a lot of opportunity in that scenario to create new APPS, new ways to interact with the EHR, and if you do that, in fact, as you make those transactions, you can almost self-document your progress note, because a lot of what you are doing there is actually explaining your reasoning, you know, what you are trying to achieve and what you are doing about it.

Rather than writing your progress note and then going out of that to update your problem list, then going out of that to update your order set, you know, it is all one process. Well, actually, let me go back to that for a second.

So, there is many opportunities for decision support in this, as you can anticipate, because the problems, you know, the goals associated with them can be kind of defined by decision support. The gaps of what you do about them, and the assessments, can basically be triggered by that.

The care plans, again, so all this process can be kind of highlighted and brought to bear by decision support.

I don't want to omit talking about safety, quality, regulation, and liability. Quality drivers are very important, and I think there is still a big difference between quality measures that are reactive versus decision support, which is proactive, but we have an opportunity to synchronize and drive off of the quality emphasis that's there.

At some point, you know, we haven't seen this yet, but when does CDS become the norm of care and where not doing it becomes an issue for liability, and also, how much of the black box is going to get regulated. So, these are all challenges I think we are facing.

The overall goal is to make CDS

invisible, I think, to guide the care process, anticipate user needs, make the right thing the easy thing to do, facilitate auto-documentation, auto-ordering, and so on, and focus on usability.

This is just, if I have time to just summarizing a couple of things.

I think rules of NLP are in the question and answering systems and info buttons, findings cases for interventions from population databases, identifying problems we have heard about, detecting adverse events, finding data needed for CDS, integrating diverse sources of data, Kavi Wagholikar was in the audience from Mayo Clinic had a post doc, developed very nice applications on pulling data from pap smear reports and colorectal prior reports to help make reliable recommendations for that.

There is a tradeoff issue that we have touched on. The documentation is easier if notes are list structured in certain CCC situations. There is more benefit from structure, so where do we draw that line.

Where is a potential role that we started touching on voice for navigation and completion of structured templates, and then I wanted to just mention roles for probabilistic techniques. I don't think I have time to actually go into it, but I talk about all of these.

Lastly, I think that I can say that the CDS landscape has great new changes, probably will continue to change. We have a lot of opportunities to rethink the car paradigm by integrating the diverse data and knowledge sources, focus on usability and on the continuity and coordination of care paradigm in designing decision support to support these tasks and developing in or out pool APPS that will actually allow us to kind of free ourselves from the lock-in of current vendor systems, and in all of these, an important role for NLP and probabilistic techniques.

[Applause.]

DR. LUO: Maybe we have time for one or two quick questions.

DR. SHAIKH: That was really a broad-ranging talk, it was very helpful in setting the landscape. My question is when you talk about the potential for the development of applications to support CDS for patients, providers, and systems, what do you see as the role of innovation and perhaps public/private sector partnerships to help stimulate the development of these CDS-related applications.

DR. GREENES: Actually, I think it's essential. We actually had a meeting in Scottsdale February 1st to the 3rd, had the VA and the DoD, Intermountain Health, Mayo Clinic, Georgia Tech, Open Health Tools, and I forget who else, Hubbard, Smart Project.

Our goal was really to see if we could agree on the need for an interoperable set of specifications beyond data interoperability, which is obviously essential, but there are a lot of things like context management, and so on, that need to be addressed.

So, we came away with that with a pretty resounding yes, and it should be driven perhaps by these large-scale care providers that could actually articulate the need for it and maybe build demonstrations, so one part of that is to kind of create sandboxes, so the consortium would be kind of a public/private development of this set of specifications.

But then the sandboxes can be semi-commercial. They basically will kind of become entrepreneurial work spaces, ecosystems, that can allow APP developers to work with middleware and lower level service and tool providers, and create kind of the technical infrastructure to allow them to work together to prototype, build applications, so in our Mayo ASU environment we are focusing on the continuity of care coordination issues, there is problem list management that they already have internally, that they are trying to come up with a universal solution for, and there is other kinds of things like that, that are driving it.

I think we will see other examples like

that.

DR. LUO: Thanks.

[Applause.]

DR. LUO: Our next speaker is Dr. Li Zhou from Partners Healthcare System.

DR. ZHOU: Good morning, everybody. My name is Li Zhou. I am from Partners Healthcare and Harvard Medical School.

So, why do we need natural language processing in CDS? This figure shows you the availability and the usage of electronics clinical notes in our ambulatory electronic health records.

You can see that the availability of the electronic clinical nodes has increased magically in the last 20 years, and in the time we have 13 clinical notes in our review chart. So, we also know like much information that could support CDS is in those textual data and therefore cannot be leveraged by a CDS system without NLP.

First, I want to talk my journey in and perspective of the two areas.

When I was a doctoral student and to pursue my Ph.D. in biomedical informatics at Columbia University, my dissertation focused on natural language processing and temporal reasoning. We use NLP system that lead to convert the narrative report into structured format, and in this way develop a system called Time Text to conduct reasoning and identify the temporal aspect of medical events.

Example, two years before admission, the patient was diagnosed with hepatitis. The patient had a liver transplant on June 1992. He underwent a t-tube study and then present a fever lasting for two days.

By conducting natural language processing and temporal reasoning, we could identify when the events occur, how long it last, and then calculate the duration lead to different medical events, and also deduce the temporal relationship such as after and before.

I joined the clinical informatics research and development team a few years ago.

The group was led by Dr. Middleton. I am particularly in their clinical support team where I have great opportunity to learn a lot of things in regard to real time clinical decision support.

We are involved in the developing and maintain clinical decision support interventions including reminders, alerts, infobuttons, order sets, et cetera.

Those clinical interventions have been implemented in both inpatient and outpatient settings, applications including like CPOE/e-prescribing, eMAR, et cetera.

In recent project led by Dr. Middleton, we also investigate and propose method for representing the clinical knowledge and also the data elements for those CDS interventions to make the clinical knowledge sharable across different clinical settings and different systems.

We also implement service-oriented architecture to provide a centralized service, to share knowledge and also achieve system interoperability.

So, I have been involved in these two areas. There is one question always in my mind, how we apply NLP into CDS. There is many areas we can apply NLP techniques as mentioned by Dr. Greenes.

Here, I will give a few examples in the following three areas: improving patient safety, enhance EHR functions, and reduce the health care cost.

For patient safety, I want to use medication list as example. We know medication errors can cause injuries, are common, and are very costly.

Adverse drug events and medication errors are estimated to cost the U.S. health care system $177 billion per year.

Medication lists within patients' records are often outdated, in complete, or inaccurate, which is a major cause of medication errors.

Active medications are often not added in a timely manner to the structured medication list.

Wagner and Hogan found discrepancies between the number of medications that patients reported taking and those listed in their EHR was one medication.

In addition, outdated medications are frequently not deleted. One study found 67 percent of medications were still active one calendar day after their inactive status was documented in the clinical notes.

Medication reconciliation applications, because of those errors, so MedRec applications have been built to address this issue. MedRec applications draw data from different results and to try to create a more complete, more accurate, updated medication list.

This screenshot shows you an example of reconciling medication lists before admission. This shows you an example MedRec application to reconcile a medication list after discharge, however, those MedRec applications use highly unstructured data.

One of our studies funded by AHRQ we

found that 30 percent of active medications mentioned in notes were missing from patients' medication list particularly those prescribed by a specialist outside the institution.

In addition, clinicians often need detailed or additional information beyond the medication list in order to make judgments, changes and other decisions.

For example, they want to understand the history and progress of the disease, also, they want to look at the consultation notes from medical specialists.

One potential application we can use extract medication from NLP output, and we can put several buttons into the existing MedRec application, for example, when a user click the possible missing medications button, because other medications extracted from notes, but missing from medication list.

The system can also provide a warning when they found a high and important medication.

Also, we can provide a note button

alongside of each medication just like the info button. When the user clicks, they will see the notes where the medication was mentioned.

However, there is many challenges to do so, how we integrate/couple NLP with CDS. One of the issues is data interoperability and terminology standard, we know medication list may be coded using an institutional or commercial terminology, while most existing NLP systems encode clinical text using standard terminologies like people mentioned the SNOMED, et cetera.

This requires the system encode information using multiple terminologies, and is able to conduct dynamic or static mapping as needed.

Another challenge is system and data integration, integrate NLP system with other EHR applications, and conduct the data integration, aggregation and summarization for both narrative data and structured data.

It would be nice to see some conduct reasoning or start doing some inference using, for example, knowledge base.

The second area I want to talk about is how to use NLP in EHR functions. Here, I use CPOE as example. One of our studies found a 7 percent of medication order entries are free text even though it would have been in place for 20 years, there is still a lot of free text entries.

We also found 9 percent of hypoglycemic medication orders were entered using free-text, and 75 percent of those free-text entries have an exact name match in our medication dictionary, and the remaining 25 percent of the free-text entries could be coded if specific formulary information was also provided.

We also found interestingly 17 percent of free-text hypoglycemic medication order entries including a misspelling. Here, I show you several most commonly misspelled terms.

CDS is not triggered when a medication order is entered as free-text.

Using similar data, we found 92 drug-drug interaction alerts were not triggered due to free-text entries, affecting 84 different patients.

196 patients who had a free-text hypoglycemic order entry also had the same exact drug entered as a structured and coded order during the study period, and 10 percent of those had identical drug entries active in their medication list at the same time. This is kind of duplicate therapy error.

Only 26 percent of those patients had diabetes recorded in their problem list, so we also have issue related to problem list if they are not completed well.

These CDS aspects are critical to patient safety, if unintentionally bypassed due to free-text medication order entries may result in potential harm to the patient.

NLP - CPOE/CDS - can provide advanced search function. Such search function should not be limited to only detecting exactly spelled medication name.

It should provide a relevant and a smart list, not just a long list to sort through.

It should also have spelling error detection and correction, for example, can provide a list of suggestions for the correction, it can do smart autocorrect.

Importantly, we need to design efficient user interface to address workflow issues, for example, auto-fill features, allow providers to create their own favorite list or pick list, however, the free-text entries should be monitored. We also should avoid navigating through multiple screens to save their time. It would be more efficient if we can incorporate speech recognition, so, for example, it may be able to reduce misspellings.

The last example is how we use NLP to reduce health care cost. There is a great need to minimize the cost of care delivered while still meeting quality initiatives. Five percent of patients generally account for up to 50 percent of the cost.

One example is how we use NLP and CDS to prevent readmissions by identify high-risk, high-cost patients prospectively.

Most current risk assessment methods use claims data or structured data. Clinical narrative reports contain rich information, however, are not tapped.

One possible solution is to combine structured data with data extracted from free-text using NLP, so we can identify the target patient population, and then employ machine learning methods like classification techniques and other probabilistic models to stratify patients, and then finally, we can provide CDS and make recommendations based on clinical data.

We give a brief summary and discuss the potential opportunities and also challenges in this field.

Here are Dr. Bates and Dr. Middleton's 10 commandments for effective CDS. As we see, there are many opportunities to apply NLP to enhance CDS.

We need to remember little things can make a big difference. We need to make the system useful. We also need to make it easy for a clinician to do the right thing.

However, there are many challenges to tackle. Speed is everything. If it took too long to work, it will be useless, so we needed to, in addition to the position recall, we need to remember we need to optimize system performance, particularly speed.

The system should be able to anticipate needs and deliver in real time. The system should bring information from free-text to clinicians at the time they need it. The application also needs to fit into the user's workflow, for example, it can be efficient user interface to present NLP output.

As we mentioned before, standards and system interoperability were important, and could we provide service-oriented architecture to provide centralized NLP service to big, diverse EHR systems.

There is many other issues like encoding organizational issues, diverse clinical domains, users meet the needs of the users with diverse rules, background and needs.

There are also other requirements in the field as well. Simple interventions work best, so we needed to simplify and condense NLP output and make it useful. Importantly, we needed to continue monitor impact, receive feedback and make improvement of those applications.

[Applause.]

DR. LUO: We have time for one or two questions.

DR. HUSER: My name is Vojtech Huser from NIH Clinical Center. Do you try to use the consumer as a source of data and maybe NLP has misspelled drug names from PHR, and even be able to view what is my current medication list on record and add the over-the-counter medications?

DR. ZHOU: This is a great suggestion. We haven't done this yet, but I think it is a very interesting research area.

DR. CARRELL: David Carrell, Group Health.

I am curious, in the CDS area, if you are seeing anything that might be called like a Google effect where we are getting accustomed to just being able to remember a few words related to something either in our e-mail, in our web searches, and expect that that should be the standard of our own performance in order to find things that we need to in the record.

DR. ZHOU: Yes. This is another suggestion, yes. I think the technology is there, we just need to bring those of technologies to retrieve, you know, relevant information, and the information that we really need in real time.

Thank you.

DR. LUO: Thank you again.

[Applause.]

DR. LUO: Our next speaker is Dr. Stephane Meystre.

DR. MEYSTRE: Thank you.

Good morning and thank you for this opportunity. I think this is a really excellent workshop and very nice mix of feedback. What I am going to tell you about for a change is NLP and clinical decision support.

I am from the Department of Biomedical Informatics at the University of Utah. We have collaborations also, many collaborations with the Salt Lake City VA, and so I would like to start about asking a question that several presenters already addressed, about why NLP could be useful, why bother about natural language processing, because there is, on one side, a really fast growth of information that is becoming available, more and more systems use EHRs.

There is more requirements, more electronic documentation. There is also a huge growth in other types of electronic information from different, new investigations, genetic testing, et cetera.

But the problem is that most of this information is not structured and coded, it is narrative text, and the existing structured information, encoded information for most of it is created for administrative and reimbursement purposes, not for clinical care.

So, if we look at the electronic health records, most of its contents is narrative text, it is documents, history and physicals, discharge summaries, orders, progress notes, et cetera, and we have some imaging reports, prescriptions -- this is becoming more structured -- lab results, administrative information that is mostly structured.

So, you see that most of the EHR content is unstructured narrative text, and this is not usable for clinical decision support directly. Clinical decision support needs structured and detailed information, information that is structured using some data model and that is coded using some standard terminologies.

Actually, most EHR content, as I said, is narrative text, and it is unstructured and not accessible for clinical decision support. Also, information needs to be detailed at different levels of granularity, and the only existing or most of the existing structured information in EHRs now is coded for public health statistics or reimbursements like ICD-9-CM or CPT-4, and doesn't allow for enough details and clinical care-oriented coding.

One way to deal with this is to use natural language processing to extract this information from all the narrative text content of electronic health record.

This is for most of it called clinical information extraction. Information extraction involves extracting predefined types of information, so it is not the whole complete analysis of everything that is expressed and mentioned in the structure of the narrative text, but it is focusing on some specific types of information of interest for clinical care.

The development of information extraction has already quite a history in the biomedical domain, clinical domain, but much more in the biomedical sides, the scientific publication side, because of available data. Clinical data, as was mentioned yesterday, is still very difficult to obtain and this availability, and also characteristics of the clinical text make it difficult to do information extraction or clinical information extraction.

Characteristics of clinical text like ungrammatical structures, telegraphic style, a lot of abbreviations and acronyms, and this is becoming even more important when information is manually entered by the healthcare provider, because they want to do it fast, and so they abbreviate, and the abbreviations are often specific to a specialty, to their institution, or even to themselves. So, it is a difficult problem also.

Now, I am going to tell you about two examples of efforts we have done in this domain.

The first one was called the automated problem list, and the problem there -- and this happened a few years ago in Mountain Health Care

-- the problem is that there was an electronic problem list that was available, but it was most of the time incomplete like this one here, and this one has already some entries, often it was just not used at all.

At the same time, there were also efforts that needed information from the problem list, implementation of CPOE, of clinical support, for example, and other applications in here, that really needed a complete, accurate, and timely problem list.

What we did is to develop a system that used natural language processing to extract potential problems from all the narrative text in the electronic health record. This system used different steps that were mentioned, discussing several examples yesterday, started with some pre-processing, detecting, and analyzing the structure of the document, the sections, the sentences, then disambiguating all the ambiguous acronyms and abbreviations I just mentioned.

Then, used MMTx. This was a job

implementation of MetaMap, the Department of the National Library of Medicine here, to map concepts with the UMLS Metathesaurus and then also did some negation detection, and finally did some post-processing, also to take into account and correct for the fact that MMTx and MetaMap are the biomedical text and not clinical text, so the ambiguity was dealt with differently.

For example, very common acronyms and clinical text, like M.D., for example, were understood as mental depression, and this is just an example. So, we added some disambiguation to make this work pretty well and reconcile negation, map to local codes, and then created, so fed back the information in the electronic health record in two formats, HL7 CDA documents that included all the extracted and coded information of the problems along with the sections encoded, and also each individual medical problem using the local information model and terminology in American healthcare.

This information was made available in

EHR, and we implemented it at the LDS Hospital in Salt Lake City, and it looks a little bit like this. I don't know if this is readable. It is not supposed to be readable, but what you see if that there is many more entries here that have some additional information here.

So, what we did is to first make sure that users knew where the information came from, and then we allowed them to change or edit the status of the problem that was automatically proposed here to allow them to assess whether it was a correct current problem or inactive problem or dissolved problem, et cetera, and so we really wanted to have the human in the loop, have the final decision for all information that became officially part of the electronic medical record.

This is something I will mention it again at the end that was really important, and then also to allow them to see and have an idea where the information came from, because sometimes they were really not aware of the information we were proposing.

Then, they could just click on the source button and then see all the source documents with the sentences the problem was extracted from highlighted in red. It is not very visible here, but if you allowed them in a few seconds to check for themselves where the information was extracted.

So, we first tested the natural language processing information extraction part of it and compared also different methods to do it, so this is the system that I am talking about.

We also compared it as a baseline with simple keyword search using all entries in the UMLS Metathesaurus, and also compared them with the individual human reviewers.

There were clinicians who created the reference center for the study, and you see that humans had an advantage. They were more precise. What they found was most of the time correct, but were not as sensitive as our NLP system. This is the advantage we brought, so add some content that humans were missing.

Then, we implemented the system at the LDS Hospital in Salt Lake City in the medical and surgical ICU and the cardiovascular surgery, and evaluated it in randomized, controlled trials, so this was an extrinsic evaluation of the impact of the system on the content of the problem list with control test group and about 250 patients.

In the end, we observed that especially in the intensive care units, we went from a sensitivity, so a proportion of problems that were supposed to be in the problem list, and actually found in the problem list of about 9 percent to 41 percent, and then if we also included problems we proposed that were not assessed by the physician, by the user of the problem list, then, it was almost 78 percent, but the specificity went down a little bit.

You see also that in the cardiovascular surgery units, there was almost no effect, and this is simply because they didn't use it. They told us, oh, yeah, we will use it, we are very interested, that's great, but they didn't use it, and so this final control, this human in the loop didn't work here. We had no impact because of that, but potentially it could go up to 88 percent sensitivity.

Another example is the i2b2 medication challenge in 2009, realizing that there was a strong need for structure information for different reasons, and was observed that the list of problems of medications was often incomplete mostly because of medications that were prescribed in another institution somewhere else, out of the system, or bought over the counter by the patients or simply prescribed before the introduction of the order entry system.

So, we also developed a system to extract all medications that were mentioned in the narrative text, in clinical text. For this challenge, what we focused on first was the medication names obviously, but also detailed information like dosage, route, frequency, duration, and a more difficult one, the reason for the prescription.

For this challenge, it was evaluated in a corpus of a bit more than 500 documents, and the system that we developed included also many different steps than are shown here, so it started with again some analysis of the structure of the documents, sections, and filtering of some sections, for example, we excluded the medication allergy section obviously, because the goal here was to find the medications the patient was taking or had taken some time, detected sentences, tokenized to POS tagging, and then also some disambiguation of some ambiguous acronyms like here MG became milligram, could be magnesium or something else, IV became intravenous, BID was BID because it was not ambiguous in this context.

Then, we extracted the medication names and potential reasons for prescription, so these were mostly diagnosis problems, and filtered some of those problems, and so we had like in this very simple example, lasix 40 mg IV BID to promote diuresis. We have a direct lasix with a

COOE here, and the reason diuresis was a COOE also from the UMLS metathesaurus.

Then, we would analyze the context, so the negation, the experience here with other medications was about the patient or someone else, if it said the patient did not take the medications, also detected allergies, we didn't want to consider a patient taking a medication when it was mentioned the patient is allergic to, penicillin, for example.

Then, extracted all the additional information, the dosage, the route, the frequency, the duration, and in the end, reconciled all this information to end up with some structured entry, for example, an entry like in this example, the name is lasix, the dose is 40 milligrams, the route is intravenous, the frequency is twice a day, duration as I have mentioned in this case, and the reason was diuresis.

When evaluating it for exact matches, so this means what we found, the terms we found corresponded exactly to the reference standard. We had performance, the recall is equivalent to the sensitivity here that range from 17 percent only to about 82 percent for some types of information, precision was much higher, and you see that for some of these categories like the duration and reason for the prescription, the performance was pretty low and actually everyone struggled with it, and even humans struggle with it.

So, for example, we added some manual annotations in our team, and our agreements at the first pass of annotators for duration annotations was only 16 percent or 31 percent for reason for prescription, which is variable, and even at the challenge level, multiple pass of annotations and adjudication of differences, annotations by experts, then, for duration, they only agreed about 40 percent of the time, about 47 percent of the time, and reason 40 percent of the time, so this is a really difficult task not only for NLP.

As a conclusion, I would like to emphasize two aspects that I think are important for using natural language processing for clinical decision support, and more specifically, when using natural language processing to extract clinical information.

This is really for humans, for users to trust the information that is extracted, and what we observed and experienced is that the fact that there is a human that eventually decides if the information is correct, it is very important, and also, to allow this human to see where this information came from and maybe how it was extracted. So, this also influences some discussion we had yesterday about what methods to use and for acceptance, and in this case, it should be something that is more transparent and allows explaining where and how the information was extracted.

With this, we observed when we implemented the system, we had very quickly a lot of trust from the physicians working in the ICU, and they even liked it so much that they ended up dictating small notes, you have the problems put automatically in the list instead of entering it manually.

