PETER SZOLOVITS: So today's topic is workflow, and this is something that-- a topic that I didn't realize existed when I started working in this area, but I've had my nose ground and ground into it for many decades. And so finally, it has become obvious to me that it's something to pay attention to. So here's an interesting question.

Suppose that your goal in the kind of work that we're doing in this class is to improve medical care-- not an unreasonable goal. So how do you do it? Well, we had an idea back in the 1970s when I was getting started on this, which was that we wanted to understand what the world's best experts did and to create decision support systems by encapsulating their knowledge about how to do diagnosis, how to do prognosis and treatment selection, in order to improve the performance of every other doctor who was not a world class expert by allowing the world class expertise captured in a computer system to help people figure out how to do better-- so to make them more accurate diagnosticians, more efficient therapists, et cetera.

And the goal here was really to bring up the average performance of everybody in the health care system. So we used to say things like, bring everybody practicing medicine closer to the level of practice of the world class experts. Now, that turned out not to be what was important. And so there was another idea that came along a little bit later that said, well, it's not really so much the average performance of doctors that's bad. It's the subaverage performance that's really terrible.

And so if you're subaverage performance leads to your patients dying, but your above average performance only makes a moderate difference in their outcomes, then it's clearly more important to focus on the people who are the worst doctors and to get them to act in a better way. And thus, was born the idea of a protocol that says, let's treat similar patients in similar ways. And the value of that is to reduce the variance-- so improve average versus reduce variance.

So which of these is better? Well it depends on your loss function. So as I was suggesting, if your loss function is a symmetric so that doing badly or doing below average is much worse than doing above average is much better, than this protocol idea of reducing variance is really important.

And this is pretty much what the medical system has adopted. So I wanted to try to help you visualize this. Suppose that on some arbitrary scale of 0 to 8, we have an usual, normal distribution, of on the left the base behaviors-- so this is how people, on average, normally behave-- we assume that there's something like a normal distribution.

So here is a world class expert whose performance is up at 6 or 7 and here's the dud of a doctor whose performance is down between 0 and 1. And the average doctor is just shy of 4. So here are two scenarios.

Scenario one is that we improve these guys performance by just a little bit. So we improve it by 0.1 performance points, I think is what I've done in this model. versus another approach, which is suppose we could cut down the variance dramatically so that this same normal distribution becomes narrower. Its average is still in exactly the same place, but now there are no distant outliers. So there aren't doctors who perform a lot worse, and there aren't doctors who perform a lot better either.

Well, what happens in that case? Well, you have to look at the cost function. So if you have a cost function like this that says, that somebody's performing at the 0 level has a cost of 1. Whereas somebody performing at the 8 level has a cost of almost 0, and it's exponentially declining like this, so that the average performance has a much lower cost than the average between the worst performance and the best performance.

So this suggests that, if you could bunch people into this region of performance, that your overall costs would go down. And, in fact-- this is a purely hypothetical model that I've built-- but if you do the calculations, you discover that for the base distribution, here is the distribution of costs. For the slightly improved distribution, you get a cost, which is 1,694 versus 781, again, in arbitrary units. But if you manage to narrow the distribution, you can get the total cost down to less than what you do by improving the average.

Now, this is not a proof, but this is the right idea. The proof is probably in the fact that medical systems have adopted this, and have decided that getting all doctors to behave more like the average doctor is the best practical way of improving medical care. Well, how do we narrow the performance distribution?

So one way is by having guidelines and protocols where you have some learned body who prescribes appropriate methods to diagnose and treat patients. So what happens is, for example, the article here from November of 2018, a report of the American College of Cardiology, the American Heart Association Task Force on Clinical Practice Guidelines, and this has been adopted by this cornucopia of three and four letter abbreviated organizations. And it's a guideline on the management of blood cholesterol. So as you know, having high cholesterol is dangerous. It can lead to heart attacks and strokes, and so there is a consensus that it would be good to lower that in people.

So these guys went about this by gathering together a bunch of world experts and saying, well, how do we do this? What do we promulgate as the appropriate way to care for patients with this condition? And the first thing they did is they came up with a color coded notion of how strong the recommendation a certain recommendation should be And another color coded or shaded level of certainty in that recommendation.

So, for example, if you say something is in class 1, so it's a strong recommendation, then you use words like is recommended, or is indicated, useful, effective, beneficial, should be performed, et cetera. If it's in class 2, where the benefit is much greater than the risk, then you say things like it's reasonable, it can be useful, et cetera. If the benefit is maybe equal to or a little bit better than the risk, you say waffle words, like might be reasonable, may be considered.

If there is no benefit, in other words, if it roughly equals the risk, then you say, it's not recommended. And if the risk is greater than the benefit, then you say things like it's potentially harmful, causes harm, et cetera. So if you were giving a recommendation on whether to spray disinfectant down your lungs, you might put that in red and say, this is not recommended.

And then here, this shading coding is basically how good is the evidence for this recommendation. So the best evidence, the level A, is high-quality evidence from multiple randomized controlled clinical trials, or a meta-analyses of a high-quality RCTs, or RCTs corroborated by high-quality registry studies. And then we go down to level C, which is consensus of expert opinion based on clinical experience, but

without any sort of formal analysis.