So, this was really important, and also another one is that with this in mind, then, performance is not as important. It should be fast, that's true, it should have acceptable accuracy, but even if it just adds a little to what is already available, it is already good.

Thank you.

[Applause.]

DR. PAYNE: Excellent work. One comment and one question. The comment is it is interesting that the ICU doctors dictated the notes to get them into the problem list. It tells you a little bit about what they are comfortable with.

The question is why would they do this, why would they be interested in adding to the problem list?

DR. MEYSTRE: Well, it is more, several

reasons, but it is obviously in the ICU it is because they saw interest in the problem lists from the beginning. Their Chair was very interested in making the problem list the central components in their workflow in their care.

So, obviously, to make it useful it needs to be used, and you see the difference with the Cardiovascular Surgery Department who implemented it, it has absolutely no impact.

Even if it performed pretty well, but no one used it, because they still continued their usual workflow that was really not relying on the problem list, but in the ICU, they were working on it, and they had a strong effort to make it more important to rely on entries in the problem list to base all their discussions, case discussions, et cetera on the problem list, et cetera.

DR. MENDELSON: David Mendelson, Mount Sinai, New York.

So, to work that out a little bit further, this validation step that you have mentioned before, probably you can't rely on these systems as least at this point in time completely, and what was the resistance in the Cardiovascular Unit, was it that it would take more time?

We have had some experience in radiology using CAD systems. Some people won't touch them because it takes more time. So, I would be interested in your views on how this was seen in your site, and how it might evolve.

DR. MEYSTRE: Initially, we had some pilots evaluation with a few years of select users in cardiovascular surgery, and they didn't complain about that, because for them, it meant one or two clicks, so it was really fast, but the main problem, as I mentioned, is that they simply had no incentive to use the problem list at all, electronic problem list, so that is why they didn't use it.

MR. SHANKAR: Just curiosity, as a student I used to wonder if I say you have one expert in the loop, if I see a patient, if I see a note that says cirrhosis of liver, as a junior doctor I would say that is anasarca, there is esophageal varices, there is chronic cirrhosis of the liver, so I would come up with at least five, six, seven different diagnoses possible in that.

Which one would you pick on, do you want all of them in the problem list?

DR. MEYSTRE: What we did is to filter all these problems to first make sure that there was no duplicates, and so we did an analysis at different levels of granularity using standard terminologies, relations, et cetera, to make sure that, for example, if there was already diabetes in the list, and we found diabetes mellitus type 2, then, we would not add it, because we would consider that it was already present.

We only included some types of clinical permission mostly diagnosis, so we had a long list of most frequent diagnoses in the domains we implemented it in, cardiovascular surgery and general medical, surgical, ICU, mostly cardiovascular. So, it wasn't everything, not all findings, symptoms, et cetera.

DR/ ZHOU: What is the time in the clinical notes, like three months ago, six months ago, a year ago?

MEYSTRE: When we learned that it was done in real time, so as soon as provider stored, recorded, or saved document --

DR. ZHOU: So, it was like 10 years ago, the notes, it is not --

DR. MEYSTRE: It was only prospective, only prospective, so as soon as the patient was hospitalized, entered in the ICU, everything was new, all new documents were processed and populated. It was only prospective.

MR. JAGANNATHAN: One more quick question here. Did you consider using nurse practitioners or nurses to wet the problem list, so that it is a little more cleaner in the workflow?

DR. MEYSTRE: No, we did not, because the problem we had it was almost only used by physicians. It was really more like a doctor's problem.

MR. JAGANNATHAN: What about medication reconciliation?

DR. MEYSTRE: They were not in the problem list.

MR. JAGANNATHAN: I know.

DR. MEYSTRE: Oh, you mean for the other one?

MR. JAGANNATHAN: Yes.

DR. MEYSTRE: Yes, in general, you mean.

MR. JAGANNATHAN: Yes.

DR. MEYSTRE: Not really, but it could be a good idea especially in some environments where nurse practitioners do most of the contact with the patient, it would make sense, yes.

DR. LUO: Thanks, Dr. Meystre.

[Applause.]

DR. LUO: Our next speaker is Dr. Mark Musen from Stanford University.

DR. MUSEN: Thank you, James.

Thank you. I really want to thank NLM and NIBIB for bringing us all together to have this interesting conversation. I want to start by confessing that there was a time when I was a non-believer. When I was a graduate student in the early 1980s, Terry Winograd had just made a splash with Volume 1 of his book on natural language processing, which he called Syntax, and it immediately abandoned the idea of Volume 2 when he said, well, we will never figure out semantics, so we should just give up, and that is when he moved into computer interaction.

Frankly, if you had asked me then if we would ever reach the kinds of capabilities that we have now in NLP, I would have been very doubtful, and as someone who has spent his career mainly dealing with decision support, actually, not dealing with natural language processing as much as I can at all, I am actually very impressed where NLP has come, and, in fact, I am actually frustrated where CDS has gone particularly in the commercial sector.

I think what I would like to do in the next few minutes is talk a little bit about where CDS exists as is actually deployed in the real world where I think NLP can be very helpful.

I was very, very inspired by Bob's keynote where he went through the large gamut of opportunity that exists in the area of clinical decision support and the ways in which our technology can contribute to better medical decisionmaking.

Then, I began to think about really where are we in the world now, and, frankly, as my slide suggests, at least in the commercial sector, and they are largely in the academic sector, as well, we tend to look for our keys where the light is and claim great success when really we are missing out on great opportunities as ways of really improving health care and health.

When you look at what the vendors talk about in terms of decision support, they advertise rules. This is actually a rule stolen from some product literature that Cerner dishes out, which talks about the kinds of rules that you can write in their particular language for handling rules.

We all know about HL7, promoting Arden syntax as another rule-based language, and, frankly, these rules provide the opportunity to say if some bad situation exists, then, do something about that bad situation in general.

This is a technology that has existed since the 1970s, 40 years ago Clem McDonald showed brilliantly how this kind of technology can really have a tremendous effect on affecting a lot of the kinds of problems that routinely affect people in ambulatory care and averting a lot of the kinds of situations that most of us are most concerned about.

We can deal with drug-drug interactions, we can make sure that patients who are allergic to drugs don't get drugs they shouldn't get. We can suggest alternative medications that might be cheaper at the time of order entry. We can make remarks about abnormal lab results. There are opportunities for immunizations and preventive services that we bring to the attention of providers.

This is all wonderful stuff, all stuff that we showed 40 years ago really is at the heart of rule-based systems when they are used really well in clinical care.

The problem is we are continuing to promote these kinds of systems, and they are promoting them to do the kinds of things they do well, while we really ignore the possibility of alternative forms of decision support that actually can address other kinds of pressing problems that we just don't look at at all.

We are really good at squinting at clinical problems and saying this is where rules might be useful, but we blithely ignore all the problems of rule-based systems that emerged when work was done in the 1980s, as rule-based systems expanded, we realized how they were brittle, how they really didn't address problems at the edge of their capabilities, and most important, how we learned how rule bases when they became large were absolutely impossible to develop, how quality assurance was impossible, and how we really needed ways of structuring knowledge in a much more coherent way, not only to make our systems more intelligent, but to make us more intelligent and be able to maintain them and to repair them.

Frankly, as the population ages, and we start talking about what the next generation of care is going to require, we are not very good at dealing with what the next kind of decision support systems need to provide. They need to be able to deal with guidelines as Bob suggested.

They need to be able to deal with treatment as it falls over time, they have to deal with interventions that are based on previous response to treatments, so if someone did well previously on some drug, it might be good to continue that kind of drug.

If someone did poorly, we know we want to try some alternatives, and most important, as the population ages, we have to recognize that 60 percent of patients over age 50 have more than one disease, and in the setting of polypharmacy, in the setting of multimorbidity, even our existing clinical practice guidelines, which were all developed in clinical trials, where patients with multimorbidity were excluded in the first place, we have trouble adopting our guidelines in ways that actually make sense were the kinds of patients we are actually treating in the clinic.

Clinical care, as it really exists, is messy. It deals with patients who have very complex problems, many complex problems, guidelines that don't always apply, and yet we are really good at thinking about those rules that tell us not to give drugs to which patients are allergic.

What we really need to be thinking about is the next generation of clinical decision support that can start to do things that are more interesting.

I know this sounds immodest, but let me at least give you an example of the kind of things that we have been working on over the years. One thing, because I think it gives a good example, too, because a lot of it was funded by NLM, so I feel obligated to talk about it.

We have a system at Stanford that Blackford alluded to earlier this morning called EON, which provides a foundation on guideline-based decision support which my colleague, Mary Goldstein, at Stanford has used to deploy in about nine different VA medical centers for a variety of chronic diseases to which guideline-based care makes sense.

This shows you actually an experimental front end for the ATHENA hypertension system, which sits on top of CPRS, which is on top of VistA at these VA medical centers, and what VistA and ATHENA allow one to do together is to look at information about patients who have hypertension and to apply that Joint National Commission guideline that Blackford mentioned, or maybe it was Frank, to automatically make suggestions about how patients might be treated if the doctor wants to assume that the guideline makes sense in that particular scenario.

How does ATHENA work? Well, it has a representation of the guideline stores a knowledge base, it has the patient data that are available to the VistA electronic health record system, and assuming that all the coded data are all you need, brings that information together to make recommendations, such as, for example, consider adding an ACE inhibitor because there is a compelling indication in this particular case of heart failure.

ATHENA does this because we have under the hood, and it is always good to show what is under the hood at this kind of a meeting, an ontology that describes what can we expect to find in typical clinical practice guidelines.

This ontology provides a framework that says guidelines will have things such as recommendation specification, it will have an action specification, it will have an algorithm that will suggest how various tasks get implemented over time, much in the way that Bob suggested in his keynote, and it will provide particular properties of those various entities in a way that allows our system to automatically acquire information about the various guidelines that might be instantiated using this generic ontology.

So, the ontology might, for example, suggest what is the information we need to describe the JNC7 guideline, what are the ways in which the guidelines apply, so we know they are actually meeting its goals, what are the kinds of drugs we want to use, what are the kinds of interventions you want to consider, and so on.

And because we can dry out grass, we can say very explicitly what is the temporal order of interventions that make sense if we want to treat patients in accordance with his rather complicated guideline.

This is not the kind of a specification that a situation actual can specify in a very clean way. This is not the kind of specification that deals with one situation and one action.

Rather, it deals with multiple situations and multiple actions that unfold in a rather complicated temporal sequence.

But once you have this kind of representation, once you have the data that are available through VistA, once you have the patient who is being treated according to the JNC7 guideline, then, you have the basis for having an automated system that can go well beyond identifying very simple problems, but can manage complex patients with multiple complicating issues in the area of hypertension over time.

That is the good news. The bad news is that to do this actually requires a lot of information that is not accessible through the coded data that we are able to bring in through VistA. So, what do we miss? We miss understanding what are the specific patient preferences that might cause the patient to choose what action over another if both actions are relatively indicated at the same level.

What are the provider preferences, what are the organizational preferences that might make one action more effective over another in a particular setting? What are all the intangibles? What is the patient's social support?

What is the patient's transportation situation, can the patient even reliably get to the clinic for evaluation? What over-the-counter medications might the patient be taking? What drugs might be prescribed elsewhere that we don't know about? What comorbidity exists?

In fact, what is the clinical practice guideline or guidelines according to which the patient is being treated? That kind of thing is all in the narrative text, and that information is not accessible to ATHENA in any way, shape, or form, and as you have just heard, a lot of what we assume is going to be available through the coded information may or may not be there, but leave that for another day.

We have gotten really good at taking the

ATHENA model and replicating it, so my colleagues at the VA at Palo Alto, led by Mary Goldstein, have taken versions of ATHENA that take the JNC7 hypertension version and replicate it for use in a guideline that treats heart failure, and have replicated it for entering information about management of hyperlipidemia and diabetes and chronic kidney disease and management of opioid therapy, and all that is really wonderful, but all of these are stovepipe systems, and as I said a moment ago, our biggest problem is dealing with comorbidity.

What we are really trying to do now is identify mechanisms that make sense for taking the complexities that we have encountered when we have to manage patients in accordance with one clinical practice guidelines and begin to identify ways in which we can bring other guidelines into the decisionmaking process.

So, with a contract that was funded through ARRA money that we recently received from NLM, we are starting to look at the ways in which multiple guidelines can be administered together.

I know the model right now is very simple. We are very excited about the possibility of looking at the problem as it emerges when you take the multiple guidelines and working in the code using the EON technology, the various ATHENA versions, run them on a particular patient who has comorbidity conditions and then think about how can we apply those guidelines, how can we consolidate the various recommendations that the guidelines individually would recommend, and then how can we identify potential interactions among those guidelines, how can we repair those interactions, and ultimately, how can we prioritize recommendations when multiple guidelines run together and obviously, none of these guidelines were created with the idea of anticipating all the possible side effects that might occur when the guidelines for other diseases are being run simultaneously.

I won't say that we have the solution here, but I think this is a really exciting area of research, and points to the complexity of the clinical enterprise, and the clinical problems that decision support needs to address, that certainly we have a long way to go in being able to deal with.

Dealing with these multiple comorbidities is hard, because as I said earlier, all the clinical trials typically will exclude patients who have other diseases, and rarely is there evidence that tells us what to do when patients have more than one condition.

Understanding the relative effects of comorbidities on functional status is really an important nuance. We may need to be able to appreciate that, yes, treating a patient in accordance with one guideline might help one problem, but might hurt some other problem, and being able to assess that tradeoff and know what to do about that tradeoff is a difficult decision problem.

That is not something that is built into any of the existing guideline structures, but when we try to do that, those decision models can become explosively complicated as we deal with all those possible contingencies.

Obviously, what we really need are ways of being able to manage the complexity in a computationally reasonable sense, and the decisionmaking basically, in the absence of any kind of formal evidence, which usually is the case when you deal with patients with multiple morbidity, has to be informed from other sources.

As other authors have said today, what it will inform that kind of decisionmaking will fundamentally be the kind of information that that we can glean from electronic health records, mainly information that we can glean in the form of NLP.

So, in order to basically have the kinds of health systems that can offer evidence-based care, not for simple situations, actual situations, not even for individual guidelines, but for the kinds of patients who typically will be overrunning the healthcare system for the next

30, 40, forever years.

We basically need to be able to incorporate all the non-coded data that exists in the EHR, we need to be able to identify from historical records, patients who are similar in their comorbidity, similar in their treatment situations, and try to understand, not in the context of controlled trials, but in the context of our experience what has taken place from which we can learn in order to make decisions about these kinds of complex patients, which are the norm in clinical practice, basically, what can we do to build the kind of clinical decision support systems that will address all the situations in which patients and providers need advice, not just those that can be framed in terms of the kinds of simple rules that are so pervasive now in the kinds of systems that are available commercially.

Last summer, there were a number of workshops that were held by the Institute of Medicine with a lot of encouragement from the ONC and other agencies on the idea of a learning healthcare system that would be able to learn from the evidence of previous practice what might be reasonable approaches, but frankly, there is no evidence, and to apply those kinds of inferences in the care of patients where there may not be a clinical practice guideline, but where providers are seeking more than just an intuition about what might be the best kind of care.

I think as we think about the complexities of the clinical arena, as we think about how we want to apply evidence-based practice to the best of our ability, and how we want to fill in the gaps, I think in working toward this notion of a learning healthcare system, we can see a lot of benefit from traditional clinical decision support systems, the kinds of clinical decision support systems that can incorporate the complexities of the guidelines that folks have talked about this morning, but also understanding how similar patients have been treated in an effort come up with the kinds of care plans that can take advantage of as much information as possible in dealing with patients who have the kinds of complexities which used to be considered rare, but now, in the current aging population, are just what we deal with all the time.

Thanks.

[Applause.]

DR. LUO: I would like to invite the speakers up to the front.

DR. SONNENBERG: With regard to looking at the experience with similar patients with the same condition or combinations, I would like to ask you to elaborate on how you use that information. Do you assume that those patients were treated correctly, or do you try to separate them out according to the different ways they are treated and look at their outcomes?

DR. MUSEN: I should actually clarify that this is not research that we are currently doing, it is that we are very excited about moving in this direction. I think what one has to do is look at a variety of parameters including the characteristics of the patient in terms of observed signs and symptoms and problems, intervention, intervention history, and try to identify what is the distance, if you will, between the patient to be treated and those historical patients who might be available in the record.

Frankly, some of the best work that I know of in this area was done 30 years ago by Glen Reynolds, when he looked at the problem of reasoning about patients whose descriptions were available in the clinical literature, not necessarily through clinical trials.

I think we can learn a lot from that work and, frankly, with the ability to get access to a host of information now through NOP. I think we now have a rich set of data that we can now use to try to do that kind of work. I don't know if any of us has really done it very well yet.

MR. WAGHOLIKAR: I understand that EON,

only a part of the guideline can be implemented, because you are dealing with coded data, and I understand part of the guideline can't be implemented because the data is not available in the coded form.

Can you estimate roughly what part, I mean the fraction of the guideline which can't be implemented because of the need for coded data, because data is locked up in the free text?

DR. MUSEN: It actually turns out that the vast majority of the data that are needed for JNC7 can ultimately be inferred from coded information, not in a way which guarantees that every possible inference that might be available in the record could be applied to the guideline, but in a way that is reasonable.

I am now trying to remember, I think it was Frank who talked about it earlier, we can take information from the coded record and use it to infer abstractions that may allow us then to predicate other actions that are specified in the guideline.

So, even though we may not have a problem code for something that the guideline might refer to, we can perform inference based on what we do know to make a guess as to whether that situation is there, and what is important, too, is that these kinds of sophisticated reasoners are able to inform the clinician how the inference was made and what the certainty is with which that conclusion was made.

These systems are not perfect, but I think it is really important to recognize that providers simply are overloaded with guideline suggestions and have difficulty just remembering what guidelines might suggest in certain circumstances, and the more that we can provide these kinds of decision support systems, the better off they will be in being able to determine whether the guideline might apply, and if so, how they want to apply it.

I think just to clarify, when Bill Tierney and his group tried to encode a guideline for heart failure, and described that work in

JAMIA about 20 years ago, they published a really wonderful paper that goes into great explicit detail about the information the guideline required, how it was not necessarily directly available in the coded record and what kind of inferences they needed to make in order to provide that information.

DR. LUO: Thanks, Dr. Musen.

[Applause.]

DR. LUO: We will move to the next phase, and we have about 25 minutes for the panel discussion.

MR. JAGANNATHAN: I have a quick question for Dr. Musen. I remember participating in a technology expert panel on quality measures, and a long-term care physician said that any patient chart which has more than 10 medications is guaranteed to have a drug-drug interaction.

As you look at all those retired people and the baby boomers, they all have a long list of problems. Are you looking at just two clinical guidelines or are you looking at like 10 different problems that these patients have which are all being treated with medications?

DR. MUSEN: That's a great question. In fact, we are looking both at the problem of the guidelines interacting and suggesting interventions that may be at odds, but also looking overall at the polypharmacy question, particularly in the elderly, that is a really serious problem, and then it becomes a matter of trying to identify what are the preferences of the patients and the providers in terms of how to trim such a list in order to give just those drugs that are the most reasonable to give, recognizing for exactly the purposes you point out, that it may be dangerous to give those long lists of drugs in the guise of settings where actually a strict application of the guidelines might suggest such enormous numbers of medications.

DR. HUSER: I have a question maybe primarily to Mark Musen, but then to the other panelists. We have heard from the keynote that we don't have a guideline execution engine out there, the GLIF project perhaps was close. You mentioned that we have a rule-based language like Arden syntax.

Where do you see the magic of what is sometimes called Text-Networks models like ATHENA is using, where do you see kind of where they go beyond the rules, where the rules become brittle, and maybe put this in a perspective of I am now dealing with pharmacogenomics and folks writing from a genomics dosing guidelines?

As an informatician who is supposed to be good at encoding knowledge coming up, I can now have a perfect opportunity to shape them towards a particular standard, and the best thing I can tell them is Arden syntax, and there is also this boundary where one pharmacogenomics guideline ends for one drug, but actually some of the genes affect other drugs, so they have problems of the boundary between guidelines.

So, maybe to summarize, so this problem of we don't have a standard out there, and the text-network models bring some magic, how do we address it?

DR. MUSEN: Well, I won't address the magic, but I guess the first question was where is the inflection point where rule-based systems become difficult to manage, and that is a hard one to answer.

Obviously, when it's one, it is very easy. When the rules are such that they do not interact, it is very easy. The problem occurs when you have to use multiple interacting rules in order to deal with complex decisionmaking processes.

I think there was a paper in AMIA about seven years ago when attempts were made at Columbia to be able to implement, in a rule-based framework using Arden, I believe, protocols for management of cancer chemotherapy.

The bottom line was that those rules individually made sense, but collectively, actually were very problematic, because of the complexities of the interactions.

When you look at what happened in the 1980s, all the excitement of rule-based systems, all of the industrial applications of rule-based systems got great press early on, and ultimately, there was an enormous retrenchment because people found that as you had more than 20 or 30 or 100 rules to represent complex situations, the management of those rules became absolutely impossible, and so as we think about the idea of automating tasks for the purposes of decision support, then, I think it is much more useful to be thinking about tasks as abstractions, which we will then implement using some programming framework, which could actually have rules under the hood, but where the cognitive tasks for the developers to be thinking about things at the highest level of abstraction, and not having to be worried about how individual production rules interact, I think Arden has an important role to play, essential role to play in all of the situations where one is dealing with, say, allergies or drug-drug interactions or contraindicated drugs.

In those situations, that can be stovepiped and dealt with individually. Individual rules are wonderful, and although a lot of people got into trouble with large rule-based systems in the '80s, there were some amazing successes with small rule-based systems in the '70s.

So, my argument would be one has to know when to use the right technology.

DR. GREENES: I would just amplify that. A guideline, to actually execute it, patients don't come ready made to fit a guideline. They come at a certain point in time, so there is an eligibility point at which maybe a guideline might apply to them.

At that point, there is some situation action rule that might apply. So, we tend to think of guidelines as, you know, if we had a standard representation of the test network, that you could take individual points of those and turn them into situation actual rules, so now you have this knowledge base of Arden rules or whatever.

I pick Arden because it's a standard Y and N, another expression language. But those then can co-reference the guidelines that they are part of or the goals directed systems that they are part of, so that, you know, if you now need to refer to the framework, the more complex management of that problem, my GPS model, you know, you could basically have a goal that that rule is part of.

I don't have the complete answer, but, you know, rather than saying, you know, it's rules versus guidelines, some rules, I mean to actually actuate a guideline, you have to decompose it into rules and order sets and other things that you are actually going to implement.

MR. SHANKAR: To continue with Dr. Stephane, there are some electronic medical record system which enforce medication association with the problem list. So, for example, if you prescribe the data, there should be a primary diagnosis of some cause for that, so if we just put everything for, say, chronic liver failure, or chronic cirrhosis of liver, and all the data, statement report will have hyponatremia, esophageal varices, should we associate them like that, or would you rather prefer to associate each of the medications for specific causes or specific etiologies, like say edema or --

DR. MUSEN: In the front of the room I am having trouble hearing you because of the echo. Perhaps you could speak a little more slowly.

MR. SHANKAR: Okay. The first question is I think the electronic medical record system, they enforce medication association with the diagnosis, whether, say, can associate all the medications I have for the chronic liver disease patient, or the chronic cirrhosis patient, or I can say the biggest part is food retention, so how do you balance it?

DR. MUSEN: I mean certainly those kinds of associations are the perfect kinds of things to encourage in a rule-based approach. Where you get into trouble are in the kinds of guidelines that deal with more nuance situations, so just to pick one, which we like to talk about a lot, because it causes a lot of trouble when we tried to do this at Stanford, we were coding guidelines for management of HIV, which would sound like a rule saying that if this is the second episode of anemia that has followed the use of these drugs, then, consider substitution of these other drugs.

That kind of a situation, which sounds like a rule, when you get down to it, is very hard to encode in a rule-based framework, because you have to figure out what does it mean to have anemia, what does it mean to have the second episode of anemia, how do you figure out whether that anemia followed some intervention which now you want to be concerned about in prescribing in the future, and when you have to have that kind of reasoning, suddenly, Arden is not a good hammer for that particular nail.

What I am saying is not that there is

only one hammer out here, and it is really complicated, what I am saying is that there is a large suite of things we want to do in clinical decision support, some of which are very simple and very repetitive, and beautifully done in a situation action framework, and other things, which are hard and which are becoming increasingly more prevalent as patients frankly get older and sicker, we are going to need to turn to other kinds of decision support technology in order to have the kind of influence in the world that we want.

As Bob alluded, many of us over the years have been dealing with ways of trying to represent these more complicated guidelines to be able to deal with these situations that elude the simple rule-based framework.

There is not a standard way of doing it. There is a lot of experience and different approaches, and I think we need a lot more of that, we need to learn a lot more from that, so that we can have more sophisticated ways of reasoning, and then when we realize we are not there yet, when we are falling off the cliff, we need to be able to fall back on the historical record to figure out, well, what is the next best kind of information that we can get to help us.

DR. MEYSTRE: Did you want to have more answer? I thought you had directed the question to me originally.

DR. MUSEN: If so, I apologize.

MR. SHANKAR: That's fine. Very good. I didn't understand very clearly what you were asking, so I am very glad. It's just that with extended, you would like the diagnosis on a primary list for purposes of associating with medications.

DR. MEYSTRE: Correct me if I don't understand, but this is about organizing something like a list of problems that includes also medications and organizing them, ordering them with specific problems. So, is your question about the functionality or about automatically providing this type of information?

MR. SHANKAR: Functionality.

DR. MEYSTRE: So, that is something that we heard several times, and we didn't really investigate it, but the need or the use of the problem list varies a lot between users and specialties.

Some use it as a very extensive differential diagnosis tool and would like to include all the relevant information in it using that as the main table of content for the whole EHR, which is the idea of the problem with the medical record.

Other ones just use it as like to-do lists with a level of extraction that makes sense for them, but obviously doesn't include a lot of information, relevant information as I mentioned.

Sometimes the medication is included. Sometimes the problem becomes something like check or verify medication effects or remove or stop medication, you know, just as with the dates, so that when they look at the problem list again, they don't forget. They will look at that and ask the patient if the medication works as expected, and if they have to change it or re-evaluate it, it really varies.

DR. ZHOU: I have a few comments. One is like how we are using our P2, mining the data like to figure out associations are current, medications and prevalence, like mining a large amount of clinical data and to find that association, return medication.

One problem of this you can spend a lot, they can spend a lot false positive. The false positive like if you want to trigger like those two medications, this is two thing co-occurrent, you know, happen together, it is not two, so need a very, very sophisticated standard models to handle this problem.

I think you can look at some reference published at the firm in Columbia, of Rechon Hurtsek [ph] paper. Another is about EHR function. We have MAPLE project, which you can use, because the problem list is not complete, so you can actually look at the patient problem list, is the patient taking some medication for diabetes, and the clinical didn't support it for the efficiency, would you like to add a diagnosis of diabetes on the problem list.

DR. SIEGEL: I have I guess partly a philosophic question, but it is something that I have struggled with, with some of the work that I have done looking at the electronic medical record.

That is, if I get admitted to the hospital, I really want the best doctor possible to take care of me, but there is no consumer reports for these decision support systems, and so as time goes on, I mean it was interesting to see the version from Cerner and from some of the established electronic medical record companies.

There is going to be more and more third parties who are going to have software that is going to sit on top of our electronic medical record, but how do I know, how does one determine the quality of a system like this, is there kind of a test set like Consumer Reports might have you kind of know what the answers are, and you just run your clinical decision support system on?

So, how would I determine whether or not the system you guys are using or anyone else is really any better, and when you train it, how do you know that your system is better next year than it was this year?

DR. LUO: Decision support system to determine the doctors, hospitals, and in a different level, right?

DR. SIEGEL: No, not so much to determine quality of doctors, I just meant in general, in other words, clinical decision support systems, how do you know whether they are doing what they purport to do, is there a Consumer Reports for that, how do you guys know when you develop it, that it is any better, and how do you know whether it is getting smart or not?

DR. GREENES: Do you work for the FDA?

[Laughter.]

DR. SIEGEL: No, but I am sure they would like to know the answer.

DR. LUO: It looks like an interesting question.

DR. MEYSTRE: I think the problem is larger than that, because even if you have a great performing clinical decision support system evaluated intrinsically, you know, in isolation with a specific standard set of decisions and data, et cetera, how you implement it in the healthcare system, the data that it has access to it, and this is precisely the problem that we have mentioned many times here, it makes all the difference, I think.

The most informative evaluation would probably be some extrinsic evaluation, the impact of such a system in a specific healthcare institution.

DR. SIEGEL: Yeah, that's kind of tricky, though, isn't it? I mean I can tell you who has the strongest test engine, I can tell you who plays Jeopardy the best essentially, because I know the answers to the questions but how can I determine what the quality of an information decision support system is.

DR. GREENES: One of the ongoing issues for a lot of APPS and other, you know, EHR improvements is to have a robust de-identified patient data set.

DR. SIEGEL: Right. That is what I am suggesting, so is there a data set --

DR. GREENES: There is an existing move afoot to create such a thing. Did you want to comment?

DR. MIDDLETON: The other thought is that I think the evaluation component, of course, is always important. One might imagine a test corpus just like the NLP folks, of course, iteratively test their engines and whatnot.

We have also come up with a measure that can be used in clinical practice. What is the ultimate correct outcome for decision support is the outcome, not necessarily the process change, as Jim Walker was talking about, you know, not necessarily he is ordering the test perhaps, but the clinical outcome more performance.

So, we have a measure called number needed to remind, which looks at clinical performance, and thus can evaluate the same CDS engine in different clinical context where there might be different reasons why clinical performance isn't being achieved, but can give you a differential measure.

DR. SIEGEL: But can you do the other side, can you evaluate multiple different clinical support engines in the same clinical context essentially, is it not possible to be able to compare one versus the other on, as Bob says, a corpus of data that one might be able to essentially compare one to the other arm.

DR. MIDDLETON: You might find differentiation with this number needed to remind idea, but I agree a separate test corpus could also be helpful.

DR. SIEGEL: If there is three vendors who all claim that they are the best, how do I know which one to buy?

DR. GREENES: At one level, you want to just know that it is performing the correct logic, so that is the simplest. You know, whether it is effective, you know, how it is implemented, how it is triggered, you know, how it integrates with the care flow, I don't know how you would test that other than what Blackford is suggesting.

DR. MUSEN: I think there are two issues here, and that is, if there is, for example, evidence that has been used to formulate a guideline, then, in an artificial laboratory setting, you can determine whether the decision support system is allowing advice that follows the guideline. That is the process that Bob alluded to as being important, but not really where the money is.

On the other hand, the only outcomes, it is very, very hard, because it is so idiosyncratic based on the particular patient, the particular situation, and even if you are doing a trial where you are looking at a decision support system on a cadre of patients, the time it takes, certainly in the area of chronic disease, to be able to know whether you have reached outcomes that are different is so long that you are in a situation where really the best that you can do is really say I have faith in the guideline representing the evidence, and therefore, I will go with process because it is expedient even though it may not be the metric that really matters the most.

DR. LUO: The guideline has gone through some of the comparison data, that evasion process, so if there is guideline, I think that will be automated, go to standard for some of the disease. If not, then, there is another process.

DR. PAYNE: One last comment. When I was with the VA, there was an enormous difference in how different sites used the same tools, how much energy they put into using them. So, it wasn't really the vendor so much as other factors, and I found that to be true for the commercial vendors with which we work, as well.

They have incredible sets of tools, some of them should be more incredible, but they are not fully utilized, and the sites are working on other things. So, it would be a tricky thing to do to say this vendor is stronger or weaker than that vendor unless you were somehow magically to be able to control for the people who are using it, their methodology, how much energy they put into it, and so on.

MR. WEIDA: Hi. I am Tony Weida from Apelon.

More than one speaker today has noted the use of standard terminologies for communicating information from NLP systems to CDS systems. I want to emphasize that what we really need are value sets, that is, terminology subsets to focus the interaction.

On the one hand, this gives guidance to the NLP systems about what we are looking for in terms of decision support, and then from a CDS perspective, it indicates what we need to find in order to render decisions.

So, when Bob spoke earlier about the difference between quality measures, which are retrospective, and CDS, which is prospective, I thought about the fact that an increasing focus is being paid today to quality measures, and as a result, more and more value sets are being developed and published, for example, we are working on developing behavioral health value sets in support of quality measures.

So, my question for Bob and for everybody is considering the differences between quality measurement and CDS, do you feel that the conceptual basis, that is, the kinds of concepts we find in standard terminologies are sufficiently similar between quality measurement and clinical decision support, that, indeed, we can develop and reuse them for both purposes.

DR. GREENES: Well, I think they are the same data classes, certainly for proactive decisions where you may want to do things that are not measured in the numerator or denominator of a quality measure.

So, you may want to check on temporal relationships, you know, existence of things, comorbid conditions, other kinds of things like that.

Some of those basically aim to try to make the decision support as targeted and as specific as you can. The quality measure is kind of a coarser net probably doesn't have all those refinements in it, and so I think when you are trying to deliver it and avoid alert fatigue and all these other thing that don't apply to your patient, nonspecific recommendations, your obligation is really to try to pull up more data.

So, I don't think it is a difference in classes that we haven't thought of that aren't in your value sets, but it is probably just going to need more parameters within those classes.

DR. MEYSTRE: I have a quick comment to make on that. I see the distinction as quality measures pretty much encodes the clinical guidelines, but as clinical decision support goes, it allows you to not only measure, given that you have a particular diagnosis, you need to treat them in a certain way, but as clinical decision support says how can I make a certain diagnosis, maybe you can help with that, too, which is not the focus of quality measures. At least that is how I have been looking at it.

MR. JAGANNATHAN: I have a comment, and this is based on the experience we had also trying to match what can be found in the narrative text using information extraction and what is eventually used and needed for calculating a problem while you are still driving clinical decision support or other applications.

What we found is that there is actually a mismatch often. Well, you can most of the time find something that matches, but what is most of the time missing is connections between the concepts.

So, most of the terminologies we use are just terminologies that are not really well organized and complete ontology. We have multiple relations that we can use to navigate different levels of granularity and details about a concept.

So, there is some of that, but we need to actually add some logic to it to navigate these terminologies, and this is probably what I think would be the most useful to add connections between the different concepts besides what is already existing and already very rich, but not complete yet, like in the UMLS Metathesasurus, for example.

DR. ZHOU: When we deal with the medication, we find like we use AXONOM. AXONOM don't have drug classes, so a lot of clinical decision support actually is based on classes. They need more work to build to this kind of standard.

DR. LUO: Thanks, everyone, for a great discussion, and we will continue the discussion while we have poster sessions during the lunch hour, and we are almost running out of time, so thank all the speakers.

[Applause.]

DR. LUO: We are coming back at 1:40. We still have 40 minutes.

[Luncheon break 1:01 to 1:52 p.m.]

DR. LUO: This is the afternoon session. Before we start, Dr. Lindberg has an announcement.

DR. LINDBERG: I won't delay your session, but I want to present to you one more insurmountable opportunity, namely, if you have some time left when your meeting here ends, there is in the library, which is to say the building you can see nearest, an exhibition called "Native Concepts of Health and Illness," which we have worked pretty hard on.

It relates to American Indians, Native Alaskans, Native Hawaiians, and what we are after is presenting the mental model, so to speak, which they have of health and illness. Arguing that, you may or may not agree with me, that if the doctor and the patient don't have the same idea in mind, the therapy doesn't work anyway, because the patient doesn't follow it.

So, these people have their own ideas. Some of them are quite worthwhile showing. That is why we have the exhibition. You would be very welcome over there, it's interactive. You can find your own way through, and, please join us if you get the chance.

Thank you.

Panel 3: Stakeholder Perspectives

Moderator: Steve Lohr

MR. LOHR: Perhaps just a word. My name is Steve Lohr. I am a technology reporter for The New York Times. The reason I am here is, just to explain, I have covered electronic health records for a number of years and have reported a fair bit on artificial intelligence techniques including natural language processing for uses in a number of fields, and so I am interested in the subject.

This afternoon's panel is entitled "Stakeholder," and we have four distinguished panelists, three from our industry stakeholder for Aetna, Siemens, and IBM, and our initial speaker is Jacob Reider, who is a hybrid in terms of his perspective, his family physician, and for several years, and is now senior policy advisor to the Obama Administration in the Office of the National Coordinator.

Jacob got into industry in a sense through complaints, I gather. He was a practicing physician, as well as an IT guru, and on his blog I think in 2004 he complained about the usability problems of an electronic health record.

At first, the vendor called his boss and tried to shut him down, and that didn't work, because he wouldn't take down his blog, and after that, they engaged him in conversation, and finally, they hired him as the Chief Medical Director for Misys, which later became AllScripts when the two companies merged.

Last year, just after he had decided to join the Obama administration in an interview, he was asked for his role as a senior advisor at the Office of the National Coordinator, and he said, "My role is to listen to the market, to doctors, nurses, hospitals, and vendors, and we are listening," he said.

Well, today Jacob is here to speak, and I am sure we will all benefit from what he has to say.

Thank you.

[Applause.]

Keynote: Jacob Reider, M.D.

DR. REIDER: Steve has demonstrated what a good reporter does, which is to do research. So, I am impressed that you did that research.

I am going to enhance some of the background, because I didn't know Steve was going to do so much research. I will talk a little bit about my background, some stuff that Steve didn't catch on, and then I am going to talk a little bit about the perspective of Office of the National Coordinator and some of the problems that we see that perhaps CDS and, by extension or inclusion perhaps, natural language processing might play a role in solving some of those problems, and perhaps some of the realities that we see.

The program said that Doug Fridsma is talking now, and Doug couldn't make it today, so I did a little image search on Doug Fridsma. This is page 1, and by the way, there is a little natural language processing pitch in here, right, because Google must be doing something.

So, here we see various iterations of Doug Fridsma and some who aren't Doug Fridsma. This is page 1. Here is page 2 and go on Google and do this.

There is the simpsonized Doug Fridsma up there, but look at this. Search for Doug Fridsma and you get me wearing my Red Sox hat with a little bit more facial hair.

So, I guess Google and the organizers today knew that I am Doug Fridsma. I grew up in Boston, unlike Doug Fridsma. On my Linked-In profile, which I know Steve looked at today, it lists me as these things: husband, parent, brother, son, family doc, blog, pioneer. I didn't say nerd, but I think it's there.

Finding benevolence in business, there is an interesting thing called Googleism, so that you can Google Googleism, and you can find it, and you can ask it something, you can give it a word or a name or something, and it actually says Jacob Reider is...looking for benevolence in business. That is from about 10 years ago.

The apple doesn't fall far from the tree. That is an apple tree in the upper lefthand corner. My dad is a psychiatrist, and I think a lot of the way that I practice family medicine, in fact, practice health IT is from a psychiatric perspective.

I am still a practicing family doctor. This is not an endorsement of any organization -- I have to say that, right -- but you can all five people who answered the Health Grade survey thought I was a great doctor, and here is Awards and Recognitions, "Dr. Reider is a Health Grades recognized doctor." My wife is a law professor, and sometimes writes recommendations for students, and, you know, Joe Schmoe was definitely one of my students, it is not always the best recommendation.

I spent some time in a start-up, and the start-up that I worked for was a company called MedraMo. We actually were doing machine learning and natural language processing about 10 years ago.

The company was acquired by Nuance as was our IP, so this is just maybe a hook to this conference. I am really fascinated by what might be feasible, and I think that is what we were doing back then was doing work that might make stuff possible someday, and I think that is what I am hearing is that we are inching ever closer toward what we might actually be able to do as we align clinical decision support, natural language processing, and perhaps, as somebody mentioned earlier, clinical quality measurement.

I did spend time, as Steve described, an NEHR vendor, and learned a lot when I did that, but my focus has always been on how we can help people be healthier, both as a family doc and as an informatician. You know, I try to keep the end in mind, and for me, that is what this is.

So, I found my way to this city and specifically, this building, which is where ONC is headuartered and HHS. For those who don't know, this is just a little explanation of where we sit and what we do.

The meaningful use stuff is comprised of two regulations. One is from CMS, and that is the Incentive Program, and the other is from us, the Office of National Coordinator, and we define the standards and certification criteria for electronic health records.

Everybody has seen this slide that we have all plagiarized into our slide decks.

Blackford, do you know who did this one originally?

DR. MIDDLETON: In the Policy office outside Washington.

DR. REIDER: So, who did it, is this Pat or somebody?

DR. MIDDLETON: One of those.

DR. REIDER: Okay, because I have to, especially with Steve in the room, I have to make sure that we provide appropriate attribution.

So, someone at HIMSS took this picture, and has made it into many, many slide decks.

The metaphor that I usually use is to driving a car, so the meaningful use program is about taking your driver's test when you are 16, and you have to do the right things. You have to be able to do a three-point turn, and you have to be able to stop at the stop light, and just do the stuff that ought to be done.

That is defined by the CMS regulation, and if you do all of that, you get money, quite a bit of it, in fact, if you are a health care provider.

As a friend of my son's found out when he went for his driver's test in his dad's uninspected car, you also need to be driving an inspected car or you will fail your driver's test, and so the analog is that in our program, you can't meaningful use an EHR unless your EHR is certified. So, the Incentive Program says you must make meaningful use of certified health information technology.

We are the service station, we, ONC. Well, maybe I shouldn't express it that closely. We define the inspection criteria, because they are testing bodies that are the service stations, because obviously, we don't scale whereas, Shell and, you know, Joe Schmoe, or this is the inspection store, they do scale.

So, we define the certification criteria, and so we set the standards by which EHRs will be certified, and by extension, then, we also define what the EHRs are capable of being meaningfully used to do.

So, if we were to expand the certification criteria in some way, the meaningful use criteria could also be expanded, and by the same token, if we constrained the meaningful use criteria, then, the EHRs may not do so much consistently.

So, as you can tell, we have a lot of leverage here, because people pay attention to the things that we say in our regulations. What happens then is that products get certified.

When a product gets certified, it joins its competitors on the CHPL, the Certified Health IT Product List, which is on our web site, and so if you are a clinician or a hospital or something like that, and you wanted to purchase an EHR, you could go to the CHPL and confirm what the salesperson has told you or invalidate what the salesperson has told you, if the EHR is certified or not.

So, these are our guiding principles, and that is a picture of my dad, by the way, the psychiatrist I previously mentioned.

So, these are our five key principles: Be a worthy steward of the country's money and trust.

Focus on outcomes. That means really understand what it is that we are doing and what the end game is, not just what is in front of us.

Build boldly on what works means let's be pragmatic, we are not going to do Chitty Chitty Bang Bang, because that is not working today, and so as we think about what we might do in the domain of CDS and natural language processing, what is working today and how can we leverage that to make it feasible for the rest of the industry.

Posture innovation obviously, and support IT health benefits for all. Sometimes we translate that internally as look out for the little guy. It is not uncommon for us to be visited by executives of large healthcare organizations or by large IT vendors, and they can get our attention pretty well.

So, it actually takes effort on our part to make sure that we actually go out into the field and I listen to my colleagues who are in one to two doc practices, or I talk to an EHR vendor that has $2 million of revenue, and not 200 million or $2 billion in revenue, because they don't have the band width. They are just treading water, they can't come to us and eloquently express things in PowerPoint decks.

So, a lot of the focus the last couple of years has been on measurement, and as we think about quality improvement and the improvement of healthcare, you will notice that a lot of the effort in our rule, and the one from CMS focused on measurement.

I think of measurement sort of like we are giving all the kids in the class a certain grade, but we haven't empowered them to get good grades. So, everybody gets a C-minus, and we expect that to motivate them to get an A-plus.