So if you look at this particular document on cholesterol it says, well, here are the recommendations on the measurement of LDL and non-HDL cholesterol. And they say here, the confidence and the recommendation is one, and it's based on B and our level of evidence. And it says, in adults who are 20 years or older and not on lipid-lowering therapy, measurements of either a fasting or a non-fasting blood-- dot, dot, dot. So you could read this in the notes later.

But notice that there are high force recommendations. There are lower force recommendations, and each recommendation is also shading coded to tell you what the strength of evidence is for this kind of recommendation. Here's just another example.

This is secondary atherosclerotic cardiovascular disease prevention. So this is for somebody who's already ill, and it's a bunch of recommendations. If you're over 75 years of age, or younger with a clinical case of coronary vascular disease, then high intensity statin therapy should be initiated or continued with the aim of achieving a 50% or greater reduction in LDLC and et cetera. So again, a whole bunch of different recommendations. Once again, the strength of the recommendation-- by the way, this is just the first page of a couple of pages-- and the quality of evidence for it.

So this is very much the way that learned societies are now trying to influence the practice of medicine in order to reduce the variance and get everybody to behave in a normal way. You've probably seen articles about Atul Gawande, who's a surgeon here in Boston, and he's gotten publicly famous for advocating checklists. And he says, for example, if you're a surgeon, you should act like an airline pilot, that before you take off in the airplane, you go through a sanity checklist to make sure that all the systems are working properly, that all the switches are set correctly, which in a surgical setting would be things like you have all the right necessary equipment available, that you know what to do in various potential emergencies, et cetera.

So here are their take-home messages, which makes sense. Here, I've abstracted these from the paper that has all of these details. So number one, you go, well, duh-

- in all individuals, emphasize a heart healthy lifestyle across the life course. That seems not terribly controversial, and in people who are already diseased, reduce low-density lipoprotein with high-intensity therapy by statins. And in very high risk ASCVD, use a threshold of 70 milligrams per deciliter, et cetera.

So these are the summary recommendations. And the hope is that doctors reading these sorts of articles come away from them convinced and will remember that they're supposed to act this way when they're interacting with their patients. This is a flow chart, again, abstracted from that paper by them which says, everybody, you should emphasize a healthy lifestyle. And then depending on your age, depending on what your estimate of lifetime risk is, you wind up in different categories. And these different categories have different recommendations for what you ought to do with your patients.

This is for secondary prevention. So it's a similar flow chart for people who are already diseased and not just at risk. And then for people at very high risk for future events, which is defined by these histories and these high-risk conditions, these are the people who fall into that second flow chart and should be treated that way.

Now, by the way, I didn't make a poll, so I'll give you the answer. But it's interesting to ask. So when papers like this get published, how well do doctors actually adhere to these? And the answer turns out to be not very well, and it takes many, many years before these kinds of recommendations are taken up by the majority of the community, so even very, very uncontroversial recommendations.

For example, I think 20 years ago there was a recommendation that said that anybody who's had a heart attack should be treated, even if they're now asymptomatic, with beta blockers. Because in various trials, they showed that there was a 35% reduction in repeat heart attacks as a result of this treatment. It took, I think, over a dozen years before most doctors were aware of this and started making that kind of recommendation to their patients.

There's something called the AHRQ, the Agency for Health Research and Quality. And until the current administration, they ran a national guideline clearinghouse that contained myriad of these guidelines, published by different authorities, and was available for people to download and use. There's been an attempt by

Guideline Central to take over some of these roles since the government shutdown the government run one, and they have about 2,000 guidelines that are posted on their site.

And these are some of the examples. So risk reduction of prostate cancer with drugs or nutritional supplements, stem cell transplantation in multiple myeloma, stem cell transplantation in myelodysplastic syndromes and acute myeloid leukemia, et cetera. And then they also publish a bunch of risk calculators that say-- I don't know what the 4T score is for heparin-induced thrombocytopenia-- but there are tons of these as well. So there's a clearinghouse of these things. And you, as a practicing doctor, can go to these. Or your hospital can decide that they're going to provide these guidelines to their doctors, and either encourage, or in some cases, coerce them to use the guidelines in order to determine what their activity is.

Now, notice that this is a very top-down kind of activity. So it's typically done by these learned societies that bring together experts to cogitate on what the right thing to do is, and then they tell the rest of the world how to do it. But there's also a kind of bottom-up activity.

So there is something called a "care plan." Now, a care plan is really a nursing term. So if you hang out at a hospital, the thing you discover is that the doctors are evanescent. They appear and disappear. They're like elementary particles, and they're not around all the time.

The people who are actually taking care of you are the nurses. And so the nurses have developed a set of methodologies for how to ensure that they take good care of you, and one of them is the development of these care plans. And then what clinical pathways are is an attempt to take the care plans that nurses use in taking care of individuals and to generalize from those and say, well, what are the typical ways in which we take care of patients in a particular cohort?

So I'm going to talk a little bit about that, and one of the papers I gave you as an optional reading for today is about cow paths, which are these attempts to build generalizations of care plans. So this is a care plan from the Michigan Center for Nursing, which is an educational organization that tries to help nurses figure out how to be good nurses. I was very amused when I was looking for this.

I ran across a video, which is some experienced nurse talking about how you build these care plans. And she sort of says, well, when you're in nursing school, you learn how to build these very elaborate carefully constructed care plans. When you're actually practicing as a nurse, you'll never have time to do this.

And so you're going to do a rough approximation to this. And don't worry about it. But for now, satisfy your professors by doing these exercises correctly.

So take a look at this. So there are a bunch of columns. The leftmost one says assessment. So this is objective, subjective, and medical diagnostic data.

So the objective data is this patient has gangrene-infected left foot-- not a good thing, an open wound, et cetera, et cetera. Subjective data, the patient said the pain is worse when walking and turning. She dreads physical therapy, and she wishes she did not have to be in this situation-- surprise. But that's definitely subjective. You can't see external evidence of that.

The nursing diagnosis is that this patient has impaired tissue integrity in reference to the wound and the presence of an infection. Now, that diagnosis actually comes with a kind of guideline about how to make that diagnosis. In other words, in order to be able to put that down on the care plan, she has to make sure that characteristics of the patient satisfy certain criteria which are the definition of that diagnosis.

The patient outcomes-- so this is the goals that the nurse is trying to achieve. And notice, there are five goals here. One is that the patient will report any altered sensation of pain at the tissue impairment between January 23 and 24. So this is a very specific goal. It says, the patient will tell me that they feel better, that there's a change in their feeling in their infected left foot.

They will understand the plan to heal tissue and prevent injury. So there's a patient education component. They will describe measures to protect and heal the tissue, including wound care by 124. So notice, this is the patient describing to you what you are planning to do for them, in other words, demonstrating an understanding of what the plan is and what's likely to happen with them.

Experience a wound decrease that decreases in size and has increased granulation

tissue, and achieve functional pain goal of 0 by 124 per the patient's verbalization. So when they come in and they ask you on that pain scale, are you at a 0, or a 10, or somewhere in between, the goal is that the patient will say, I'm at a 0, in other words, no pain. Now, what are the interventions? Well, these are the things that the nurse plans to do in order to try to achieve those goals. And then the rationale is an explanation of why it's reasonable to expect those interventions to achieve those goals.

And the evaluation of outcomes says, what criteria or what are the actual outcomes for what we're trying to achieve? So that gets filled in later, obviously, then when the plan is made. So if you look at a website like this, there are templated care plans for many, many conditions.

You can see that I'm only up to C in an A to Z listing from this one website, and there are plenty of others. But there is an admission care plan, adult failure to thrive, alcohol withdrawal, runny nose, altered cardiac output, amputation. I don't know what an anasarca is-- anemia, angina, anticoagulant care, et cetera. So there are tons of different conditions that different patients fall into, and this is a way of trying to list the template care plans.

Now, this paper is kind of interesting, by Yiye Zhang and colleagues. And what they did is they said, well, let's take all these care plans and let's try to build a machine learning system that learns what are the typical patterns that are embedded in those care plans. But they didn't start with the plans. This is retrospective analysis.

So what they started with is the actual records of what was done to each patient. And so the idea is that you get treatment data from the electronic health record. Then you identify patient subgroups from that data, and then you mine for common treatment patterns. And you have medical experts evaluate these, and these then become clinical pathways, which are this generalization of the care plans to particular subpopulations of patients.

So the idea is that they define a bunch of abstractions. So they say, look, an event is a visit. So, for example, for an outpatient, anything that happens to you during one visit to a doctor or to a hospital.

So it's a set of procedures, a set of medications, a set of diagnoses. And by the way,

they were focusing on people with kidney disease as the target population that they were looking at. So then they say, OK, individual events are going to be abstracted into these supernodes, which capture a unique combination of associations of events associated with some visit.

So you might worry that this is going to be combinatorial, because there are many possible combinations of things. And that is, in fact, a bit of a problem, I think, in their analysis. So now, you have these supernodes, and then each patient has a visit sequence, which is a time-ordered list of the supernodes.

So every time you go see your doctor, you have one new supernode. And so you have a time series of these. And then they do the following thing.

They say, gee, when we talk to our doctors and nurses, they tell us that they care mostly about what happened at the last visit that the patient had. But they also care a little bit less, but they still care about what happened at the visit previous to that, but not so much about history going further back. And so they say, well, in a Markov chain, we only have things depend on the last node in the Markov chain. So let's change the model here so that we will combine pairs of visits into nodes so that each node in the Markov chain will represent the last two visits that the patient had.

So this could, again, cause some combinatorial problems. But here's the image that they come up with. So there are individual items.

Is it a hospital visit, an office visit, a visit for the purpose of education? Are you in chronic kidney disease stage four? Was an ultrasound done?

Were you given ACE inhibitors? Were you given diuretics, et cetera? So these are all the data that we mentioned.

They treat that as a bag. And then they say, OK, we're going to identify all the bags that have the same exact content. An asterisk, they didn't look, for example, at the dose of medication that you were given, only which medication it was. So there are some collapsing that way.

Then the supernodes are these combinations where we say, OK, you had a particular purpose, a particular diagnosis, a particular set of interventions, a

particular set of procedures. And again, we list all possible combinations of those, and then that sequence represents your sequence. These are aggregated into supernodes. That represents your visit sequence, and then these super pairs are this hack to let you look two steps back in the Markov chain.

And so they wind up with about 3,500 different of these super pair nodes. So it is combinatorial, but it's not terribly combinatorial in their data. They then compute the maximum of the length of common subsequences between each pair of visit sequences. So they're going to cluster these sequences.