I think of CDS as perhaps some of the tools that would enable them to get A-pluses, but anyway I am digressing.

Historically, quality measures were done using humans who read paper records and then made decisions based on little survey tools that they had about whether quality care was rendered. It is a retrospective view as we heard earlier. We are looking in the rear-view mirror, sometimes, a long time ago, especially if we are using paper records, and the abstractors look at what kind of care we rendered, 6, 12, 18 months ago.

Then, we develop perhaps an intervention for how we can solve the problem, and that whole iteration takes, you know, 36 months is not uncommon from discovery of quality issue to its resolution, and by then, it might have gone away anyway, or we have created other ones.

Decision support, of course, is prospective, and I think we had some discussion during the last session on that, and how decision support may be a prospective, and I think there are, as Juggy mentioned toward the end, there are different qualities to decision support where it is not just the mirror images of clinical quality measures.

I like to talk about the other decisions of CDS, and that is consumer decision support. This is our Surgeon General at the launch of our consumer health program last fall.

So, when you think about CDS, let's not

just think about the provider as the recipient of CDS. It is also the patient or the caregiver or other members of the care team. In fact, the physician may be the rate limiting step here, and as we think creatively about how we are going to solve the healthcare problem, why is that, you know, a mammogram reminder should ever be presented to a physician.

We can look at organizations that are doing things progressively. It is really idiotic to remind me about my patients' mammograms, so I can ask somebody else to ask somebody else. Let's just ask the patient, and if our systems know how to do that, there is no reason for us not to do that.

So, I think we need to think more and more about being patient centered. I can remember back when I started in the vendor world, there was all this talk about being physician centered, and that was like a revolutionary thought to stop being hospital centered. We are going to stop thinking about the hospitals. We are going to go downstream, we are going to think about the doctors.

I can remember talking to the CEO and I said, "You know, you are actually not back to the root of the issue yet," and he looked at me quizzically, and I said, "You know, the patient is more important than the doctor."

So, this is a bit of a portrayal of what we might see, and it is a caricature, of course, and oversimplifies, but if we go from research to guidelines to CDS, and then to measurement.

CDS is the part that, in general, hasn't been present in the last decade or two, but it is not a waterfall, it might be better thought of as a circle, right. So, we go research goes to guidelines, goes to interventions, we hope, goes to healthcare and life. That was a difficult circle to draw, because it encompasses so much. Then, we measure it.

The reason I didn't say physician or hospital or care provider is that ideally, our intervention may have greater scope than just the care providers as I talked about a second ago, and then we measure it. That might be part of very tight loop, so you could say this loop is a five-minute loop, or it could be a five-year loop, or something in the middle. There is nothing that necessarily dictates how slow or fast it must be.

An important piece as we think about all of the connections here, and certainly very important for us at ONC, is the adoptability or the usability of the systems. So, we hear a lot about how CPOE is this horrible thing that we have imposed on the industry.

I had a very pleasant conversation with a physician just last weekend who said, you know, the meaningful use requirement for CPOE, which is Computerized Provider Order Entry or Computerized Physician Order Entry, or something like that, is a really terrible thing. She thinks it doesn't -- our presumption that the provider should be interacting with a computer is an inappropriate one, and she can think of lots of ways that this would work better, so she encouraged me to rethink whether CPOE should be part of meaningful use. Certainly, that is fair game.

A key part of her concern was that it is not usable, that CPOE doesn't fit with the way that she works, and this is a bit of a user experience framework that I like to use as think about usability. So, when I hear complaints or concerns about usability, this is the framework that I consider.

So, functional means it does what it was intended to do. Reliable means it does it that way every time, and so systems that go down, or break, or, you know, are down every night for backup.

My first EHR experience, I am a bit of a night owl, so I would often, unfortunately, after my kid's soccer practice and whatnot, would make it back to my computer at about midnight to finish my progress notes for the day, so I am one of those unfortunate few who couldn't finish during the day, but our system would go down at

12:30 for backup, and it would stay down for most of the night, about 4:30 it came back up.

So, if I wanted to finish my notes before the next day, which was an important thing for me to do or else I would get too far behind, I would have to wake up at 4:30 and finish my notes.

That wasn't a reliable system. Usability is it meets my expectations, it is easy. I can interact with it without taking proton pump inhibitors.

Convenience means it anticipates my needs, and pleasurable is self-evident.

So, this is a functional system.

This is a functional system.

This might be a usable system.

That might be a pleasurable system.

So, we can look at any kind of system, whether it be EHR or any other kinds of new technology, and we can say that folks will actually express emotions about them, so I did a similar session with a bunch of medical students last week, and I said, "Raise you hand if you have a Smart phone." Of course, they all went up. I said, "What emotion do you have?" Of course, all the ones who purchased this company's products said "I love it."

Well, when do we get there with electronic health records? But more importantly than loving something is that there are some things that we have are safe. So that is another focus that we, at ONC, have. Many of you know we have commissioned a report from the IOM, which was published last November, that looked at the safety of health IT systems.

So we are also thinking -- and, of course, usability and safety are tightly aligned, right, if the system isn't usable, it is actually easy to make mistakes, and if it is easy to make mistakes, we actually have health IT-caused patient safety problems.

Health IT can also be the solution to patient safety problems, so we need to better understand safety.

This is an example of a fairly unsafe

environment where if bicycles are riding across roads very quickly, they might get smooshed by cars. These sorts of things slow down the bike riders, the little cattle crate systems, and in my area in Upstate New York where I live, they have actually put these in many of our bike paths, and the reason for that is that people were getting smooshed by cars when they went too fast.

The same can happen in health IT systems, and the reason I put this up is that often we will hear from end users that the system is too slow, and so the faster the system goes, in fact, in some cases, more errors occur. So, raise your hand if you have Auto Suggest in Outlook and sent the wrong message to somebody.

So, sending the wrong message to somebody may actually have career-ending implications. But sending the wrong prescription to someone or looking up the wrong patient may actually kill somebody. So, we can end our careers with Outlook, but if the system causes it to be too fast, it can also be unsafe, so we think about this, and I think decision support is really important domain for us to think about this, because all of those Auto Suggest systems use algorithms that are decision support, the system is helping you do something faster and more easily, and if you can do it too fast, that is a problem.

We participated in a session last summer at NIST where there was an example about which side -- it was OR form that a doc was supposed to fill out, and it was like which foot are you going to amputate, and the question is do I design this using a radio button where there is a 0.4 percent chance someone is going to click the wrong one just because they are in a hurry, or do I make a free text field and force them to type in LEFT.

The latter is actually the right answer, but in this attempt to do things quickly, many vendors are now making the former selection, because their customers are saying help us do it fast, help us do it fast.

Then, there is the truck that is going to smoosh somebody.

There is a problem that we have been seeing -- and I loved the discussion earlier about value sets -- in the quality measure domain. So, this is my rendition of the quality chasm 2.0.

Quality measures say things about what ought to happen, and then, of course, there is the exclusion and the exception concerns, and for those who aren't entrenched in quality measures, I will give a quick Quality Measure 101.

Quality measure asks did this happen, and yet when I am judged on whether I have done something or not, so did I do the foot exam for my diabetics, and it turns out that if I am a physician and I am being judged or, in fact, paid based on my performance, a percentage of my patients won't have gotten a diabetic foot exam done, especially the ones who don't have feet.

So, the absence of feet would be an

exception to this, but yet I need to tell my EHR that the patient has no feet. So, there is this balance between the quality measure world that says, oh, you know, we need all of these exceptions to be expressed in the quality measure, so that we can help providers not get dinged where they shouldn't get dinged. In that case, the quality measure is very complex and has lots of expectations of data that may or may not be in the electronic health record.

That is why I have the quality measure expectation is on the left of this chasm, and the capability of the EHR, now, the fact that the patient has no feet may or may not be in the EHR, there are many exceptions, you know, of medication, a patient preference, the fact that the pharmacy was out of the medication at the time, and that is why I didn't prescribe it, there are lots of reasons why a patient won't get something that are exceptions, and the EHR may or may not have that information.

As you think about the number of elements that could go into a clinical quality measure, the folks that make these measures, there are lots and lots of things that they could say in a quality measure, so I think of these as many thousands of crayons, and yet the EHR may only have a subset of those things available to it. There may be a vocabulary that expresses only a subset of the things that the clinical quality measures would express, so that is the big gap.

Yet, we are trying to solve the problem with the tool that we have. There is the EHR capability, so my attempt to think about how we might fill this gap, so there on the left is the quality measure expectations and capabilities, that we have standards. We have standard terminologies. We have standard value sets. We have standard methods of capturing and expressing information.

We might actually have to modify the expectations that we have of the information, so perhaps the exceptions for our quality measures shouldn't be so expansive, and at the same time, we may need to enhance the EHR capability.

So, we, ONC, can work on enhancing the EHR capability. I see Marc in the front row, so I will pick on him. Marc is the CMAO at Siemens, and we will hear from Marc briefly. I could say to Marc, "You must always capture the fact that the patient had a foot amputation anytime there is a patient who is diabetic."

So, whether or not this patient has feet needs to be captured. Marc could think about ways to implement that in his system either using natural language processing, where they would go through narrative notes or dictated notes or some other evidence of that, maybe CPT code 17 years ago from that amputation, but I may require that of Marc explicitly, and I could say for certification he absolutely, positively has to identify this event.

I could even name a SNOMED CD concept for this thing, yet, at the same time, there may be 3,000 of those things, and I can't expect Marc to add 3,000 by 2014 EHR certifications, so I might pick 1,400 of them, and yet I will then need to talk to the folks who make the measures and say you folks had 3,000 of these things, pick the best 1,400 and tell me what they are, so I a can impose them on Marc.

This is the rock and the hard place that ONC finds itself in today, and then usability and workflow always needs to be part of how we think about this, because these are not just forms that we are instantiating in EHRs.

In fact, if we allow or encourage the vendors to instantiate hard coded forms into their systems, we will destroy the usability in the workflow, and this is hard wiring, so any of you who have been paying attention to the HIT Policy Committee, which if a Federal advisory committee that tells what to do, you have heard them complain about hard wiring, and so this is a case where if I told Marc there is a SNOMED CD concept, Marc would actually -- I think I have the picture -- make one of these, and then his physicians will jump off a bridge, because this is terrible usability, and yet this is what happened in many cases during Stage I of the meaningful use program, which was started in 2011.

So, currently, today, what we have, and part of why we have so much hard coding, is because we have CDS and CQMs embedded inside of EHRs, and so Marc has to hire a herd of programmers -- I am really picking on Marc, but he is in the front row, he is used to it, too. He is a big guy.

So, these pieces are part of the same system, and so you have to hire a bunch of programmers, the time that it takes to actually iterate and implement these things is quite long, because you actually have to hire humans to interpret what these concepts are, and then put them into your systems.

This is what we hear from our physicians, right, "I am a doctor, not a data entry clerk," and yet we have to think more about where the puck will be, and not where it has been, and so we think about the levers that are available to us in the Federal Government specifically at ONC.

Standards are a lever. It is one of the two things that ONC defines when we create the standard and certification criteria, and so we look to standards for, say, now information is shared, how information is incorporated into an EHR, is there a standard for how a clinical quality measure or a clinical decision support intervention, and the CDS folks in the room know that in our proposed rule, we refer to these things as interventions, not rules, and the reason for that is that we want to think about how it is exposed and interacts with the user, not about how it is instantiated in the system in electronic form.

So, it might be a rule, but it might be something else, and we don't want to make any assumptions about how that is encoded. So, are there standard ways for us to do things that would enhance the likelihood that they would accelerate?

Then, we have regulations, policies. These are also standards. They are standards of behavior. So, we have technical standards and standards of behavior. These are the tools that we, the government, have to actually accelerate process, because when you standardize things, they can go faster, and that is what I tried to say here.

When I take the train the BWI to D.C., it is much faster than the bus, and the reason for that is that there is a standard.

Standard things, you can also plug in and unplug, and so this is a standard for how we, you know, put things in sockets in the United States, and so if we were to standardize both clinical decision support and clinical quality measures, and have them expressible in machine-independent, endpoint-independent formats, so that Marc and his colleagues from Next-Gen, and his colleagues from EPIC, could consume these pieces of clinical content autonomously from the core system, they could actually iterate them much more quickly, and it wouldn't be so hard coded, hard wired.

So, do we have an opportunity for this as your data entry mechanism, and instead of the form or the template? Maybe that's dictation, maybe it's something else, mind reading or interaction with the patient, or something else.

Is this another way for us to understand how it is that the patient's care has been rendered? I interacted with a company a couple weeks ago that is actually creating a system for physicians and patients to interact with each other electronically, a secure messenging system.

But what they are doing is fascinating. They are not just doing the messages. They are doing analysis, they are doing machine learning, and they are actually looking for outcomes, so the system automatically sends a message to the patient 24 hours after the last interaction, and the patient says, "Hey, Doc, you know, my back hurts." The doc gives him advice, and then 24 hours later, the system automatically follows up with the patient and says, "Hey, are you better?"

And the patient says, "Yes, I am better." Then, what the doc gets is a better understanding of the outcomes.

They are reading the text of these things, and that text actually carries with it a set of diagnoses and a set of actions, a set of interventions, but then we can learn from as we iterate forward.

So, as we think wheel, and think about all of these things, we are thinking very hard about how we can write our regulations without make presumptions about how things are done. We want very much to accommodate new innovative ways of doing things that may include decoupled CDS, that include natural language processing, that may include innovative ways of interacting with patients rather than just providers.

Here is CDS 1.0. I don't need to read it, I don't think, it is one of many. We don't really think in those terms. This is maybe a 2.0 just right, just in time information, right? Where I need it, when I need it, what I need, and the system has to do that properly.

So, without really complex thinking, we can't develop it just at the right time and the right place.

This is my mascot for Jr. JITI.

I am finished. Thanks for your attention and I look forward to talking.

[Applause.]

MR. LOHR: Jacob, thanks very much.

We are running a little over, so let's save our questions for the end if that's okay.

Our next speaker is Dr. Gregory Steinberg of Aetna.

Dr. Gregory Steinberg, Aetna

DR. STEINBERG: Yes, I am Greg Steinberg. I am currently head of Clinical Innovation at Aetna, and I am going to speak to you from the perspective of the payor and indirectly from the employers who arguably are involved in the mundane, but not trivial, task of actually paying for some of these products and services that we are talking about.

I will say from the outset that I think that a lot can and has been done already with existing data sources, so-called administrative sources, and high-quality CDS systems that are around today.

Clearly, there is the potential to do more and better with natural language processing, and I will try to cover some of that, and I will also speak briefly about some high-level strategic thoughts from the Aetna perspective, particularly relative to the new paradigm of accountable care organizations.

When I talked about data, current administrative data, what that means at least to us is obviously claims, diagnostic and procedure claims, but also data from the PBMs, lab data, so we are able to get, and have done for many years now, not just the fact that you had lab tests, but we know what the lab result actually is.

For some years now, we have also had what we are calling patient self-reported data, so we have a personal health record where data is inputted by patients electronically and comes into our database. In addition, when we have care management nurses speaking to patients, they are entering that data, and that data is also part of the longitudinal health record.

Somebody talked about the lack of availability of longitudinal health records, and I would posit that these are actually available and have been available in the payor world for many years.

More recently, and we will talk a little bit more about this later, but the health information exchange capabilities, which allow us electronic access to EMRs is going to be an integral part as we go forward of this expanded database, and what we have in the middle, that sort of green pepper mill thing is the clinical decision support engine that takes all that data and then depending on how the algorithms inside the CDS system are configured, does a number of things.

It will either deal on the left with

patient-specific so-called precision alerts, gaps in care that we have been talking about, people have mentioned patient engagement, so versions of our gaps in care are fed back actually in real time to the personal health records and to the patients, therefore.

Different versions of clinical decision support rules can inform real time analytics. We have talked a little bit about quality measures, registries, et cetera, and more importantly or more recently I should say, workflow rules have been devised to help in terms of care coordination again relative to accountable care organizations, patient-centered medical homes, and the like.

A little click down in terms of the clinical decision support system. This will go quickly. Again, we get all the data that I talked about, it creates this pretty robust longitudinal patient-centric electronic medical record, which is then applied against a digitized version of the evidence-based medical literature, and what comes out the other end, if you will, is a gap analysis on a patient-specific level that looks at the difference between the care that a patient is actually receiving as reflected in the data and the care that they should be receiving as reflected in the literature.

That little nugget of information is then encapsulated into various formats that will be transmitted, that are transmitted in various ways to both doctors and patients.

Clinical decision support, like a lot of things in the world, vary in terms of its robustness and capabilities, and this was mentioned earlier. A lot of decision support sort of looks like this, and this is obviously relative to diabetes and it is fairly basic and fairly uninteresting, and more importantly, not really a reflection of the true variety and complexity of the situation.

For us, clinical decision support is this, and it is not just that there is more stuff, which there is, but this is actually a representation of how things interact with each other physiologically and how physicians actually think.

It goes I think also to the point that was made by a few people, that clinical decision support systems are not necessarily able to take into account the complexity of comorbidities, multiple medications, and so forth.

That is not necessarily true at least with respect to the "do not harm" capability, so you can clearly have -- and these are sophisticated rules that require sophisticated rule authoring capabilities, but they exist -- so you can clearly have, going along the top, you can clearly identify a diabetic, let's say, with a hemoglobin A1c who is not on any medication. The American Diabetes Association would posit that that person should be on Metformin, but there are a whole host of conditions where Metformin may not be a good idea.

You can write rules that will proactively look for those conditions and not send the output of that rule to a physician or a patient if any of those conditions are present by either lab data, codes, et cetera, combination, and there are number of those - statins, ACE inhibitors, so you get the idea.

Give you a sense of the fact that we have been doing this for real on a fairly large scale, so these statistics are relative to 2011. On the left are some of the sources for evidence-based standards that we use, a fairly large group of full-time docs and pharmacists are involved in building and maintaining these rules, the type of rules, the numbers of rules are in the middle there, and then it gives you a sense of some of the activity -- and this is just in one year of the rules, of the messages that were generated to the providers and/or patients.

One of the points was made about, you know, you see how there are many more patient alerts than there are doctor alerts. That goes to the fact that, as somebody said, it is probably not a good idea to send mammogram alerts to docs, so we don't, but we send them to the patient.

The little thing at the bottom just makes the point we have the opportunity to do a chart analysis with a large teaching hospital in New York City where we had access to the medical records, and we were able to compare both the diagnostic validity of the rules and the diagnoses that we were imputing and the actual clinical content of the rule and had greater than 90 percent concordance.

So, what have we learned from working with the systems over the years, and those the areas where natural language processing may help us. The real world data that we deal with, although it is useful, very useful, it has some issues, right, so diagnostic claims clearly per se are often inaccurate due to a whole host of errors, not must miscoding, but ruleouts, okay. If I want to rule out diabetes, it is the same code as if the person actually has diabetes.

You can construct rule logic both

inclusionary criteria and exclusionary criteria to mitigate that, but it is an issue, and then clearly claims lag.

The sensitivity/specificity problem, again, this was mentioned by a number of speakers. When you are dealing with patient-specific alerts going to docs, you had better be right all the time. Docs hate wrong alerts, they really hate it, and they had no problem telling us that you have lied to them, and clearly the potential for alert fatigue.

So, you can do that with increasing your specificity, but if you are going to do that at the expense of sensitivity, on the other hand, and again this was mentioned earlier, when you are dealing with population-based quality measures, you probably have to dial it the other way, particularly on the numerator, because many providers are going to be measured and potentially paid based on how these rules are structured.

The last thing at the bottom of the

slide, you know, absence of evidence is not equal evidence of absence is a real problem, so how do you get to errors of omission, you know, things that are not there, that should be there.

In our current paradigm, we are using eligibility data essentially, and as a surrogate for the fact that if something is not there, you know, if it really had been there, we would have seen it.

That is true a lot of the time, but it is not true all the time, and clearly, can lead to false positives and clearly an area where natural language processing can significantly help.

So, is there any data that these CDS systems that do not use natural language processing, only these other sources of data I have talked about really work?

These are two publications based on one study, and this is one of the only randomized, prospective, controlled trials of CDS that I think is in the literature. So, they took 40,000 members in a health plan in Cleveland and basically randomized them. Half of them got the clinical decision support system, half of them did not, everything else was well matched, it was a one-year prospective study, and the predefined endpoints were, in addition to the number of errors that were found, looking at the left panel, hospitalizations and paid claims, real money. These were statistically significant changes, hospitalizations reduced by 8.4 percent, paid claims reduced by about $8.00.

The panel on the right took the same data, but had an extra year of data afterwards, after the study was over, because what happened is that the health plan that was involved figured, based on the results of the panel on the left, that it was unethical to not have the control group have CDS, so the control group received clinical decision support in the year subsequent to the study, and the authors of that study published in the Journal of Health Economics looked at that.

So, they looked at charges, not paid

claims, they found again, they confirmed that charges were significantly reduced, hospitalizations were reduced. Interestingly, the hospitalizations were all to do with the areas where the rules were involved, so there was heart disease, diabetes, and the like, and then the causality thing was interesting.

So, what happens, you had the two groups started together in terms of charges in the first year. They diverged during the study year, and then in the year after when both groups now got the CDS system, the charge differential disappeared, which led to the conclusion that there was probably a causal effect.

I said that we would talk a little bit about accountable care organizations. So, what this slide is really talking about is that we are undergoing a fairly significant paradigm shift right now where we are moving from patient-specific reactive care to population management and proactive care, and moving from paying people more for doing more to paying people more for doing better. This is a fundamental tectonic shift that we are involved with.

The way we have tried to start to look at this and try to cobble together a set of solutions that might help is in addition to the decision support stuff, that is the bottom right bubble there, where you have gaps in care and population-based decision support tools and workflow tools, we have the health information exchange capabilities being added to that.

We have on the top there Medicity, which is a large health information exchange, the largest in the country, has the largest footprint anyways, which also has this iNexx capability, which is a cloud-based application store, like the Apple Store, where a whole host of applications can be developed quickly and sit on this and be downloaded securely behind an individual provider's firewall in their office.