They define a distance function that says that the more they share a common sequence, the less distant they are from each other. And the particular distance function they used is the length of each sequence minus twice the length of the common subsequence, the longest common subsequence, which seems pretty reasonable. And then hierarchical clustering into distinct subgroups, they came up with 31 groups for this group of patients, and here they are.

And what you see is that some of them don't differ a whole lot from each other. So, for example, these two differ only in that the patient got some medication and diuretics in one case and just that medication in the other case. So these are-- it is a hierarchical cluster, and the things lower down in the clustering are probably fairly close to each other.

Nevertheless, what they're able to do, then, is to estimate a transition matrix among these supernode pair states, and they can look at different trajectories depending on the degree of support for the data. So you can set different thresholds on how many cases have to be in a particular state in order for you to take transitions to or from that state seriously. One of the critiques I would make of the study is that they had way too little data, and so many of the groups that they came up with had relatively small numbers of patients in them, which is unfortunate.

Now, once you have these transition matrices, then you can say, OK, for cluster 29, which was this cluster, so there were a grand total of 14 patients in this cluster. They were all at chronic kidney disease stage 4, so quite severe. They were all hypertensive. They were all on ACE inhibitors and statins, and everybody in that group had that categorization.

So if you look there then you can say, OK, for all the things we know about that patient, what are the probabilistic relationships between them? And what we find is that-- man, I can't read these. So these nodes imply other nodes, and the strength of the arrows is proportional to their width. And so this is a representation of everything that we've learned about that cluster, but remember, only from those 14 patients. So I'm not sure I would take this to the bank and rely on it too intensely.

But they then, by hand, abstract it and say, well, let's look at an interpretation of this. And so if they look in typical patterns that they see in that cluster, they say, hmm, we see an office visit in which the patient is on these medications and has these procedures. Then they're hospitalized. Then there's another-- let's see. No, I'm sorry.

Yeah, yellow node is an office visit. So they're hospitalized. They then get an education visit, so that's typically with the nurse or nurse practitioner to explain to them what they ought to be doing. They have another hospital-- they have another office visit.

They have a hospital visit. They have another hospital visit, and then they die. So that, unfortunately, is a not atypical pattern that you see in patients who are at a pretty severe state of chronic kidney disease. And we don't know from this diagram how long this process takes to take place.

So I have some questions. There are a lot of subgroups. Some of them were fairly similar to others.

They have between 10 and 158 patients in each subgroup. So I would feel much better if they had between 1,000 and 15,000 or something patients in each group, or 150,000 patients in each group. I would feel much more believing in the representations that they found.

And the other problem is that even within an individual subgroup, you can find very different patterns. So, for example, here is a pattern where, again, a person has a couple of office visits. They go to the hospital. Or they go to the hospital twice with slightly different-- yes.

So this person at this point is in acute kidney injury. So you can get there either

directly from the office visit or from an earlier hospitalization, and then they die. And so this is part of that pattern.

But here's another pattern mined from exactly the same subgroup. Now, this subgroup has 122 patients in it, so there's a little bit more heterogeneity. But what you see here is that a patient is going back and forth between education visits and doctor's visits, back and forth between doctors visits and hospitalizations, then a hospitalization, then another hospitalization, but they're surviving.

So it's a little bit tricky, but I think this is a good idea, but there are probably improvements that are possible on the technique that's being used here. And, of course, much more data would be very helpful in order to really delineate what's going on in these patients. Here's a similar idea that I was involved.

Jeff Klann did his PhD at Regenstrief, which is a very well-known, very early adopter of computerized information systems in Indiana. And so what he started off-- and he said, hmm. You know the Amazon recommendation system that says you just bought this camera lends, and other people who bought this camera lens also bought a cleaning kit and a battery that goes with that camera, and so on? So he said, why don't we apply that same idea to medical orders?

And so he took the record of all the orders at Regenstrief, and he basically built an approximation to the Amazon recommendation system that said, hey, other doctors who have ordered the following set of tests have also ordered this additional test that you didn't order. Maybe you should consider doing it. Or conversely, other doctors who have ordered this set of tests have never ordered this other one in addition. And so are you sure you really need it? So that was the idea.

And what he did was he focused on four different clinical issues. So one of them was an emergency department visit for back pain, pregnancy, so labor and delivery, hypertension in the urgent visit clinic-- so the urgent visit clinic is one of these lower-level non-emergency department, cheaper, lower level of care, but still urgent care kinds of clinics that many hospitals have established in order to try to keep people who are not that sick out of the emergency department and in this lower-intensity clinic-- and hypertension, and high blood pressure, and then altered mental state in the intensive care unit. So people in the ICU are often medicated,

and they become wacko, and so this is trying to take care of such patients.

They used three years of encountered data from Regenstrief. And for each domain, they limited themselves to the 40 most frequent orders, and, again, low granularity. So, for example, drug, but not the dose of the drug for medications, and the 10 most frequent comorbidities or co-occurring diagnoses.

So this is an example of wisdom of the crowd kind of approach that says, well, what your colleagues do is probably a good representation of what you ought to be doing. Now, what's an obvious pitfall of this approach? I'm just checking to see if you're awake. Yeah?

AUDIENCE: Just reinforce whatever's [INAUDIBLE].

PETER SZOLOVITS: Yeah, if they're all bozos, they're going to train you to be a bozo too. And there's a lot of stuff in medicine that is not very well-supported by evidence, where, in fact, people have developed traditions of doing things a certain way that may not be the right way to do it. And this just reinforces that. On the other hand, it probably does reduce variance in the sense that we talked about at the beginning. And so, as a result, it may be a reasonable approach, if you're willing to tolerate some exceptions.

My favorite story is Semmelweiss figured out that having a baby in a hospital in Vienna was extremely dangerous for the mother, because they would die of what was called "child bed fever," which was basically an infection. And Semmelweiss figured out that maybe there was-- this was before Pasteur. But he figured out that maybe there was something that was being transmitted from one woman to the next that was causing this child bed fever, and, of course, he was right.

And he did an experiment, where on his maternity ward, he had all of the younger doctors wash their hands with some sort of alcohol or something to kill whatever they were transmitting. And their death rate from this child bed fever dropped to almost 0. And he went to his colleagues and he said, hey, guys, we could really make the world a better place and stop killing women. And they looked at him, and they said, you know, these hands heal, they don't kill.

Many of them were upper class or noblemen who had gone into this profession. The

idea that somehow they were responsible for transmitting what turns out to be bacteria was just a non-starter for them. And Semmelweiss wound up ending his days in a mental institution, because he went nuts. He was unable to change practice even though he had done an experiment to demonstrate that it worked. So this is a case where the wisdom of the crowd was not so good and led to bad outcomes.

So like Amazon's recommendation system, it automates the learning of decision support rules. And what's attractive about this is that because it's induced from real data, it tends to deal with more complex cases than the sort of simple, stereotypical cases for which people can develop guidelines, for example, where they can anticipate what's going to happen in various circumstances. So he used the Bayesian networking model that used diagnoses possible orders and evidence, which is the results from orders that were already completed.

There's a system out of University of Pittsburgh, called Tetrad, that implements a nice version of something called Greedy Equivalent Search, which is a faster way of searching through the space of Bayesian networks for an appropriate network that represents your data. So it's a highly combinatorial problem, and the cleverness in this is that it figures out classes of Bayesian networks that, by definition, would fit the data equally well. And it does it by class rather than by individual network, and so it gets a nice combinatorial reduction.

And what Jeff found is, for example, in the pregnancy network, these are the nodes that correspond to various interventions and various conditions. And this is the Bayesian network that best fits that data. It's reasonably complicated.

Here are some others. This is for the emergency department case. So you see that you have things like chest pain and abdominal pain presenting diagnoses, and then you have various procedures, like an abdomen CT, or a pelvic CT, or a chest CT, or a head CT, or a basic metabolic panel, et cetera, and this gives you the probabilistic relationships between them.

And so what they were able to do is to take this Bayesian network representation, and then if you lay a particular patient's data on that representation, that corresponds to fixing the value of certain nodes. And then you do Bayesian

inference to figure out the probabilities of the unobserved nodes, and you recommend the highest probability interventions that have not yet been done. So it's a little bit like, if you remember, we talked about sequential diagnosis. This is a little bit in that spirit, but it's a much more complicated Bayesian network model rather than a naive-based model.

And so the interface looks like this. You have-- it's called the Iterative Treatment Suggestions algorithm, and it shows the doctor that these are the problems of the patient, and the current orders, and the probability that you might ask to have any one of these orders done. And what they're able to show is that this does reasonably well. Obviously, it wouldn't have been published if they hadn't been able to show that.

And so what you see is that, for example, the next order that's done in an inpatient pregnancy using this Bayesian network formalism has a position of about fourth on the list. So their criterion for judging this algorithm is, is it raising the things that people actually do too high on the list of the recommended list, on the recommended set of actions that you consider doing? And you see that it's fourth, on average, in inpatient pregnancy, about sixth in the ICU, about sixth in the emergency department, and about fifth in the urgent care clinic.

So that's pretty good, because that means that even if you're looking at an iPhone, there's enough screen real estate that it'll be on the so-called first page of Google hits, which is the only thing people ever pay attention to. And, in fact, they can show that the average list position corresponds to the order rank by frequency, but that their model does a reasonably good job of keeping you within the first 10 or so for much of this range. I'm going to shift gears again.

So Adam Right, you've met. He was discussant in one of our earlier classes. And Adam's been very active in trying to deploy decision support systems. And he had an interesting episode back in-- when was this-- 2016. So it must have been a little before 2016.

He went to demonstrate this great decision support system that they had implemented at the Brigham, and he put in a fake case where an alert should have gone off for a patient who has been on a particular drug for more than a year and

needs to have their thyroid stimulating hormone measured in order to check for a potential side effect of long-term use of amiodarone, as well as to have their-- ALT is a liver test, liver enzyme test. So they needed both of those tests. He was demonstrating this wonderful system.

He put in a fake patient who had these conditions, and the alert didn't go off. So he goes, hmm, what's going on? And they went back, and they discovered that in 2009 the system's internal code for amiodarone had been changed from 40 to 70-99. Who knows why? But the rule logic in the system was never updated to reflect this change.

And so, in fact, if you look at the history of the use of amiodarone-- by the way, it's an interesting graph. The blue dots are weekdays, and the black dots are weekends. So not a lot goes on in the hospital during the weekend.

But what you see is that-- I don't know what happened before about the end of 2009. They probably weren't running that rule or something. But what you see is sort of a gradual increase in the use of this rule, and then you see a long decrease from 2010 up through 2013 when they discovered this problem.