Layered on top of that as well are tools to help the patients who, iTriage is one of the leading consumer mobile labs that provides navigation, symptom, clinical decision support, et cetera, and then wrap all of that around with health plan services, because these accountable care organizations are going to have to start dealing fairly seriously with all of that risk that is something that is sort of new and foreign to them.

So, double clicking a little bit more, and I apologize, it's kind of busy, but what this is trying to say is this is one more level of detail, so at the bottom you have the doctor provider, provider data on the left, you have the hospital data on the right. It is being connected through the Medicity Health Information Exchange grid structure on the bottom.

It gets fed up into the analytic engine, which does a number of things. It sends out the care alerts and the decision support straight up. It talks to the patients. It provides workflow tools for the doctors to look after the patients proactively, and a whole host of reporting capabilities that are sophisticated and that are going to be needed in order to survive and thrive in the ACO world.

One way where natural language processing could clearly help is being able to take all the unstructured data that we have all talked about earlier today and just add that to the existing data.

Clearly, that will allow, not just the clinical decision support functions that we have talked about, that function better and more accurately, but there are probably other ancillary benefits in terms of efficiencies, in terms of utilization management, prior authorization, et cetera.

This is my last slide. I think, you know, we have come a long way. Just with the existing data sets that we have, with the existing clinical decision support tools that we have, unstructured clinical data with the held of natural language processing will allow us to get over the bridge. That parenthetically happens to be originally where I live in Pennsylvania, it is the oldest privately owned bridge in the United States.

Thank you for your attention.

[Applause.]

MR. LOHR: Thanks very much.

Our next speaker is the previously introduced Dr. Marc Overhage of Siemens.

Dr. Marc Overhage, Siemens

DR. OVERHAGE: I feel kind of funny talking from Siemens' perspective. I spent 25 years at the Regenstrief Institute with Clem McDonald doing clinical decision support, and so that is sort of my long-time home, and I have been at Siemens for just a year now.

It was a little bit hard to think about what would be useful to talk with all of you about, and so I thought what I would do is poke at a few of the areas that we are bumping into as we are doing our work where NLP and clinical decision support sort of collide in some way or another, and, in particular, some of the things that maybe we haven't spent too much time on.

The message that this is important has gotten through to people. This slide is actually one that Herr Professor Requardt, who is the head of our board for healthcare, Siemens used in London a few months ago at a stock analyst meeting.

This was the only slide that he used talking about the future of the company, and talked about how unstructured data, disease models, and therapy interact today, and how the evolution is to structured data, patient models, individualized therapy, and a knowledge domain in the middle going forward.

So, people even at the senior leadership levels in healthcare and technology I think are getting this message and beginning to understand where all of us are trying to go with this work.

This is just Marc's mental model of how some of these pieces fit together, and I am not going to tell you about all of it. The red stars represent places where NLP in particular kind of pops up, and I am going to highlight a few examples of that.

I think of it as coming from two directions. One is the lefthand side which we have had a lot of conversation about today, which is patient data getting turned into structured data, so that it can be used and learned from.

We are actually running a lot into the righthand side of this equation, which is how do you populate that right set of whether it is production rules or whether it is some kind of expert system or whatever it might be, how do you build the knowledge base that needs to underpin that. I will give you a concrete example of that in a minute.

One of the things that I get to brag about is work that my colleague Bharat Rao has done, which is, you know, one of those interesting weighing commercial successes of NLP, which you might not know exist, but Bharat looked at the problem that we just heard a little bit about quality measures, and the challenge of much of that documentation is in unstructured text of some kind, in this case, left ventricular systolic dysfunction that is needed for a particular quality measure defined by our friends at CMS in a very specific and concrete way, and he system does the usual things to find and drive that data, and provide feedback to the user about the context where it came from.

There is a couple of things really neat about this system. One of the things that I wanted to highlight with this was the learning aspect of the system. What I mean by that is there are several hundred users across the United States, several thousand across the world who are using this system every day in hospitals and health systems to help capture data for their quality reporting.

All of the cases where they choose to change the machine's interpretation get fed back into the learning algorithm on a daily basis and refine it, so that the processing that happens on an ongoing basis is improved by the feedback of these thousands of individuals who are solving their day-to-day problem, which is how do I capture this data for a particular quality measure, so one bit of genius that Bharat and his group did. I will come back to another one in a minute.

A second way that we bump into NLP a lot -- and this is almost a different class of problem -- which is we spend a lot of time thinking about the discharge summary or the radiology note.

We also find need for short snippet, which is a whole different set of challenges and problems, and I think we will probably hear a little bit about question answering in a little bit. That is another domain where short input has to be interpreted.

Another one that we have spent a lot of time on, and this is actually carried over from some of the work I was doing at Regenstrief, is the fact that you often have, in laboratory results, as an answer a short snippet of text, so, for example, if you are trying to do public health reporting of reportable conditions, you might have a test for shigella, which might say shigella isolated, great, and not too hard to do.

But it is much, much more common to have something that says no shigella, salmonella, or E. coli isolated, and you have to recognize those three observations, the fact that it is negated, and all you have are those six words, or seven words, I can't count, but, you know, very, very short snippets of text, and this is just for the purposes of public health reporting what we have done a Regenstrief was sort of three layers, and that was the theme I have heard throughout the workshop here, of sort of numerical results, those are easy, it's the threshold and things, discrete results, like it's positive or negative, and then there is a much harder category of results where the answer to the question is something that is a short snippet of text that needs to be turned into structured content in order to be able to process it.

The second thing that Bharat Rao did with this so-called remind platform that I thought was intriguing, and there were a couple references to throughout the discussion, is the need particularly with unstructured data to begin to combine and reason probabilistically about the data that we got.

So, the classic example that he walked me through when he first described this was a patient who is in the hospital, you have his ambulatory physician's note, which says the patient tried to quit smoking two years ago, but failed, and restarted, and you have the information from the patient's PHR that says, you know, it's a year old, that says, you know, I am not smoking anymore, and then you have data from the admitting nurses' note that says the patient smokes a pack and a half a day, but intermittently.

How do you put that together to decide whether the patient smokes or not? Taking into account the temporal pattern, we heard a little bit about that from the Partners NLP group in the last session, as well as the reliability of the reporter, you can begin to construct probabilistic inference about those individual elements and decide what you want the answer to be today, and like I said, in particular, in unstructured data, but in structured data, as well.

A third snippet is within Siemens, they call it "analyze as you type," but there is another short text aspect of NLP where a user who is actually interacting with the system might be recording data, entering symptoms, findings, and diagnoses in a free text format, for example, and the system is actively processing as each word is added to that text, taking into account what was there before, as well as the patient and the provider context, information in order to derive useful output in various ways.

So, here is a concrete example of this from our work at Regenstrief where we were applying this. On the lefthand side is a free text box where user is free text, narrative text, right, the narrative text box where a user is just entering things.

On the righthand side, as that text is being processed, in this case, the assessment and plan, the system is matching that up with concepts of things that might be reasonable to order.

So, for diabetes, they noted that an A1c ophthalmology consult and electrolytes might be appropriate, so those three things show up over on the righthand side to make them easy for the user to go ahead and order, somewhat analogous to the work that Tom Payne described of adding problems to the problem list, just make it easy to get to those things, and not to forget about them.

Another way that we are leveraging this data and actually trying to improve the data capture is by building individualized patient models, and that comes into play in particular in the capturing of the data.

If we have very good models for that individual patient -- this happened to be one for hypertension, one of my good colleagues Glenn Fung did -- but as you begin to have this, you can play that back into the data that you choose to use out of the patient's record whether it is for clinical documentation or other sorts of uses. So, that is another place that we are leveraging it.

Sort of the other side of the equation, the righthand side, we focus on for a minute. I haven't heard a lot of conversation about today, but it is one of the places that we are spending a fair amount of energy for a variety of reasons, which is working from these unstructured as well as structured data sources and trying to drive through to usable knowledge, and ways to curate and manage that knowledge, so that we can then begin to use it for actual reasoning and for semantic sorting of data and things of that nature, and I think we will probably hear a bit about that in a few minutes.

In particular, one of the things that we try to find is sort of what I call the art of the possible, in other words, while these approaches are imperfect in many ways, our ability to find semantic structure within whether they are guidelines or journal articles, and so on, that can begin to inform how we drive the decision support, have been interesting, probably the most sophisticated application of this that we have done to date has been where we are trying to pull together several aspects and taking data in the semantic structure and ontology derived from the cardiac literature, along with structural observations from CT, MRI, ultrasound of the heart, and then using those, construct dynamic models of the cardiac circulation.

So, structural observation, data from the literature about how these things interact and create dynamic models of the heart that can actually be used for decision support for cardiac surgeons who are operating on complex cardiac abnormalities to say, well, what if we only opened up the valve 10 percent instead of 25 percent, what might that do to the cardiac circulation and in neonates in particular, but even in older adults, for example. While experience is a good guide, having the data and the model that underpins it can be incredibly powerful decision supports that say I twiddled the dial this way, where am I going to land.

That might lead them, for example, to choose a less aggressive procedure initially, reassess, and then follow up with a more aggressive procedure if needed, and there are other examples of that.

So, that kind of decision support for therapeutic uses can be driven by these data driven out of these various sources.

On a broader scale, we sort of look at things like text mining in NLP as one source along with image segmentation, formalization of treatment plans, and so on, that sort of create a continuum. This is not a product, but a prototype called Medico, that tries to start to pull together all these different uses of NLP and decision support, stole a little bit from the work that Blackford and his group did in terms of Smart forms, took it in a little bit different direction.

But on the top lefthand panel, there is patient data with the associated images linked together through a semantic network, so that the abnormalities in the image and the data in the patient's record are linked together.

Some historical data. Somebody talked earlier in the middle righthand panel about what happened to the last thousand patients who looked like this, so the comparison recommendation is based on similar cases driven out of the local record system.

In the bottom left, the knowledge network that is a linkage back to the clinical guidelines and the literature that support those things, and then finally, on the bottom right, taking advantage of both publication streams and patient care patterns to link to practitioners that might have particular expertise in this patient's care, so this is called Medico, and is part of a European project that we are just -- actually, just in February had the five-year kind of wrap-up for, but it represents sort of an effort to pull together all those different pieces of decision support, many of them leveraged by NLP to create kind of a dashboard, if you will, in my mind proportioned wrong for illustration purposes, you know, the patient is sort of not overly emphasized.

But it starts to hint of the possibilities as you pull all of the view things together in order to enable decision support and leverage by NLP across the board.

So, thanks very much for your attention and look forward to the discussion.

[Applause.]

MR. LOHR: Thanks, Marc.

Our next speaker is Dr. David Gondek of IBM.

Dr. David Gondek, IBM

DR. GONDEK: I am David Gondek. I was part of the IBM Watson team. I was responsible for the machine learning in the Jeopardy system, and now I am their technical lead for the healthcare adaptation, so I am responsible for driving the accuracy numbers up as we look at can we use Watson in the healthcare space.

I want to thank the organizers for offering the chance for me to talk. I also want to thank NLM. It is always good to come down here. We are actually really a big closet NLM fans back in the lab. We have a list called -- there is a COOE for that, where people post COOEs that they find.

I think the discussion last week was about someone found sense of impending doom, which didn't sound so good, and then he found impending doom, and you are saying that is one that you don't want to see in your record.

I probably should start off by positioning, what is Watson, because we have heard a lot of discussion about Watson already.

I think there is a lot of hope for what it can do. This being an IBM talk I also need to show you an architectural diagram.

Roughly speaking, the architectural diagram of Watson looks like about this where you have a whole bunch of components on the left, and in the middle, a Watson occurs, and then you get an output.

This is both funny and scary. It is funny because a lot of people actually think that Watson can do almost anything, which it can't. We are working on it.

It is also scary in that this diagram is not so far from the truth in that, when we built Watson, we relied on a lot of the best state of the art in technology, so we were using technology from the latest literature, our parser had been in development at IBM for 30 years. When we looked at the medical domain, we started looking at using UMLS, which has been a huge amount of effort.

So, it is really true that Watson is an

ensemble system and depends a lot upon the components that are already there.

Now, the original system was playing Jeopardy, and some people say that is when a miracle occurred, but we looked at why, what does it take to play Jeopardy, and that actually dictated a number of the choices for the system, and I thought it would be interesting for you to think about whether those choices are useful in the medical context, because I think what you will see is due to the background there, we are taking a somewhat different approach than some of the other approaches you have heard about.

In particular, we looked at both unstructured and structured approaches for Jeopardy. I think everyone here is most familiar with the limitations and benefits of both. The group that I am in came from a background of dealing with unstructured data, so dealing things like semantic search, classification, that sort of thing.

I think we are all very familiar with

Keyword Search where it has very broad coverage, you can go in it very fast, can be very timely, but, of course, the precision is very low, and there is basically no semantics, so there is very little semantics when you do a Keyword Search.

You can contrast that with more knowledge base approaches or rule-based approaches, which you can do very elaborate reasoning, very impressive reasoning. You have precise semantics, so when you get an answer, you can often come up with the proof. It may not be readable by a layman, but you can at least come up with the proof of why the answer is there.

The question is of liabilities when you often have to hand-construct the models, they can be fairly brittle, so it can be hard to map into the models, so if you are dealing with natural language, it can be hard to understand how to map it to concepts, it can be hard to keep them up to date, and the costs can be very costly.

So, the Jeopardy system is kind of a combination of these two approaches. It wasn't a choice of one or the other. We basically merged some of the best approaches we could get our hands on in Search with some knowledge base techniques.

The hope then, is that you can actually get the coverage you would need for like, open domain tasks like Jeopardy where they might ask about anything, and then for certain areas where you realize the semantics will help, you can invest in those specific areas.

I am not claiming that Watson is able to deeply represent very complex medical guideline. That is not the goal. Rather the goal is to give Watson the shallow semantics and shallow reason techniques it needs to interpret a large number of guidelines or of diagnoses where we can get the coverage without maybe getting quite the same explanatory depth.

One of the important reasons why we think we had success with this was the architecture, and I am on a team with about 40 researchers focusing on the underlying technology of Watson, and probably about half of those are algorithms.

The other half are working on systems, so that means I am developing the architectures for plugging the algorithms together. That means working on scale out, how can you make it fast. It means working on doing things like dealing with normalizing your data. There is a lot of stuff like that.

But if you look at the stack of Watson, the Jeopardy system was built on a power 7 system. It used UIMA. I think other people have already mentioned UIMA, and then on top of that, we built the statistical learning framework, and then all of these NLP techniques.

So, you look at the NLP techniques that we developed, and they are both rules based and statistical, we will use anything we can get our hands on. You have already heard about some things like question, parsing, sense parsing, disambiguation or any detection coding, relation extraction, try and identify the relation between two things in text.

Linguistic frame extraction is we actually try to build knowledge bases from large corpora, so in the Jeopardy case, we mined over terabytes of web data. It was about 10 percent of the Library of Congress, and that was to get both facts in a knowledge base, so we know who invented what. It is also to understand about usage, so we can understand how is language used, what are selectional restrictions for classics, any of these participate in this sort of relation.

We can learn hyponymy relations. This is a way of actually extend your knowledge base or just perhaps your hand built knowledge bases don't have enough coverage, at the cost of some noise.

Also, we have a textual entailment framework, and the purpose of that is to identify whether one passage or set of passages justifies an answer to a question.

So, just looking at lexical mass, something simple as a Word Master, these lexical terms match or not, kind of like keyword search. You can look at deeper analysis like using the product (inaudible) structure in illogical form, it can look at coding, it doesn't have to.

So, what that allows you to do, then, is you have this ensemble of different approaches that will trade off sensitivity and specificity, so some things are just purely lexical, others things require that you actually code correctly and are able to identify the relations.

So, it is this ensemble of techniques that we found had the best performance.

Now, the way we were able to do this is with some of the facilities for integration. I mentioned UIMA before. The Watson system was based on only text data, however UIMA allows to use multimodal analytics, so you can deal with things like images or speech.

What it means is that there is share and process system, so if you have a component, if you are an NLP researcher, or if you have a knowledge base in a medical domain, to add it is fairly simple, and you just have to write it to the UIMA API, and then plug it into the system.

Now, that is not enough, because what is the system going to do with that. It is not enough just to add your new parser or add your new ontology, but you just have to, the system has to figure out how to use it, when to use it, when to trust it, when not to trust it, and so that is where the statistical integration framework comes in.

So, we have a framework there, where you can register all of the scores from these analytics, and it trains, so using training data, and in Jeopardy cases where Jeopardy questions and answers, it learns which components are more reliable for which question.

Then, at apply time, when you are actually running the system, it can use those trained models to know how to combine the output of all of your different NLP techniques. The net of this is that means that you don't have to understand the entire system to improve it. So, you can write a specific component, plug it in, let the system learn how to weigh it.

It is something that we are working on all the time is how to extend that to more aspects of the system, everything from scoring passages, to identifying the types of things, to evaluating source quality, do I trust the source or not.

Another important aspect of the project was that we had to have an experimental evaluation, we had tools and we had standard test sets for evaluating whether a component was adding to performance or not.

This was hugely important because it helped to tell people what to work on. The biggest problem we have is not thinking of ideas or even doing the work, but triaging what is important at this point, what is going to have the most effect on our end-to-end accuracy.

So, I think one of the most important things we learned in developing the system was by having the standard set of questions, in our case it was Jeopardy, if you asked me, I could tell you here are the five most important things we have to work on, because I can look at the set and say what is the head room for possible improvement.

If you work on parsing, how much improvement could that have, if you work on adding semantic relations, 10 semantic relations, how much impact did that have.

Having a shared, end-to-end test set or benchmark allowed everybody to tune their components to best help the system and also help tell them what to work on.

Finally, it is a probabilistic computation, it produces a confidence in all of its answers, so each answer can give you a confidence with the probability it is correct.

It can also tell you why is the answer there, so what are the different sorts of evidence that it used. It can also point to the specific pieces of evidence, so one nice thing is because we often work from text, we can actually pull up the passages, say here is the passage we used, and you can look at it and see whether it correctly interpreted it.

We have begun looking at it I think to the medical domain, and we wanted to find some set of questions, so I think Dr. Siegel brought this up earlier that it would be nice to be able to evaluate different tools. I certainly agree that you would want to evaluate that in a clinical context, but also very useful for developers is to have a standard benchmark set, because what we want to do is know, how do we tune these algorithms to best improve the performance.

So, as soon as we looked at the medical domain, the first thing we did was say we have to find some data. The data that we initially found was this American College of Physicians, Dr. Dilemma questions, which are the sort of Trivia game that they play with medical residents and professors.

These are things like skin rash

associated with Lyme Disease, or type of murmur associated with this conditions is heart systolic and increases in intensive valsalva. So, these are fairly simply questions, they have an agreed correct answer.

They typically have a single answer, not always. Most of the time it is only relevant information is given, and all of the information you need is given should you need the specified answer, but it was a nice test to see how is the system going to perform and what do we need to start working on first.

In the graph on the right there, I am showing you the performance of the system. That lower purple line is the Jeopardy system, so nothing was changed from the Jeopardy system. It is like we literally took it off the set and just gave it these medical questions.

Now, what the graph is showing you is you can answer between zero and 100 percent of the questions. The system ranks the questions based on confidence, so if it answered all 100 percent, the Jeopardy system would get 20 percent correct. If it got to pick the 10 percent it was most confident in, then, we would get something more like 40 percent correct.

Now, what the different lines are showing you is that as we started to improve the system, so we did things like add sources, that is the red line. We added a few of the medical references texts we could get out hands on.

We retrained, so instead of having a system which was trained on Jeopardy questions, we trained on Dr. Dilemma questions, and then we began the work of functional adaptation.

The issue is eventually you get about as many sources as you are going to get. Eventually, machine learning saturates, more training doesn't help you, so you have to work on improving what capabilities the system has. You have to work on the medical reasoning.

That is the blue line right there, so we are right at about 50 percent, which some of the medical teams tell us is about where they are, but I don't have firm numbers on that.

Here are some examples of questions and our answers. So, nasal mucosa, atrophy, foul-smelling crust in the nasal passages, benign cause of congenital and direct hyperbilirubinemia, so these are some of the things that are somewhat useful. I think what might be even more interesting is the technology that goes into solving them.

So, you can use it for question and answering, you can also ask, well, what are the components that we put together to answer these.

Another thing, too, is that from Jeopardy, we kind of inherited this interaction model, if you get a question you give an answer, and we also inherited this focus on giving the correct answer ranked first in your answer list, so making sure that your top answer was correct.

That might not be as important in other domains, so you might be interested in something like this question where the correct answer is in second place. For people who can't see the question, cause of dysphasia that can be relieved in some patients by lifting their arms over their heads or with the valsalva maneuver. Watson said cancer but it did have the correct answer in the second place.

So, how does it actually come up with those answers and justify them? For a question like this, what neurological condition contraindicates the use of bupropion?

We have an NLP stock, much like you have seen. Those are standard things, things like tokenizing, entity detection, parsing, negation detection, relation detection, frame instruction, and so forth.

So, this is the parse of the clue, a syntactic parse, but we will start to understand how these terms relate. We also do things like named entity detection, so we all say ah-ha, the bupropion has COOE, so we know what that is.

Now, we have things like relation detection, so we train semantic relations (inaudible) ahead of time on the most important medical relations for the set, and so we might identify that aha, that is asking for a contraindicator relation. If you have a knowledge base, you can just look that up. Suppose I have UMLS semantic relations, something like that, I can look up that drug, I can try to see if I have associated contraindicate relations for it, and come up with the answer.

One thing that we learned from the Jeopardy experience was that errors propagate, so if you don't get the parse right, then, you don't come up with the correct relation, and you don't segment your argument correctly, you are going to get mistakes, and so the system does not depend on perfectly extracting everything and perfectly coding everything.

We also have a number of passer scores, so we will use text, we will get unstructured content, and you can see different passages here which justify the answer, so it will say things like bupropion is contraindicated in epilepsy, seizure disorder, anorexia, so that is very nice passage, and the parse matches pretty closely with some of the algorithms we use.

You would also have to know things like, well, it has a whole list of conditions here, which one of them is a neurological condition. Well, then you have to go to the instruction resource.

Another question is like Wellbutrin, contraindicated in adults with seizure disorders. Well, you have to know that is a commercial name for bupropion, so you have to go to your background knowledge again.

So, we are always sort of going back and forth between using structured techniques and the structured approaches. What the system allows you to do is basically just plug those in and to some extent train itself on how to use those.

This is nice because it means you don't have to code everything perfectly, and then what we will do is, everything in the NLP stock on the question and everything in the NLP stock on the supporting passages there, we will try to match if we can, so best case we will extract the very rich frame or relation of matches. Maybe we are not able to correctly extract those, well, then, we will use things that are not as sophisticated, things I might use, the named entities that appear in both, things I might use some of the parse, but not the whole parse, does not require a perfect parse.

Then, you have this whole ensemble of scores, which it is going to put out a score on whether it thinks the answer is correct or not, justified by this passage, and the system is going to learn how to combine those.

We have to do this, because we couldn't stick with the purely knowledge base approach because of the propagation of errors.

So, what this means is another important asset that we have been working on is this matching framework, which allows you to basically plug in different algorithms for doing these sorts of matchings on the parse and on the relations and concepts that occur.

If you are going to match things, it can get quite difficult and it is hard to find training sets, and in the medical domain it is very subtle, so, you know, if I am looking for something like enamel erosion, is that the same as tooth enamel erosion.

Well, yes, that is probably always the same thing, but if I am looking for like a yellow, foul-smelling otorrhea, is that the same as a purulent debris. Does it matter if it is yellow or not? If it increases with valsalva versus decreases with valsalva, that is very important.

So, being able to do this matching between two sets of signs and symptoms is actually very subtle, and hard to find training data for, because it really requires some understanding of the domain.

So, this is one of the first tasks we have taken on. Now, it is trying to come up with a training set for this, come up with some -- we have rule-based techniques, we have statistical techniques, we try to code these, so we will run MetaMap and we have an (inaudible) detector we will run over it, we will also not try to code it, we will just try to identify this is a symptom, and syntactically, what are the modifiers, I will just use them.

So, again we are trying to be robust, and not let errors propagate throughout the system.

Look at typing, so in this case, heart is a strategy, or upper airways is a location. We need to know that. We mine that from text. We also use UMLS.

We deal with more difficult passage justifications, so something like attacks of Meniere's disease or precipitated by this dietary indiscretion, passage says a low-salt diet might also help in alleviating the symptoms, so the question is talking about what precipitates an attack, the passage is talking about what alleviates it, so you have to do some reasoning there to decide whether that is justifying or not.

Now, another thing that opens up when we start looking at using Watson in the real context is you can interact with the user, so Watson is able to score its confidence in these interpretations, so it could ask the user, I am not sure what this phrase means, do these two phrases mean the same thing, does contraindicates mean the same thing as should not use.

Based on all of those NLP techniques, we can start to do things like run them over the EMR, we can try to do diagnosis, which uses a lot of the things I talked about, the matching, the extraction together to see if you can come up with the correct diagnosis.

You can also do things like question and answering, so you can use the EMR as your source and say what is this patient allergic to, or what medications have been used for neuropathic pain for this patient, is there any family history of heart disease, that sort of thing, so you can actually pull out and summarize relevant factors for a patient.

Another thing we are trying to do is trying to create a factor extraction timeline construction technique which will show you the timeline. Here is the signs and symptoms the patient has had and when.

We are thinking of using Watson in more of a dialoguing fashion, then, you can start to have Watson come back and ask you questions, so if Watson finds some evidence for an answer, in this case this is a Lyme disease example. This is from an earlier version of the system.

It will come back and say, well, here are some other factors I saw in the description of Lyme disease, circular rash, fatigue, headache. What else should the person have? So, we start to think about using Watson in dialoguing fashion. That means that we are changing the paradigm of Watson from being a question in and an answer out, to letting Watson itself generate questions and get those answered.

So, that is really where the focus of the effort now is on both identifying what are the important missing information, the gaps, and then also what is a convenient way, how do we score those, how do we present this to a user, so that they can actually answer in a useful way for Watson.

To sum it up then, in going beyond Jeopardy, we are dealing with much more complicated artifacts and much more complicated reasoning required to answer the questions. Instead of these simple factoid questions, we are dealing with these very large, as you know, EMRs.

We are looking at instead of having a question in, answer out, how do we interact with Watson, how does Watson propose candidates, how does it identify gaps, how does it ask you questions, how do you encode all of that for Watson.

Then, one thing I haven't talked as much about is the explanation, so when Watson come up with the answer, you can actually score, well, here are the pieces of evidence that gave me the most evidence for these aspects of the answer, so

I can tell you why do I think the condition is harsh, why do I think it increases in intensity with valsalva. I can show you the best scoring passage you can find for that.

Finally, this also affects how Watson learns, so up until now we have always trained Watson in the lab, created a version of the system and then deployed it in the Jeopardy case.

What this allows you to do instead as you are able to interact with Watson, is to do more of the online learning that people have talked about where Watson can look at the responses it is getting, can use that to help tune its algorithms and hopefully increase its accuracy over time.

So, I think to sum up, we are still working very much on the functional stage at this point, and hopefully, we will get to something that is eventually usable and maybe even enjoyable. Thanks.

[Applause.]

MR. LOHR: If the panelists will come up.

Your description of Watson always makes me think about somebody, a long-time predecessor named Fred Jelinek, pioneered in voice recognition, explaining efforts to have machines do what people do. His simple explanation was airplanes don't flap their wings.

We will take questions, if you have one, identify yourself. Please.

DR. RESNIK: Hi there. Thank you for a phenomenal panel. This is really two questions, but I think I can link them through the notion of incentives.

So, for NLP to be valuable in this context, narrative text is important because you have to have narrative text to operate on.

The future of NLP is pushing, like many other things, is driven by data, moving in toward big data, but there are a couple of things going on. On the input side, there is the potential for throwing the narrative text baby out with the bath water in the push to try to accomplish structure data for meaningful use.

One of the things that is very visible here, and in my experience as well, that accessibility of data to do especially on the clinical NLP side you are seeing this in the industry side.

In academia, and other forms of research, it is much more limited except in the context where somebody is affiliated with an academic medical center, for example, so the question is, one, how do you incentivize EHR vendors to recognize and preserve the importance of narrative text? That may be an ONC question.

And how do you incentivize the folks who are doing all this cool work in industry to broaden the perspectives and find ways of engaging more of the natural language process in the community than just the people who are already in bioinformatics departments or associated with academic medical centers?

DR. REIDER: Okay. I will try the ONC part.

I don't know that we want to incentivize

NLP per se. I think we want to allow for NLP per se, which means that we need to be careful not to define how the data gets entered, so we don't need to say it must be structured when it comes in, so much as it needs to be structured at some point.

So, it could be gobbledegook when it goes in if that is how the providers want to put it in, so long as the system, perhaps with NLP, could cause it to be ungobbledegook. That is the technical term I am using. So, I look to the NLP experts to maybe answer the second half.

DR. STEINBERG: From the perspective of the payors, all I would say is we recognize the incremental value of NLP. What we are struggling with is how do -- you know, which we have talked about the old day -- how do you actually get it into a form that we can use.

We have the infrastructure, we believe, to be able to use it, where we are not quite there yet, and hopefully, you know, with technologies like Watson, but perhaps others as well, this will allow us to do this more effectively.

But the will is already there, and it is there, people are voting with their feet and with their wallets, I mean ultimately, the people who are paying for these services want the services to be good, and they realize that for the service to be good, for the output to be accurate, having more data is better.

So, they are already there intellectually. I think it is a matter of the missing link is getting from the unstructured data, such as it is today, to a way that we can use it.

MR. JAGANNATHAN: I have a followup question on the same topic. If you look at the certification criteria for EHR, it doesn't really say that you can use NLP, and if it had some indication that you can use NLP, and it is brought to light as part of the certification criteria, not all EHR under Siemens excluded will not be going around saying physicians should be entering into all this wonderful little check boxes.

So, I know you don't have to explicitly say anything, but at least if the specification has in it some bias towards allowing free text and narrative text, it will make a big difference, and if you go and look at the missed criteria for usability, it doesn't have those wonderful pictures you just put up, and it really reads like they want these things to be entered one by one, and they are worrying about the safety and things like that, so nothing in the specification for certification of EHR really allows you to -- if you are a EHR render to think of things to be narrative text.

That is a failing I think on the part of the certification part of it. At least that is my humble opinion.

DR. REIDER: Noted.

MR. WEITZMAN: Steve Weitzman, Dataform Foundation.

My question is to Dr. Gondek. Where

would you use natural language processing in the learning health system described this morning?

DR. GONDEK: I think that I said a little bit during the talk. We don't expect the NLP and Watson or the Watson system to have the sort of deep understanding that a practitioner would have to understand context to represent the interaction of many different factors.

What we do think that it is good at is running up a lot of data. We think it has richer matching and richer understanding something like keyword search. So, we are what is so interesting then is if you are searching for something, you don't know exactly how it is represented, you don't want to miss it, so it is an important question to you.

So, we talk about things like these high value questions.

MR. WEITZMAN: Can I follow up with one question? Can you use the IBM system to code medical records again using UMLS, and give me the codes, and embed them into my medical record?

DR. GONDEK: Yes, the system does coding. It does coding. It has not been our main focus, because we think that from our experience it is always difficult to get the ontology as rich as you want, and you are never going to get it as rich as you want, and you are never going to be able to match it exactly like you need when you are using it.

So, I think some of these examples that were given today were that you have to triage. You have to pick what the 1,400 codes that you are willing to implement.

Part of the Watson view I think is a longer term picture. We are kind of getting to the point where the system itself can make these, do these two things mean the same thing, maybe not make a binary decision, but sort of have a richer representation and try to understand some of the same subtleties that humans are doing when they are reading the text.

DR. MEYSTRE: I have another question about Watson also, and a very nice presentation, by the way.

In general, it seems that humans have a hard time to deal with probabilities. So, I am wondering how you present answers to clinicians who are answering questions, do you just pick the best, the highest probability, or do you give them all the information and then try and figure how they deal with this information with these probabilities.

DR. GONDEK: It's something that we don't have a user interface or a CDS system at all. You know, our expertise is really much on the NLP side, and we are trying to bridge to that.

What we do, though, is I think you are absolutely right, that you look at how Watson parses its knowledge bases when it does, and we use statistical techniques that are based on thousands of passages, and you can't show the doctors thousands of passages.

On the other hand, it is also an asset in that whereas typically, a knowledge base is going to have a binary rule, we have probabilities associated with these, so we know what is more common, we know when it was said, we know what sources that were said.

So, one thing that we are looking at doing is if I give you a fact, so I say that this is a treatment for this, can Watson find the best passage justifying the fact.

What that means is even if Watson gets to a very complex statistical technique, it can retrieve a passage which should ideally be convincing to the user, and that requires setting up training data, so we need to go to assess is this a convincing passage, is it a reliable source, is it timely, and so forth.

It is something that we haven't, you know, in just answering questions, you don't deal with that, you just care if the answer is correct or not, but now we are actually extending the framework to evaluate is this evidence reliable or not, is this the best passage that we could have shown.

DR. MIDDLETON: Loved the presentations,

and the whole panel. My question, though, is really for Marc and David perhaps. I wonder about the edges of reasoning between the NLP methods, the statistical methods, how will we get at the deeper reasoning methods like, you know, anatomic reasoning, or pathophysiologic reasoning or other forms of systematic reasoning in either semantic modeling underneath record or the semantic and knowledge modeling underneath Watson.

DR. GONDEK: I think that the models that seemed to be emerging across the board are these layered kind of approaches or ensemble approaches where there is no one of them. Then, you have to combine those and sometimes iterate through them until you get to some kind of stopping rule.

So, I think that when we will get to those is when we have gotten some of the basics better baked, so that then those things become more useful as a supplement to those.

DR. OVERHAGE: I will also add that I really appreciated Marc's talk because of the sort of in-depth models they were building and trying to deal with the interaction between models. What I see is I see that work approaching it from a rule-based background and making those rule bases richer, more looking at the joints between them, how they interact, whereas, I see our work as coming more from the unstructured side where we don't have those rich models, and this was a conscious choice we made, so we took those Dr. Dilemma questions.

We could have said we are not doing Dr. Dilemma, we are going to do diagnosis of a specific condition, and then we would have taken a very different approach. We would have built the hand-tuned models using machine learning, and so forth.

The fact that for the team, we decided we are going to take this general task, means that the type of techniques we are developing are different, which are much more based on trying to interpret those models from the language on the passage side.

There is no way it is going to be as good as a human at this point, so I think that I wouldn't look to Watson to be doing that sort of deep, having large decision trees sorts or reasoning. Rather, I look at it more for coverage and see if that shallow reasoning gets smarter over time.

DR. MIDDLETON: It's interesting that you use the probabilistic reasoning, as well. Is there a utility function considered in terms of, you know, guiding search or guiding question and answering, because it seems like you might be able to prioritize certain directions and paths with either a single attribute or multi-attribute utility model.

DR. GONDEK: Yes, definitely. One thing we do when we get a question is we break it up into factors, and that is based on syntactic parse, also identities factors, and we learn from data which are the important factors to answer.

So, maybe if you have, you know, you have some symptom that is not very specific, you wouldn't be using that, or not very sensitive, you wouldn't be using that.

Now, we have those techniques identified, working on parts of a passage, you know, what are the key terms, how important are they for answering questions.

The probabilistic -- one thing that happens in a probabilistic computation is that it is able to take different interpretations of the question, so we can say we are going to interpret this as a relation, or we are going to interpret this as contraindicates relation, and see what happens, try to explore different interpretations and then in the Jeopardy case, we didn't get a chance to revise our answer, it was basically question and answer out.

But what you could do here is because of the probabilities, you can say here are what I think are the important factors, here are what I think are the important relations, and then a user could say, well, no, actually, I disagree with that, that is not as important as your weighing it.

MR. SOBOROFF: Ian Soboroff from NIST.

One of the neat things from my perspective about Watson is that it built on the shoulders of a bunch of giants. One of those giants is NLP at 40 or 50 years of research on that.

Another is about five or six intelligence community advanced research projects that funnel millions of dollars, listed all the technology behind them, extraction, question and answering, and couple with evaluations that showed how you could tell if those things were working, how you could measure the progress of them.

It seems to me -- I am not a doctor or not that kind of doctor -- that the community in this room doesn't see how clinical decision support makes the same leaps, gets the same way forward except from the companies, the Siemens or Aetna, who have budgets and research teams and access to huge amounts of data.

But I wonder what is the research program that gets CDS to a Watson?

DR. OVERHAGE: I might take a little bit of a stab at that, and we heard a little bit about some of the kind of grand challenges from some of the other speakers.

It seems to me that there is a couple of major pieces, and Bob had his 10 areas, or whatever, but one area of research is clearly in the human-computer interaction aspect, which is still a huge challenge, how to deliver this in a way that is not interruptive, whether it is GPS or however you want to conceptualize it.

That is an area that clearly is still an unsolved problem.

The second I think big insult problem is related to a lot of what we have been talking about this afternoon, which is how do you represent the knowledge or the rules or whatever it ends up being in a way that are editable and creatable, which is I think a real challenge. Even if we come up with this really cool way to do it, it is going to take us 10 years to train up a cadre of people who understand it conceptually, have the clinical knowledge and the technological translation abilities, so I think that is a huge area.

So, even if we have the perfect mousetrap of how to do it, I don't know that we would have the human and process engineering solved about how to do it.

The third thing I think -- and I am simplifying it in some ways -- the third line of research is I think how we integrate these things into workflow and process in a high reliability way.

I think we have learned a lot about some of the other areas, but how do you make that sing in a collaborative, multi-user -- and Bob alluded to this a bit -- environment, and you did also, Greg, you know, how you make that fit into an environment where it is not just one person that you are supporting, but a team and a collaborative that you are supporting.

It seems to me those are three big areas, and the answers on any one of those can happen independently, they don't all have to move together to get progress, but, you know, it seems like those are the things that I would if I were advising somebody what kind of research program, and CDS, those would be the three big ones.

DR. STEINBERG: I would just add I think you were really getting to the funding issue, I think, or that was at least part of the question, right?

At least to the funding, I think the public/private partnership route is really where I believe this is going to have to go, and is already moving.

I mean as an example, Aetna is involved in the mini-sentinel project along with a number of academic institutions, and it is that kind of public/private partnership where there is some funding from the Government and some funding from industry towards hopefully -- and this is going to sound like mom and apple pie -- but towards the greater good, that I think that is going to be the impetus.

DR. RIPPEN: I guess the other thing is to consider the perspective of how best to leverage it, because it can be a pretty amazing tool. The question is what is the best application.

So, for example, if we know 93 percent accuracy, that it relates to billing, which is better than most coders at times, is that really good for coding, and then also billing, but then also for reporting, for example, quality reports.

If you start talking about 93 percent accuracy, and now you are doing clinical decision support with no learned intermediary, where now you are making medical decisions, you may now sway, because generally, we don't tolerate 93 percent if there is some assistive device.

So, again, I think how do we apply it, how do we leverage it, how do we understand kind of the benefits and the risks, I think is an important part of it, too.

DR. STEINBERG: Well, to be fair, I

think, first of all, we would never say that we do clinical decision support. We are providing information to a provider, and he or she makes a decision based on their knowledge and judgment as to whether or not they want to proceed.

We say, look, we have access to information whether it is on the data side, the knowledge side, or both, that you may not have. It is here for you to figure out whether you want to go forward or not, but we are not -- as one doctor sort of called us up and said, look, I will listen to you if you can tell me the color of the hair of the patient in front of me. So, we can't do that.

DR. REIDER: Although I would argue that there are EHR systems, perhaps not anyone from the table here -- that I have seen, that actually do a little bit more than that, and so you may, in a hospital setting, have, if you are familiar with the way the systems work in some hospitals, there is dependent and independent protocols, so an independent protocol is something that a nurse or the system can do.

If a patient meets certain criteria, without an order from a clinician, the nurse can actually do something, and I have seen systems that actually do that work, so a patient will have a certain diagnosis, like diabetes, the diabetic order or meal plan gets automatically ordered with no human intervention, and so where do you draw the line between the order for the diabetic meal and the TPA that you administer for the patient's MI.

Certainly, there is a line somewhere in there. So, sometimes decision support in this might get into the discussion earlier about someone regulating this in some way. I won't name the three-letter agency, but this is important and of deep interest to us and, in fact, the public.

I will also pitch something that I neglected to say earlier when I said "Noted," officially, I did not note that. We are in the rulemaking process, which means that our proposed rules have been published, and our final rules have not.

The comment period is still open, so if anybody has opinions on how ONC and/or CMS should enhance the likelihood that NLP becomes incorporated, go to regulation.gov by May 7th, and let us know.

MR. LOHR: I think we are running a little over on time, so if you have any questions or comments, the panel is still here.

Thanks very much.

[Applause.]

DR. LUO: We will take a 5-minute quick break, and we will come right back, and at the same time, we will set up the station for the next session.

[Recess.]

Panel 4: Future Challenge and Opportunities

Moderator: Dr. Vinay Pai, NIBIB

DR. PAI: This will be the last panel for this workshop. This panel will be having one speaker as the keynote, and then we will have a session where all the speakers of the day are going to sit together and people can ask questions.

This panel is about the future challenges and opportunities for natural language processing and clinical decision system support.

The keynote speaker is Dr. Jon White, who heads the Commission Technology Section, as you can see up here, for AHRQ. He did his training in family medicine at the University of Virginia, and also residency at Lancaster General in Pennsylvania.

He has won the National AAFB award for excellence in graduate education.

[Applause.]

Dr. Jonathan White, AHRQ

DR. WHITE: Thank you so much and thank you for sticking through to the end of the day. It has been quite a day. There has been a lot of great, great discussions.

I promise you I will not overburden you with deep thoughts. This is the end of the day, and this is meant to kind of pull it together and try to look at a little bit further ahead.

I do want to say that this is a great honor. This is the first time I have had the opportunity to speak at Lister Hill, so thank you very much for being wonderful hosts.

Preparing for this talk was actually a lot of fun, because it got me thinking about language and how we use language, and the ways in which we try to interpret it, and I decided the right approach was going to be to use other people's language, because my language is terrible.

You will see liberally sprinkled through here quotes about language that I hope will kind of guide us on our talk.

I was most excited about this quote. This is Galen 2,000 years ago talking about language. I thought, oh, my word, this is fantastic.

"The chief merit of language is clearness, and we know that nothing detracts so much from this as do unfamiliar terms." My God, 2,000 years, there is nothing new under the sun if Galen talked about this 2,000 years ago.

In terms of thinking about challenges, research opportunities and where NOP fits in all of this, I tried to frame it up in my head, and I thought, well, you know, I come from AHRQ, it's about quality.

So, I wanted to talk a little bit about where we are trying to get to. Like I said, I promise I will not overly burden you with this, but I want to talk about two things: National Quality Strategy and our National Healthcare Quality and Disparities Reports.

We have talked a lot about quality for a long period of time. The excellent Dr. Reider put up Crossing the Quality Chasm 2.0. The topic of quality has been chewed over significantly, but prior to about two years ago, we didn't have one solid thing that we could say this is what we are trying to do, and this is what we are getting at.

The Affordable Care Act did establish such a thing called the National Quality Strategy builds on the work of a lot of different folks, many of you who probably know them, and moves to a patient-focused approach. It can be found out, if you want to look at the whole thing, you can find a link at the bottom.

There are three aims, six priorities, and 10 levers by which you get there. I am not going to read you this word for word. The three aims you have probably heard of: Better care, improving overall quality. The second aim is healthy people, healthy communities. The third aim is affordable care. We can get deep into those if you want to, but everybody is again like mom and apple pie as one of the speakers talked about.