Now, why a decrease? I mean, it's not a sudden jump to 0. And the reason was that this came about-- first of all, it came about gradually, because the people who had had this drug before that change in the software had gotten the old code, which was still triggering the rule. It's just that as time went on, more and more people who needed the test had gotten the drug with its new code. And with that new code, it was no longer triggering the rule.

And then this is the point at which they discovered the bug, and then they fixed it. Of course, it came right back up again. Oh. Well, I'll talk about some of the others as well.

So this was the amiodarone case. So it fell suddenly, as some patients were taken off the drug and others were started with this new internal code. And as I said, the alert logic was fixed back in 2013. Yeah?

AUDIENCE: So I don't know how hospital IT systems work, and it might vary from place to place. But is there ever a notion of like this computer needs to be updated for the

software, but that one already got updated? Or are they all synced up so that they all get updated at the same time?

PETER SZOLOVITS: They tend to all get updated at the same time. There are disasters that have happened in that updating process. Famously, the Beth Israel was down for about three days. Their computer system just crashed.

And what they discovered is that they had this very complicated network in which there were cyclic dependencies in order to boot up different systems. So some system had to be up in order to let some other system be up, which had to be up in order to let the first system be up. And, of course, in normal operation, they never take down the whole system, and so nobody had discovered this until there was-- Cisco screwed them.

There was some fix in the routers that caused everything to crash, and then they couldn't bring it back up again. And so that was a big panic. John Halamka, who's the CIO there, is a former student of mine. And after this all played out, I asked John, so what's the first thing you did when this happened? And he said, I sent a couple of panel trucks down to the Staples warehouse to buy pads of paper, which is pretty smart.

So here's another example. This is lead screening. And so this was a case where there is a lead screening rule for two-year-olds. There is also one for one-, three-, and four-year-olds. And there was no change in screening for one-, three-, and four-year-olds, but the screening for two-year-olds went from 300 or 400 a day down to 0 for several years before they noticed it, and then went back up to the previous level.

And they never did quite figure out what happened here, but something added two incomplete clauses to the rule having to do with gender and smoking status. But the clauses were incomplete, and so they were actually looking for the case of neither the gender nor the smoking status having been specified. So smoking status for a two-year-old, you could imagine, is not often specified, but gender typically is.

And so the rule never fired because of that, and they have no idea how these changes were made. There's a complicated logging system that logs all the changes, and it crashed and lost its logging data. And it's a just so story.

Chlamydia screen-- this was human error. And so they wound up-- they found this very quickly, because they had a two-month-old boy who had numerous duplicate reminders, including suggestions for mammograms, pap smears, pneumococcal vaccination, and cholesterol screening, and a suggestion to start the patient on various meds. So this was just a human error in revising the rule, and that one they found pretty quickly.

So that's amusing. But what's interesting is these guys went on to say, well, how could we monitor for this in some ongoing fashion? And so they said, well, there's this notion of change point detection, which is an interesting machine learning problem, again.

And so they said, well, suppose we built a dynamic linear model that includes seasonality, because we have to deal with the fact that a lot of stuff happens Monday through Friday and nothing happens on weekends? And so they created a model that says that your output is some function, f, of your inputs, plus some noise. The noise is Gaussian with some variance, capital V, and that x evolves according to some evolution that says it depends on the previous value of x, plus some other noise, which is also Gaussian.

So that's the general sort of time series modeling approach that people often take. And then they said, well, we have to deal with seasonality. So what we're going to do is define a period, namely a week, and then we're going to separate out the states on different days of the week in order to give us the ability to model that seasonality.

I worked on a different project having to do with outbreak detection for infectious diseases, and there the periodicity was a year, because things like the flu come in yearly cycles rather than in weekly cycles. And so that idea is pretty common. And then they built this multiprocess dynamic linear model that says, basically, imagine that our data is being generated by one of a set of these dynamic linear models. And so we have an additional state variable at each time that says which of the models is in control to generate the data at this point.

And so if you have the set of observations up to some time, t, then you can compute the probability that model i is driving the generator at this point. And so you can

have three basic models. You can have a model that says it's a stable model, in other words, what you expect is the steady state. So that would be the normal weekly variation in volume for any of these alerts.

You can have a model which is an additive outlier. So that's something that says, all of a sudden, something happened, like that chlamydia screen or one of the other things that had a very quick blip. Or you can have a level shift change, like the change that happened when the screening rules or the alert rule for amiodarone stopped firing, because it went from one level to a very different level over a period of a relatively short period of time.

And then what you can do is calculate the probability of any of these models being in control at the next time, and that's called the change point score. And you can calculate this from the data that you're given. And of course, they have tons of data. It's a big hospital and lots of these alerts go on.

And if you plot this, there's the data for a time series. So you see the weekly variation. But what you see is that the probability of the steady behavior is quite high except at certain points where it all of a sudden dips. And so those are places where you suspect that something interesting is going on.

And similarly, the probability of a temporary offset goes up at these various points, and the probability of a level shift goes up at this point. And you can see that, indeed, there is a level shift from essentially 0 up to this periodic behavior in the original data sequence. And so they actually implemented this in the hospital, and so now you get not just alerts, but you get meta-alerts that say, this kid ought to be screened for their lead levels, but also the lead level screening rule hasn't fired as often as we expected it to fire.

Yeah, so there are a lot of details in the paper that you can look up, if you're interested. And what they find is that, if you look at the area under the delay false positive rate curve, so you're trading off how long it takes to be certain that one of these conditions has occurred versus how often you cry wolf, and you see that their algorithm does much better than a bunch of other things that they tried it against, which are earlier attempts to do this. And these are all highly statistically significant, so they got a nice paper out of it.

In the remaining time, I wanted to talk about a number of other issues that really have to do with workflow. So we've talked about alerting, but there are an interesting set of studies about how these alerting systems actually work. So there was a cool idea from the Beth Israel Deaconess Hospital here in Boston where they said, well, what we really need to do is to escalate alerts.

So, for example, it's quite typical in a hospital that, if you're a doctor and you have a patient who you have just sent their blood to the lab, and let's say there serum potassium comes back as 7 or 8, that patient is at high risk of going into cardiac arrhythmia and dying. And so your pager, in those days, goes off, and you read this text message that says, Mr. Jones has a serum potassium of 8. You'd better look in on him.

So what they did was very clever. They said, well, the problem is busy doctors might ignore this. And so we'll then start a countdown timer. And we'll say, did Dr. Smith actually come and look at Mr. Jones within 20 minutes?

And if the answer is no, then they send the page to the doctor's boss that says, hey, we sent this guy a page, and within 20 minutes he didn't look in on the patient. And then they start another timer. And they say, if that boss doesn't respond within an hour, then they send a page to the head of the hospital saying, you're her infectious disease people are doing a lousy job, because they're not-- or in this case, you're endocrine people, or whatever, are doing a lousy job, because they're not responding to these alerts.

Now, how do you think the doctors liked this? Not much. And there is a real problem with overalerting. And there is no general rule that says, how often can you bug the head of the hospital with an alert like this before he or she just says, well, turn off the damn thing, I don't want to see these?

And clearly, if you set the thresholds at different places, you get different results. So, for example, I remember Tufts implemented a system like this back in the 1980s, but they would send a page on every order where any of the lab results were abnormal, and that was way too much. Because a lot of these tests generate 20 results.

Normal is defined as the 95% confidence interval. What are the chances that out of

20 tests, which aren't really independent, but if they were, one of them would be pretty guaranteed to be out of range for most of the patients? And so basically every test generated an alert to the doctor. And the doctors did threaten to kill the people who had implemented the system, and it got turned off.

A system like this, if you set the threshold to be not abnormal, but life-threateningly abnormal, and if you set the rate and the time durations such that it's reasonable for people to respond to it, then maybe it can be acceptable. When we did this project on looking at how an emergency department could anticipate a flood of patients because it looked like flu season was starting, for example, the question we asked is, how many false alarms a month can you guys tolerate? And they thought about it. And the ED docs got together and said, three times a month you can cry wolf, because we really want to know when it actually happens. And we'd rather be prepared, and we can tolerate a 10% error rate on this prediction. But I don't know what it is in this domain.

Another interesting study was-- it's become quite popular. I got a bunch of emails from my doctor today, because I had ordered a refill on some prescription, and he wanted to know how it's going, and blah, blah, blah. So the BI asked the question, what fraction of those messages are never read by the patients that they're sent to? Which is an important question, because if you're relying on that mode of communication as part of your workflow, you'd like it to be 0.

It turned out only to be 3%, which is remarkably good. That means that most people are actually paying attention to those kinds of messages. Then I wanted to say a few words about the importance of communication and then finish up by mentioning some so far failed attempts at really good integration of all different data sources.

So as I said, the BI started in 1994 with a system that said, if you're taking a renally-excreted or a nephrotoxic drug, then we're going to warn people if there is a rising creatinine level, which is an indication that your kidneys are not functioning so well. Because, of course, if the drug is renally excreted, that means that if your kidneys are not excreting things at the rate they're supposed to, you're going to wind up building up the amount of drug in your body, and that can become toxic. So they saw a 21-hour, so almost a full day, reduction in response time from the medical

staff given these alerts versus what happened before. That's remarkable.

I mean, saving a day in responding to a condition like this is really quite an impressive result. And they also saw, in terms of clinical outcome, that the risk of renal impairment was reduced to about half of the preintervention level. So that earlier response actually was saving people's kidney function by getting people to intervene earlier. I found it interesting they said 44% of doctors found these alerts helpful, 28% found them annoying, but 65% of them wanted them continued to be used in a survey.

Enrico Carrera is one of my heroes. He used to be in the UK. He's now in Australia. And he had this very deep insight back in the 1980s.

He said, you know, all you computer guys who are treading on this medical field think that all of the action is about decision-making, but it's not. All of the action is really about communication, that health care is basically a team sport. And unless we spend much more time studying what goes on in communication, we're going to miss the boat. And then mostly, we didn't pay any attention to him, but he's kept at it.

So he said, well, how big is the communication space? So he cited a 1985 study that said that about 50% of requests for information are ones that people ask their colleague for versus 26% that they look up in their own notes. So if a doctor is on rounds, walks into a patient's room and says, I want to know has this guy's temperature been going up or down, a quarter of the time he'll look at notes. And half the time, he'll turn to the nurse and say, is this patient's temperature going up or down? So he says that's interesting.