Underlying those aims, well, how do you prioritize within that what to do, and they said, well, here are six priorities. Again, not word for word, but I want to point these out, because they come up later in the Quality and Disparities

Report.

Reducing harm, patient engagement, effective communication and coordination of care, effective prevention and treatment practices, why these are best practices, and making quality are more affordable.

Again, makes some sense, but it helps you say okay, if we want to do this or do that, how do we prioritize.

So, I am not going to get into the 10 levers, but will mention that Health IT is specifically mentioned as one of the 10 levers that we are supposed to use to move quality ahead.

So, again, why do we do all this? To try to improve the quality, better care, healthy people, healthy communities.

This is something that AHRQ does every year. Most recent editions were released this past Friday, National Healthcare Quality Report and Disparities Report, they are sent to Congress as part of AHRQ's authorizing legislation.

Overall quality improves very slowly, a little trudge at a time.

Notably, this year we noted in the Disparities Report, though, that access to healthcare is not improving for most racial and ethnic groups, and, in fact, in some places it is moving backwards. So, that is worth taking away for you.

On average, most people get most of the care they are supposed to get most of the time. Cardiac care has actually been a bright spot, the place where we have kind of been making some significant gains.

If you care to go look at National Healthcare Quality Report, Disparities Report, there is data on the adoption of EHRs, which is something that we are including. We say this the mark of an organization that has good information tools, that has high quality care.

Again, I just want to point out those six priorities. The Quality Report and the Disparities Report are going to be aligned with those, so it all good and well to talk about the strategy, but again you can't prove what you don't measure, so we are going to start measuring relative to that.

I told you I was not going to overburden you, but that is kind of what you are aiming at.

This was a great writer that I stumbled across in looking for this.

We talk a lot about quality and all the different components of it, and stuff like that, and sometimes I get befuddled, so I ran across this, I thought, oh, perfect.

"It is the great mystery of life itself which is at the bottom of all the mysterious language we are obliged to employ concerning it." That is what I think about when I talk about quality too much, it is like, you know, quality is at the bottom of this.

So, those are the challenges.

Research Opportunities. I don't know if there are any of you in the audience who are looking for new research opportunities, but if there are, let us talk about a few of them.

I am very pleased to be able to tell you that just today, posted on the AHRQ web site, there is a new evidence report, basically, a systematic analysis on the impact of clinical decision support systems done at Duke University, a number of folks in the audience participated in the expert panel to help guide the work.

An article was published yesterday on line in Annalsofinternalmedicine@annals.org, and basically find that after review, 150 different studies that are out there, clinical decision support is shown to improve process measures, not just in the academic centers where they have been shown to previously, but across a variety of different systems in a variety of different settings.

So, this is good. This is the first time we can take a look and say, you know, this does make us better. Where there is limited evidence, okay, is in terms of clinical outcomes, economic outcomes, workload, that says "electronic," it is supposed to say "economic," and other types of outcomes.

So, getting all the way down to living longer, suffering less, that sort of kind of an outcome, getting better value for your dollar. It doesn't say one way or the other, it doesn't say it doesn't improve those but there is just not enough evidence about that.

So, there is a research opportunity for you to take a look at, this is important, how it affects clinical outcomes, economic outcomes, workflow outcomes.

Basically, there are also several features of implementation that are identified as leading to improved impact of clinical decision support.

So, AHRQ.gov, or if you to annals.org, there is a nice article about it.

So, that is one opportunity right there for you.

Clayton Christensen. I have thrown up a lot of -- and there is going to be a lot more, you know, people up there, but you have got to have a little kind of pop business psychology in here, too.

I actually really like this. There is a lot of good things to recommend it in The Innovator's Prescription, which is a good book by Clayton Christensen, but this one is "The graveyard of failed products and services is populated by things that people should have wanted...understanding the job that customers are trying to do is a major issue in every healthcare innovation."

Okay. So, he talks about a milkshake, and when you buy a milkshake, what job are you trying to get the milkshake to do. Let us not down that road, but when we talk about decision support and, you know, this is great, it is going to change the world, if only people would use it. Why aren't they using my decision support, right?

Well, there may be a reason why they are not using your decision support, and it is probably related to the job that they are trying to do, and how does it help them do that job.

So, let's talk about another funding opportunity and another opportunity for future research. The funding opportunities that we put out actually a little over or almost a year ago, but that are really good, they are a little more basic science than even AHRQ is used to. AHRQ is used to a lot of applied stuff, a lot of demonstration stuff, how to improve quality.

These are opportunities to define what is that job, right? Understanding the job to the customers, and we talk about customers. I am not just talking about the doctors, but I am talking about everybody in health care because we need to provide decisions for it, not just doctors. You heard some of that from the folks that have gone before.

So, two funding opportunities, Understanding Clinical Information Needs and Health Care Decision Making Processes in the context of Health Information Technology, not I have a widget, and I want to do something with this widget, understanding the information needs and the decision making process, because now we have good information tools, and we have good information systems.

How does that change our job as clinicians, as patients, as caregivers, right? When you mom calls you and says, "I had this weird test result," how does that change because of the information tools and systems that are available to you.

So, those opportunities are out there, they are going to be out there for a while, and we hope folks will come in with good opportunities for that.

Coldridge. "The best part of human language, properly so called, is derived from reflection on the acts of the mind itself." So, these words that we are trying to process, okay, and by the way, I am going to get into the issue of whether or not it is actually natural language, because I don't think it is.

But the words are really our thoughts made manifest in a lot of different ways. So when we make those thoughts manifest, are we doing them in a way that helps us support decisions. Another opportunity for people thinking about how to do this.

The IOM issued a report in late 2011 called "Clinical Practice Guidelines we can Trust." Some of our colleagues have worked extensively on that report. Two recommendations that are key in there for the CDS community: Guideline developers should structure the format, vocabulary, and content of CPGs to facilitate ready implementation of CDS.

So, this is I am going to take language and I am going to get a big computer somewhere, and I am going to process the hell out of it. I am going to figure out patterns that are in there.

We need to push on the other end is what the IOM is saying. We need to take the language that we are using and we need to structure it more. Now, I am not going to beat on that anymore. You have heard a lot of that here, but the IOM itself is saying, you know, this something we ought to be doing.

The second recommendation, also kind of key, guideline developers, guideline implementers, and decision support designers should collaborate in an effort to align their needs with one another, because you don't now, or not as much as you should. Okay, there is some.

As we try to take clinical knowledge and translate it into how we support decisions, we don't talk enough, so that is something that we need to do, opportunity to move ahead.

Last part. Streaking along here.

Natural language processing.

Lavoisier. Any chemists in the room? Yes.

"It is impossible to dissociate language from science or science from language... To call forth a concept a word is needed; to portray a phenomenon, a concept is needed. All three mirror one and the same reality."

So, language, okay, that we use to describe events and healthcare conditions, what is going on, is meant to represent something, and it is representing health, and it is representing sickness, and it is representing that things are happening in health care and in people's lives.

I don't like the term "Big Data." I think it is fuzzy. I think it is overused, and I think that people, you know, like the cloud. I think there is a lack of precision in that term, but I did make a slide title "Big Data."

So, what do we hope to gain from natural language processing? Well, we have all talked about it here. There is a huge swath of healthcare data that is not structured, it is in narrative form. This is how I used to do it. It goes onto the dictation and then comes back.

We hope to get information that is out there from the system, pull it in. That is for the afferent loop. Process it and spit it out the other end.

There is two things that I think that we are trying to get from this. One, we are trying to look across all this big data that we haven't been able to analyze before, and we are trying to discern patterns in it. We just heard some of that up here, is that, you know, try to identify issues that us mere mortals that can only keep 4,000 patients in our head at any one time might not be able to discern otherwise, and, you know, all the issues with collecting data to try to be able to determine effects, maybe we can find it in the narrative.

So, I think there is some promise to be had there.

And then there is the efferent loop. There is John sitting there and see a patient, and he can't remember something, so he turns to Watson or whoever, and he says, "Watson, tell me," and Watson says, "to what you said, and I think you ought to do better" da-da-da-da.

So, there is the either and out to whether it is the clinician or whether it's the patient or whether it's the policymaker or the caregiver or whoever, there is a loop out, and rather than have to look for it in a certain way, we would like to have it come back to us in a way that we can understand, or at least that Alex Trabek can understand.

So, bear with me for a moment. I know that the Watson that we just discussed is not Arthur Conan Doyle's Watson. I know it's the IDM Watson. But six months ago, I am at with Jon-- you were sitting next to me, Blackford at AMIA. He is talking about Watson, so he sketched up Watson. And yes, that is the sketched Watson. The number is imaginary down there, I know it was 77,000, not 36,000.

I am looking at it and I am thinking Dr. Watson, Dr. Watson, I am trying to imagine going to see Dr. Watson, I am thinking, you know, there are some things that Dr. Watson might help me with, but there is something missing just from my, you know, all the days of seeing patients and being trained to be a clinician, and my days of being a patient, there is something missing.

So, I started sketching on the other side, and I started sketching Sherlock, and I am like, oh, okay, so where am I going with this.

The question was asked, you know, so Watson is a first year medical student now, right, was the way you phrased it, and what will happen when Watson gets through medical school. I really think that Watson and technologies like Watson have tremendous promise, okay, and I think David correctly observed here, and I don't know if you are still here, or if you have gone off to talk to people, but correctly observed that really there are certain things you can expect from it, and certain things that you can't.

This is something that a lot of folks discuss when we get down here, so what is it in the counterpart to Watson in the clinician that you need. My sense is a couple things. I threw this up here. There is actually a quote from Sherlock Holmes.

It says, "There are 50 who can reason

synthetically for one who can reason analytically."

I started trying to pick that apart, and like I said, we can talk about it more, but the idea is that an analytic thinker or an analytic reasoner starts with a conclusion or the premise and then works backwards, decides how it happened and how did that dead body get in that closed room sort of a thing versus A equals B, B equals C, therefore, A equals C. That is synthetic reasoning. You pull together a lot of things.

It seemed to me -- and it was actually reinforced to me by the discussion we just had about Watson, that, you know, Watson is pulling together a lot of data and looking for matches and then looking for patterns in that, and trying to apply it to what you are seeing there, and help you determine that outcome, and, you know, all the different ways we just talked about Watson helps guide you down the path, that we shouldn't expect it to, you know, go a whole lot deeper.

There is also, you know, we talk about the difference between science of medicine and the art of medicine. There is still a lot of art in medicine, and frankly, there is a lot of art in the language that we use in healthcare.

All of you know this. Part of our limitations in natural language processing is the imprecision with which we use our language, and frankly, to go back to Lavoisier, the imprecision that we have in our concepts which are constantly evolving, and our understanding of the phenomenon, so to the degree that we don't have a good solid grounding in those, we can't expect a good national language processing to be able to do things for us kind of automagically, right?

That said, there are amazing things happening. I look forward to a discussion about it further.

Finally, where do we go? What do we need to do? I was digging around and my Bartlett's Familiar Quotations, and I ran across T.S. Eliot.

I am not sure who the poets are here, but

I will read it.

"Poets in our civilization, as it exists at present, must be difficult... The poet must become more and more comprehensive, more allusive, more indirect, in order to force, to dislocate if necessary, language into its meaning."

Really, we are getting at the meaning of language, and that is what we are trying to pull out of the processing part of it, okay, is the meaning, and that translates into better health care, which is better quality, which is ultimately what we are trying to get at, so I thank you very much for your time and attention, and I will look forward to a great discussion.

[Applause.]

DR. PAI: Are there any questions?

[No response.]

DR. PAI: Can I get all the speakers from the previous sessions together for the final discussion?

Basically, we will have some kind of

discussion or where we see NLP and CDS going, and I guess one of the questions I would like what future do you see for that, and what are the advances needed in broad fields for healthcare to benefit.

The other question is where do you see like NIH funding coming in or helping out for this process to move forward.

Blackford, do you want to start?

DR. MIDDLETON: I am sorry, I didn't catch the whole question.

DR. PAI: My question is like where do we want to see like NLP and CDS go in five to 10 years from now, and the second question would be where do you see like NIH's role in moving it forward.

DR. MIDDLETON: I am happy to start. You know, it has been such a terrific day listening and learning from both NLP investigators and CDS investigators. I guess if I was really to try to paint the picture 10 years out, I am not sure if it's the Greek oracle model Randy Miller has written about, I am not sure if it's the dialog with Watson or if it's a dialog between the patient and Watson, or if it's a three-way dialog between patient, Watson, and the physician.

I think the NLP role obviously is to both inform models, inform knowledge structures, and drive correlations. Just like genomewide area association studies, GWAS studies, we are still trying to connect the dots in language as Jon White just pointed out, and find meaning in words, and find the connections in the corpus of words.

So, 10 years out, I think we have to have a much more informed clinical practice. There is just simply way too much to know. We need to have the tools assisting the clinician in synthesizing and summarizing the patient's state. I think the physician needs to have many, many more predictive tools to help him or her interpret the patient's trajectory through a plan in health and wellness.

I think the same set of tools actually

probably will be offered to the patient, him or herself when appropriate obviously or to a caregiver or proxy for the patient, because after all, that is for whom it really matters.

The knowledge base, gosh, when I was listening to David Gondek, if I understood all the correctly, and there was a lot, you know, perhaps that thing will become really the medical Syborg. Maybe it really is the terminator in the best sense of the word, because if it starts to learn and it continues to learn, then, you know, we will be way beyond Neva Ponti's inflection point or the singularity as Ray Caswell wrote about, and the machine will be way smarter than all the doctors, and that might be okay.

It really might be okay, but we will be still in a consultative role, and I know, as Jon White said again, as both patient and caregiver, the hands-on role is still going to be important. It will be different perhaps than the cognitive and hands-on role we have now, but I think we need to have the NLP, the CDS, the synthesis, the summary, and this ever-expanding knowledge base, and oh, by the way, that will lead to discovery in ways that we haven't thought of yet.

DR. WHITE: I want to try to focus you back on the whole job thing. There is science to be had here, right, understanding the semantics of language and the meaning behind it, but then there is what you are trying to do, okay, in healthcare.

As a patient, I have got a couple jobs, right, if I am not well, I want to get better, and how can NLP tools help me do that. I don't think we totally know yet, I don't think we have clearly stated that.

As a clinician, I have got a job to do. Now that role may change over time, but I have got a job to do. What do I need that is in language form now, not in structured data form, that I need to get at through analytic tools, to be able to do my job, and give me the information that I need as a payor.

What job do I have as the steward of

folks' resources to help them get better care, what am I trying to get at. I can see a lot more clear directions there, right, how am I going to sift through all this data, I need good tools to be able to do that.

Or people who are setting policy for a care organization or other kind of large organization or purchaser of care. I think if you get back to that and you say here are problems in healthcare, and here is how NLP can fix that, I think a precise statement of that will get you a long way.

DR. GREENES: I think one of the reasons that we are struggling with this question is that we are still in the process of kind of incrementally reengineering a legacy healthcare environment.

So, a lot of information that is now captured in prose that we would like to have structured isn't, a lot of the nuances that we would like to be able to capture that are and always will be in natural language haven't been approached.

So, what Jon is saying is put needs first, you know, the design you would like to have, and if we could sort of suspend healthcare for a decade, and design a system we want, could we get there by incrementally improving the system we have, or should we be rethinking, and then creating that goal architecture, that goal environment, and then trying to move the system toward it.

We are not going to suspend it, so can we come up with that vision and then move toward that?

DR. REIDER: I am trying to think about the original question, which I think if I am remembering right, was what sort of research might we think about doing. I was struck by a hallway conversation a few minutes ago that reminded me of some of the opportunities for using the nuances that are often missed when we try and do structured text entry.

To Marc's point earlier, about looking

under the streetlight, so we often look where there is lots of data, and I was impressed actually at the Watson team, you know, is looking where there is data, but in fact, there are interactions that healthcare providers have with each other, similar to the vignette that I described where the patient and the provider are electronically communicating, and I thought about the signout event where we actually are very efficient in conveying information to each other, right, so signout for those who don't know is when I am on call for the weekend, and my colleague tells me about the seven patients that he or she is especially worried about, or I am a resident and I am signing out to the night resident who is on call tonight, and I tell them about the 26 patients in the hospital, and often I will even write little vignettes about them and bypass my information systems.

Those actually might be fascinating places for us to look for very efficient use of language in conveying important things about patients, so I guess that might be an interesting place to look that is not quite under the streetlight, because we are probably not capturing that.

I think maybe the other related question is are there barriers to these sorts of things, so as we move toward too much or as we move toward the capture of too much information, I can think of two reasons for it.

One, we have incentivized it, right, with billing based on capturing too much data, and we also have a legal system that docs are afraid of, and therefore, are accustomed to capturing too much data, because they think that that is necessary to avoid being sued for some reason, so are those barriers, and if so, how do we handle them?

DR. CORN: Thank you. You throw away the lawyers.

DR. PAI: I have another question, which I want to be a devil's advocate. Suppose we get this whole thing done, and NLP-CDS works great, and this goes back to the analogy of the GPS system. Supposing the doctor, the NLP-CDS system shuts down on him, there still are the analytical ability left to him or her to make decisions.

DR. MIDDLETON: I have a story to tell. So, I bought a GPS system for my wife when we moved to Boston from the West Coast. She thought it was pretty cool, because, you know, you can go to downtown Boston and get lost, and it is horrendous if you are not from there.

One day she told me, you know, this GPS thing it really stinks. I said, "Well, what is the matter?"

"Well, I am putting in the location whenever I have to go downtown, and I am getting taken about a mile away. repeatedly." I thought, oh, you just are not using it right. I got my 20 lashes for that, but after a few more episodes of it not working, I figured out that it really was a knowledge base update problem.

There had been a knowledge base update, I neglected to put it in, and thus, once the new knowledge base was uploaded to the system, of course, everything was fine, and the GPS worked, and I had meals at nighttime again.

But to address the question I think oftentimes we think that these cognitive aids become crutches. I have never actually ever found that to be the case. Using a calculator in high school, using any kind of computing tools I could get my hands on to facilitate my learning in college and graduate school, et cetera, and ever since I think the experience suggests that it may change one aspect of reasoning or decisionmaking to have a decision aid or these cognitive kind of support tools in your hands, but on the other hand, it dramatically extends another form of reasoning, that is, the reasoning that maybe considers all possibilities or the miracle of different functions and formulating, and whatnot, so I actually don't think it is a loss necessarily, I think it is a win and loss, that these decision aids will extend reasoning and our capabilities in some interesting and yet to be described ways, while we may actually forget the Kreb cycle who really give a damn.

If I can reason about the Kreb cycle, and its relevance to the disease at hand, then, that is fine.

DR. SONNENBERG: I will just add a comment. I agree with what Blackford just said about the systems not really completely being a substitute for our own reasoning, but I think one thing that they can do is if they function well, they can teach us better ways of doing things as we use them, and that will make us better off when we find ourselves without them temporarily.

DR. CORN: I would like to make a comment about Watson suddenly rising up, and, you know, the New York Times after that Jeopardy thing, I think captured it pretty well. It said Watson doesn't even know it won.

[Laughter.]

DR. CORN: I would be interested simply because it is the end of the day, in a little comment from any of the panel members here, because you all have rather top/down positions in policy or in some of these other things, there is a little bit of a tone here of you all are going to figure out how to do it, and you are going to make them love it. I think that Bob's idea really crystallized that very well, if we could only suspend everybody for 10 years, and then we will produce Camelot, and everyone will live in it.

I understand the point that you are making, of course, we are working with hostile system, but I don't think that we are going to get very far sitting here and implying vaguely that the medical professionals, the doctors and the nurses, are in somehow lead ice.

I haven't heard enough about how we work with them to see what it is that they want to do their job better.

DR. GREENES: I have an anecdote about that. When our organization moved to Mayo Clinic, you know, my department, immediately, I began to have a parade of docs that were frustrated with their shall be nameless information system.

They couldn't do X or Y or Z, and were very frustrated with it, and they came up with ideas about how they could have a work-around, or how they could do something that they can't do with that system.

The message I got was they really want to have an information system that would be useful to them.

DR. CORN: I am not talking about information systems, I am talking about helping them to do a better job by giving them clinical decision support.

DR. GREENES: That is part of what they are asking for. They are asking for ways to assess the data that they have got, they have got ways to enter the data that they need to enter, ways to manage the patient's problems. You know, these all decision support potential targets.

I view the more advanced decision support is just an extension to that, that they are not even thinking of yet, because they can't do them in their environment.

So, what you want to do is change the dynamic where instead of the IT system being something that is thrust upon them, have it be something that they really can help, can buy into, and help foster the nature of, you know, that they can actually help bring about.

DR. REIDER: I agree with you, I think the direction that you are going, and I am not sure it is in the scope of NIH right now. I think this is actually more in the scope of, I don't know, Wharton, or Harvard Business School. This is a Marketing question, Marketing with a Capital M. I leaned when I lived in the product world that Marketing with a Capital M is understanding the needs of the market, and then addressing them.

When I was in high school, I thought that marketing was market communications, which is pitching the stuff we made. To some degree, I was reflecting on that when I was sitting in the back earlier, that to some degree this is a solution looking for a problem.

One of the real problems that the market understands, so that we can then work on addressing them, and over time perhaps the two may meet. Calling an NLP and CDS is actually the right way to stop the conversation at the beginning.

We need to go out and lead with our ears, as Steve accused me of doing earlier, and really listen and deeply understand what it is that the market needs, and then work very slowly and deliberately to see if some of the stuff that has been build over the years has application rights. So, Siri is a good example of something that the market understands, right, so if I ask my mom, has she ever used natural language processing, she would say no.

If I asked my mom if she has used Seri, she would say yes, and so obviously, we know from all of the anecdotes and the jokes on Saturday Night Live, that Seri isn't quite perfect yet, but it is an application that is the sum of all of its work that has actually met a consumer need, which is make some stuff easier.

What are the other stuff that help care providers need to be easier and better?

DR. WHITE: Just real quick. I put up the Clayton Christensen report, so I am with you. I want to urge and a caution that if you only think about the needs of certain sectors of the market, you run the risk of meeting those needs, so my pop culture reference here is the Simpson episode where Homer finds his lost brother, runs the car company, and he says, "Homer, I want you to design me a car."