Paul Tang did a study in the '90s that said that in a clinic, about 60% of the time is spent talking among the staff, not doing anything else. Enrico and one of his colleagues said that almost 100% of non-patient record information, in other words, the thing that's not in the written health record, is done by talking. That's almost tautological, because where else would you get it? And then Charlie Saffron at the BI did a time and motion study and was looking at, I think, nursing behavior, and saying that about half their time was face-to-face communication, about 10% with electronic medical records, and also a lot of email, and voicemail, and paper

reminders as ways of communicating among people.

So this was a study looking at-- this is that 1998 study by Colera and Tombs. And they're looking at a consultant, the house officer, another consultant. These are British titles, because this was done in Australia-- a nurse, et cetera.

And they say, OK, among hospital staff-- I think this was in one shift, I believe, I should have had that on the slide-- this is the number of pages that they sent and received. So they range from 0 up to about 4. The number of telephone calls made and received-- this ranges from 0 up to 13.

Oh, here's the length of observation. So this was over a period of about three hours for each of these patients. And this is the total number of events.

So think about it. In 3 and 1/2 hours, the senior house officer had 24 distinct communication events happen to that person. So that means, what, that's like 7-- yeah, like 7 an hour. So that's like 1 every 10 minutes, roughly.

So it's an interrupt-driven kind of environment. Here's one particular subject that they looked at, three and a quarter hours of observation. This person spent 86% of their time talking. 31% were taken up with 28 interruptions. So even the interruptions were being interrupted.

25% were multitasking with two or more conversations. 87%, face-to-face or on a phone or a pager. So most of that is talk time. And 13% dealing with computers and patient notes.

So the communication function is really important. And I don't have anything profound to say about it other than I'll put up a pointer to some of these papers. But the kinds of things they're considering are, well, we could introduce new channels, or new types of messages, or new communication policies that say, you know you may not interrupt the person who's taking care of patients while they're doing it, or something like that. And then moving from synchronous to asynchronous methods, like voicemail, or email, or Slack, or some modern communication mechanism.

Let me skip by these. Next to the last topic, quickly, how do you keep from dropping the ball? So there are a lot of analyses that say that the biggest mistakes in health care are made not because somebody makes the wrong decision, but it's because

somebody fails to make a decision. They just forget about something. They don't follow-up on something that they ought to.

The patient is going along, and you think everything's OK, and you don't deal with it. So inspired partly by that escalation of pagers that I read about at the Beth Israel, I said, well, this sounds like what we really need is a workflow engine that's approximately a discrete event simulator. So has anybody built a discrete events simulator in this class? It's a fairly standard sort of programming problem, and it's useful in simulating all kinds of things that involve discrete events.

And the idea is that you have a timeline, and you run down the timeline, and you execute the next activity that comes up. And that activity does something. It sends an email, or it shoots a rocket, or whatever field you're doing the simulation in. But most importantly, what it does is-- the last thing it does is it schedules something else to happen later in the timeline.

So, for example, for something that happens once a day, when it happens, the task that runs schedules it to happen again the next day. And that means that it's going to be continually operating all the time. So the idea I had was that what you'd like to do is to say, if at some time, t, I have a task that says do x or asks z to do y, or both, then the last thing should be at some time in the future schedule another task that says, is y done? And if not, then go notify somebody or go remind somebody.

And as far as I know, no hospital and no electronic record system has any capability like this, but I still think it's a terrific idea. And then I wanted to finish with a pointer to a problem that is still very much with us. So in 1994, some colleagues and I wrote this thing we called "The Guardian Angel Manifesto." And the idea was that we should engage patients more in their own care, because they can keep track of a lot of the things that systems didn't do a very good job of keeping track of.

And the idea was that you would have a computational process that would start off at the time your parents conceived you and run until your autopsy after you died. And during this time, it would be responsible for collecting all the relevant health care data about you. So it would be your electronic medical record, but it would also be active.

So it would help you communicate with your providers. It would help educate you about any conditions you have. It would remind you about things. It would schedule stuff for you, et cetera.

So this was a nice science fiction vision. And in the mid-2000s, Adam Bosworth, who was a VP of Google, came to me. And he said, you know, I read your thing. It's a good idea. I'm going to do it.

So Google started up this thing called Google Health, which was more focused on being at least the personal health record. They did a pilot with 1,600 people at Cleveland Clinic, and then they went public as a beta. And three years later, they killed it.

And they had a bunch of partners. So they had Allscripts, and Beth Israel, and Blue Cross of Massachusetts, and the Cleveland Clinic, and CVS, and so on. So they did their job of trying to connect to a bunch of important players. But, of course, they didn't have everybody.

And so, for example, I, of course, immediately signed up for an account, and the only company that I had ever dealt with out of that set was Walgreens, where I had bought a skin cream one time for a skin rash. And so my total medical record consisted of a skin rash and a cream that I had bought to take care of it-- not very helpful. And so nobody, other than these partners, could enter data automatically, which meant that you had to be even more anal compulsive than I am in order to sit there and type in my entire medical history into the system, especially, because if I did so, nobody would ever look at it.

Because if I go to my doctor and say, hey, Doc, here's the Google URL for my medical record, and here's the password by which you can access it, what do you think are the odds that they're actually going to look?

AUDIENCE: 0.

PETER SZOLOVITS: 0. So the thing was an absolute abject failure. And people keep trying it. And so far, nobody has figured out how to do it, but it's still a good idea. With that, we'll stop on workflow.