Homer goes off and he designs a car, and it is this big monstrosity that has got a bubble and like a coffee maker, and doughnuts, and the guy is like, "You have ruined me, you have run my company to the ground."

You know, I worry that if you ask one specific statement on the market you are going to get, you know, the Homer.

DR. REIDER: Wants and needs are

different, right? So, we don't build them what they ask for, we build them what they need, which means we have to actually think after we have asked them. Anybody who has worked in product management know that, right?

We ask our customers, and then we actually act on what they didn't say, because otherwise, Steve Jobs never would have built an I-pad, right?

DR. MIDDLETON: To add to the levity, I will add a serious note.

DR. WHITE: I am serious.

DR. MIDDLETON: One of the challenges -- it's a great question -- I think one of the challenges is -- or two thoughts. One is that actually, I have seen now in sort of three evolutions of EMRs that I have been personally involved in, that once the tool is well understood and really you know how to use it, then, the physician takes over, and the clinician, who is interested in patient care, interested in the well-being of his patient, interested in populations and asking questions assumes the professional role of caring about his patients using the data to do so better.

I have seen that three times over, not without hiccups, but three times over. The problem is we stili live in this world in this country, and it's got a say in Washington, you know, he who pays for healthcare IT is not he who gains.

So long as we have an asymmetric risk and reward function for investing in health IT, where the bulk of the benefit, 89 percent in our calculations goes to the payor among others, you know, when the physicians are footing the bill, it is just not appropriate.

The physician is not motivated to optimize his use of health IT. The payors are motivated to optimize the physician's use of health IT in a strange dynamic that the payor and the physician live in.

I think we have to think about, you know, fundamentally health reform and health IT in the same breath in this country, because in 10 years we could be in the same place. We could have cooler tools and maybe smarter tools, but if we are not incented correctly, and if we are not worried about value instead of volume, physicians may still say who cares.

DR. STEINBERG: Right, and I think that is what is going to happen. I mean there is going to be -- it is happening now, but it is going to happen at increasing speed in the not too distant future.

Physicians and payors are going to come together as one. Right now it is an antagonistic relationship for a lot of reasons that we are all familiar with. It is going to become a symbiotic relationship, and that is what is going to drive a lot of this.

DR. WEITZMAN: How do we motivate the patient to drive what all of you want to do? I mean right now the financial incentives are being allocated 4 1/2 billion I read yesterday, has gone to the payments out of ONC's, the doctors, and the hospitals so far on adoption of VHRs, but I don't think that the patients know or understand the benefits of EHRs and everything we are trying to do here, in every conference I go to here, and every other organization I go to. I still don't see the bottom-up push where the patients are shown look, I solved your problem because, and then he goes to all his docs, since I got four of them, it's 68, and I have got everything, a quadruple bypass and diabetes and everything because I was a kid who ate candy all the time.

The thing we got to do is get the patient to say to the doc that doesn't have the EHR system, I am going to another doc, because I want to get the best available medicine, and we haven't made the patient the driver, and I come out of 10 years of working for the advertising industry, and my law firm was general counsel to the advertising federations and Association of National Advertisers, and we know how to make consumers want things in the advertising industry, and ONC has got to make the case.

I hate just to point to you, but ONC, we have got to make the case to the patients that they are getting the benefit.

DR. REIDER: We actually do have a marketing program that is outreach to patients. Now, is that our greatest lever? You saw me put the slides up, right? So, we have a couple of levers, and our levers are our regulations. Do we have efforts? We have a woman that is our consumerista. She focuses on our consumer program, so it exists.

DR. WEITZMAN: You don't have the budget of Kellogg, you don't have the budget of General Foods.

DR. REIDER: Of course, we don't, right, nor should we, because that is not our greatest lever, but do we have some effort there? Yes, we do.

DR. MIDDLETON: The other observation to make, though, is that if you ask -- and surveys have done so -- you ask patients, you know, does your doctor have an EMR, is your record electronic, and patients typically say yes, it is, and they are surprised when they hear about the slow and gradual penetration of EMR. Point No. 1.

Point No. 2, you know, I think the consumer accountability thing, or the personal accountability thing is going to play itself out. It is different in this country than lots of countries around the world, and others can attest to this perhaps.

You know, our own sense of entitlement and all that kind of stuff, in this country, is part of the healthcare reform challenge, but the consumers are voting with their feet in their pockets, with their feet, a third of Americans see an alternative care provider every year.

With their pocket, you know, there is 20,000 now or more, I-phone medically oriented APPs, you know, that people are downloading and using, so something is happening. In some ways, I have heard one person say, you know, we have had the Arab spring, and the consumer movements, and everything like that. Some have suggested perhaps we are in this medical spring. It is actually the springtime of a conversion or transition to a new model of care that is not volume based, it is value based, it is not physician oriented, it is consumer oriented, et cetera.

DR. REIDER: I guess I would ask maybe Greg if you folks and your colleagues are using the data that we are actually making available about providers who are using EHRs, because right now I can to your web site, I can probably go to your web site and find myself, and find certain factors about myself, and what kind of care I provide, which hospitals I have been to, and perhaps whether I use an electronic health record and maybe in a decade it's an electronic health record with clinical decision support, and maybe it's with clinical decision support from Partners or Mayo, so there might be certain attributes of the kind of practice that are actually publicly available data that ONC and CMS can make available that payors or Ladies Home Journal or the New York Times would make available to folks to help in their decisionmaking about which provider they have.

DR. WEITZMAN: I just note that we have one press person in this audience sitting next to me, and that is one of the things that I have found that we sometimes are missing at lots of these meetings here at the Library, I must have attended about 10 in the last year since the first discussion of Watson.

We need to get some press people at these meetings to translate what we say in more difficult jargon for this audience.

DR. STEINBERG: Yes, I would agree with that.

To your point, I am not aware of us having access to whether or not they use electronic health records with clearly the quality measurement relative to physician movement is alive and well, or maybe not well, but it is alive, it needs to get better because it is not accurate, it is not well, and that is I think where a lot of what we talked about here today could and should make all of that better.

To get back to what is ultimately going to drive all of this, the change in behavior, whether it is the physicians or the patients, it is money. It is money. It is not complicated.

The minute you start paying doctors more, as I said, for doing better, and give them the right tools to really measure that in a way that they believe in and agree to, the behavior will change.

The minute you make patients pay more out of their pocket for unhealthy behaviors, their behavior changes. It is not complicated.

DR. SONNENBERG: I just wanted to comment that a lot of the ARRA meaningful use requirements are focused directly on patients, for example, the requirement to provide a visit summary at the end of the visit, that was one of the most difficult things for us to implement in our practice, physicians really resisted it initially, but it is one of the things the patients appreciate the most, and we have gotten a lot of very positive feedback about the value of walking away with a summary of what was discussed and an accurate medication list.

We have had a number of patients who have switched their care to our practice specifically because we have the electronic medical record, and they like the fact that our providers in different specialties communicate with each other and share information.

DR. WHITE: You know, that's the point. This patient, do I really care if my doctor has an EHR? Well, I do, but that is because I am the Director of Health at AHRQ.

I care that my doctors talk to one another, I care that my chart is not missing when I get there, I care that, you know, my information is captured and that my prescription gets to where it needs to, and by the way, I also happen to care that it is Tier 1's, that have

Tier 3 medicine, that always irritates the hell out of me, because maybe that information wasn't available at the time.

I mean do I care they have an EHR? No. Do I care that they provide better care, that I can e-mail my doctor and that I can get a timely answer to my questions, I care about that absolutely.

DR. PAI: James, you have got a question?

DR. LUO: I want to get back to research again.

Part of this goal of this workshop meeting is to ask where we are and where we want to move to, so I would like to hear panelists' comment on what are the opportunities for the future, and what are the new research directions can advance this field, and how to make the impact on the healthcare.

DR. WHITE: I will offer a brief recap of what I described. I think that understanding information needs is key, okay, whether it is again, you know, whether it is the patient or whether it's the clinician, or whether it's somebody else, understanding what information they need, okay, is important.

Then, they are saying where you get that information. If I have got a certain set of data, that is my streetlight, okay, any information that is not there, where do I have to go to get it.

So, I think that is a fundamental need.

I think that in terms of driving the evidence, okay, that shows that these things make a difference, I t think more research that actually investigates the link to outcomes, which is complicated. There is a reason why we don't have a ton of studies of related outcomes is because processes are easy to measure, and the confounding factors that lead down to outcomes are great. Thank God for smart researchers.

Finally, this is more of a practical question, but pulling together those who create the medical knowledge, okay, and those who try to translate that, so it supports care, is it whole right now? To get to 10 years from now, that has to happen now.

DR. MIDDLETON: I love Jon's list, and I will be submitting a proposal before the end of the day.

I think you have heard a lot of the key pressing issues through the course of the day, today and yesterday, I mean several of us have chatted about this knowledge representation problem, still many different ways to approach that, and perhaps different ways necessary for different approaches to inference, but we need to arrive at a stable knowledge representation, so we can begin to build the corpus of a knowledge base that will be then suitable to broad-based inference and whatnot, and then think about the sharing problem just as I alluded to in my little diagram. You know, it may be one thing to use this knowledge base in EPIC, and another thing to use it in Siemens, another thing to use it in Watson, but they all should be potentially uses.

The CDS inferential problem, right now we are taking such baby steps with our rule-based systems, you know, this is really, as we heard people discuss, situation action rules of the most mundane simple order, we need to think about how to incorporate patient preferences and utilities, what is a utility model for a patient, what do they really care about with respect, what do they really care about with respect to their genetic testing or inference around disease and treatment.

The same might apply to physicians, oh, by the way, it is not just one preference model or another, they really both have to be considered.

I think the other things we are doing now is it is stateless reasoning, it is really just a cross-sectional snapshot, situation action rules or production rules apply to today's chart as opposed to considering the patient's trajectory through health and disease, how do we actually do much more stateful inference to really think about a patient's long-term trajectory.

Workflow insertion points, Bob Greenes and others have done a lot of thinking about situational factors. Decision support has to be provided at the moment of opportunity, the so-called teachable moment for the docs and educators, you know, but at what point does the doc really have the light bulb go off, that, gosh, I need to think about something else, I may need to go there, or the patient for that matter, too, what is the cognitive model that suggests there is a point of uncertainty at which decision support can really be applied and really be useful, and where does that occur in the workflow, because it is different at different times, pre-visit, during the visit post-visit, in the middle of visits, you know, in the intricacies of care.

The data package problem, if I send to Watson, you know, a chart, what form does it have to go in, as a CCD, or a green CDA, or CCD-plus, or a VMR, or whatever, you know, what is the model of that package that has to be shipped to an inference engine, so it can be inferred upon, and some interesting result come back, and what is the nature of the result, you know, what is the recommendation, are there ways to standardize that, what is the explanation, who are the actors and targets for the intervention, et cetera.

Then, just lastly, I think we really, you know, As the National Research Council report that Bill Sted and others worked on, you know, it suggested that we have really this transactional approach to our HIT now, and it is not sensitive at all to the cognitive models that physicians actually have.

I don't reason about a hemoglobin A1c result. I use the hemoglobin A1c result to reason about diabetes. We don't really design our systems to take advantage of immediate pathophysiologic state representation and allow clinicians to do second order analysis of the patient information to actually reason much more efficiently and effectively.

You know, that is a simple example, there are many more examples, I don't understand well, but the cognitive models of what the physicians is trying to do is not well understood. That feeds right into the physician information needs, are certain cognitive models or thinking patterns associated with different information needs, et cetera.

DR. CORN: I would like to follow up a little bit on your answers in James' question. Would you say on the whole, then, the research questions are more and behavior cultural, societal issues than in technical?

DR. MIDDLETON: I wouldn't. I think it is distributed both, maybe if you forced me to guess, maybe a third technical issues and a third cultural, and a third sort of knowledge, modeling, and engineering. Maybe that is part of technical, so I think it is at least half and half.

DR. PAI: Any other comments?

DR. WHITE: In addition to the behavioral, psychological, cognitive modeling that we were also talking about, there is a systems modeling component of this. We are breaking it down by the individual user.

We are not taking the opportunities that Jacob alluded to in his talk when he said why the hell am I getting a mammogram reminder, right, was what you said?

We have actually done research on this. We did a demonstration in a number of primary care practices where it was called standing orders, where basically, any member of the staff that had clinical -- if there is an individual that needed prevented services, any member of the staff, it would pop up in front of him, that said can I get that schedule for you now, because it didn't have to be Dr. Reider or Dr. White to be able to do that.

So, there is a systems modeling component of this. You and I were just discussing earlier that we have been calling NSF, they have that behavioral component, they have the computational component. They also have a systems modeling component that we have been working with closely, and we are the healthcare people.

I think all of those get more deeply at the issues that are down there.

DR. REIDER: I will say that ONC is interested in what works, because on some level, we went to say to AHRQ and perhaps also NIH, these are the long-term objectives, right, these are the kinds of things that we expect the market to need, back to our market conversation, and therefore, figure out for us, researchers, what is going to work, so that we can then implement those things as perhaps the standards, right, these are the things that may accelerate the implementation of the stuff that works.

So, both from a standard perspective, from a technical view, and also from behavior, right? So, these are the kinds of things that we would motivate vendors to do in our regulations, and the technical standards that we want to require, so that things go faster.

DR. WHITE: When I talk about AHRQ, and

what AHRQ does, and I contrast it to NIH and NIH does and CDC does, I talk about AHRQ in terms of health services research, and we do research about health services.

If you look across all the different components of NIH, they both fund a lot of health services research, right, just like we fund diabetes research, but it is in the context of health services.

But I think the different institutes can also get more deeply into the particular issues that that institute is there for, right?

The informatics components of those issues, they have deep knowledge about that domain, so let's meet at the interface between and betwixt I think that different research agencies have different roles to play, NSF has a different role to play, CDC has a public health role to play, that are all related to one another, but they definitely have their own twists to it that meet the needs of the constituencies that our different agencies serve.

DR. PAI: Any other questions?

I had one question about the clinical decision support, somebody had raised the question about how do you rate different CD systems, and I was wondering, does it make sense to have a centralized database for academic researchers, where people can test their systems against multiple kinds of structured notes and have it open for everybody to use. I mean industry has it, but something like Watson will benefit from having a bigger database to work with, too.

DR. WHITE: Is the gentleman from NIST here? Do you want to talk at all about TREC, do you know about TREC?

MR. SOBOROFF: In fact, I do.

DR. WHITE: I manage the group that runs TREC. TREC, for those of you who don't know, is an evaluation workshop series for information retrieval, has some siblings in natural language processing also, where we make data available and structure user-focused tasks around that data.

So, for example, question and answer came out of TREC, and eventually gave birth to Watson. IBM spent all the money for the Watson parse. I don't want to take too much credit, but the idea of the technology, how you would measure it, how you would actually write that technology came out of TREC.

The challenge that researchers have in this domain right now is that there is a colossal amount of data and no one can get to it, and if I have two people who got to some data, they can't actually talk about what each other did.

There are really good legal and privacy and IRB reasons for this, which I don't want to for a moment imply that computer scientists like myself think are a barrier to progress, but the challenge is if we can solve this, what you call secondary use scenario, you will change the state of the art in 5 to 10 years completely.

Every single problem I heard talked about today done, but -- but you have to solve the data problem, and the data problem solution is not something that people at the research level can solve. It is not people at the Mayo level can solve it. It is policy level people to stand up and say, well, if we are going to push the state of the art in how computers support clinical decisions, for example, part of that might be NLP, part of that might be databases, part of it -- all kinds of stuff.

If we going to push the state of the art, it happens, we really want to make a jump, we are going to create this phenomenon where people, more than one set of eyes can look at the data at once and then people can compare results between, and you can actually measure progress.

DR. WHITE: I apologize for not calling it TREC. This is what I get for calling on e-mail, but not actually coming in person.

DR. GREENES: When you said "solve the data problem," I would just like to understand a little bit more about what you mean. Are you talking about being able to amass aggregate data, avoiding and addressing the privacy issues, or are you talking about solving the structure?

MR. SOBOROFF: I am talking about solving the data access problem. So, for example, in TREC, we have a medical records TREC, so we have a task around cohort finding, and reducing some de-identified medical records.

Now, apparently it was okay to have de-identified data out for research as long as not too many people knew about it, but if too many people know about it, that's bad, even if it has been de-identified. That is the problem that needs solving.

DR. CORN: Forgive me for a moment. That's true, the TREC people have had a hard job, and they have been doing it. I called a number of people that they were having trouble getting the data from, and there is certainly truth to that point of view, but a number of them were, in all fairness, a little concerned about the large number of people to whom the data would have been distributed for purposes of the TREC thing, and they told me at least that in terms of their lawyers and their IRBs, they were able to make agreements with two, perhaps three organizations whose structure and whose people they knew well and had confidence in.

It is not lack of confidence in TREC, but the fact that such a large number of relatively unknown groups would be looking at it. So, it is not venal, it is perhaps unnecessary cowardice.

DR. SOBOROFF: I think that the central part of the problem is everybody is trying to solve this problem at the little person level, so every little university with a medical school, is doing medical informatics and trying to solve this problem independently of everybody else.

Everybody's IRB is asking this problem on their own, everybody, but certainly those natural language processing information people are confronting it, because we are just learning what IRB stands for.

If we have to keep solving the same problem, then, we can't get it solved.

DR. METEER: I would like to speak on the same point. I am from Brandeis University. If you look at the history of speech recognition, and named entity extraction, and all of these different natural language processing, you will see that every time data was collected, released, and let everybody in the community work across that data, and then evaluate them, you see movement in performance at every time, and it is well plotted, and unfortunately, these are fields that need a lot of data, and I have been in the field for many years, and we really just kind of played around for about the first decade of my work in this area, until we became data driven, and evaluation driven.

I know if you just pick a couple places and say, well, we are going to let them work on it, you are not going to move the field forward. We need to figure out how to get it, so that we can work comparatively, and then we get together.

I mean this is how Darpa solved it, well, didn't solve speech, but moved speech to the point by saying okay, you all are going to work on the Wall Street Journal, and then you are required to come to this meeting and say what you did, that you got 0.4 percent improvement because you did adaptation, and everyone had to share, and we actually got speech to the point where the apple marketing machine made everybody want it, but we were able to do that.

So, we need that evaluation. We also need to figure out how to get our components into pipelines to have an extrinsic evaluation, just as, you know, we want to evaluate performance, but then what do we do, how do we get maybe the software to use, so we can take a piece of that problem and say, well, look, I know it worked to this degree, is it really useful in that context, so those I think are the two blocks that the natural language processing, particularly the University community is facing.

MR. MARCUS: This is a bit of a tangent and this may have been discussed earlier, so forgive me, but I need to put a plug in for the behavioral and social sciences.

If you are conceptualizing more information, better medical decisions, and that is the basis, sometimes it is thought of as a rational sort of way to think about things, if that is the conceptualization, and I have seen more computer scientists and more bioinformaticians, et cetera, think that if you just have better information, you have better decisions, I am here to tell you that there is a whole field of knowledge that will tell you that humans act on more than just rational based knowledge, and as you build expert systems, I would encourage you to contact your local behavioral scientist.

[Laughter.]

DR. WEITZMAN: To address the previous question, I commend to all of you the Caffman Foundation Report of their task force that came out last week. It is Caffman with two f's and 1 n. In there, there is a set of recommendations including four lawyers, two from Duke, one from Yale, and one from somewhere else, who are criticizing HIPAA as an impediment to the kind of research that we all would like to do with large data sets and at a meeting back in January, Senator Daschle made one comment which was beautiful. "We have got silos and stovepipes, and no cooperation."

DR. REIDER: I would ask about patients. Raise your hand if you would donate your data to such a data set, of you had control of it.

My mom said the same thing, and we can, you know, brainstorm about what might be the policy levers or options or whatever, but I would say there are certainly a lot of enthusiasm for patients starting to take control here, and if patients could take control and then donate it all to NIST or Brandeis or wherever, that actually might be a viable option.

DR. SOBOROFF: Is anybody here who is not a physician, but who has spent a considerable amount of time in a hospital would certainly agree, being the patient makes you realize that you don't control any of that data, and that is scary for anybody, so I agree let's solve that problem.

At the same time, the people who are working on the data issues, the secondary use issues, are removed from -- they are not trying to find out privacy revealing information, and I understand that this is a very difficult line to understand or to draw, but let's think outside the box, because the HIPAA privacy, the HIPAA de-identification stuff doesn't work, and because nobody believes they have safety behind that number.

DR. PAI: Blackford, do you want to make closing comments?

DR. RESNIK: Actually, I was just going to just put the cherry on top for what Ian had to say. Thank you for that, and, Marie, thank you for reinforcing that. I just wanted to repeat what Ian said at the beginning, which is this is not a problem that is going to be solved by the researchers. This is a problem where we need the people who are several levels above us to be addressing this, and it seems to me that this room is as good a place as any to start, so it would be great.

Let us know if there is something we can do to help, but otherwise, we are just speaking into the wind.

DR. MIDDLETON: On that provocative comment, I am going to take 60 seconds to close. Thanks to all of you for staying to the very end. Thank to the sponsors of the conference, NIBIB, NLM, James Luo, Victor Venaypi, Milt Corn, and Don Lindberg, and to the incredible array of speakers yesterday and today, Phil, your talk was extremely interesting, nice to make your acquaintance, Chris Manning, Carol Friedman, of course, George Hripcsak, others I am going to forget today.

Thanks to all who have traveled long and far, and I won't name names, but some are gone already.

I suggest we meet again, maybe about two or three years out, but it seems it needs to be the CDS community, the NLP community, and the behavioral scientists, thank you for the very important and provocative comment, but to orient all this together, we really do need to think about those behavioral dimensions and cognitive issues that actually orient us to all of our decisionmaking and perception.

Thanks to all of you and see you then the next time.

[Whereupon, at 5:01 p.m., the meeting was adjourned.]