PETER SZOLOVITS: So last time we talked about what medicine does, and today I want to take a deep dive into medical data. And I'm going to use as examples a lot of stuff from the MIMIC database, which is one of the databases that we're going to be using in this class. Some of you are probably familiar with it, and some of you are not.

And there are, I hope, some takeaway lessons from this discussion. So for example, a few years ago, when MIMIC-III was about to be released, I was playing with the data, and I looked at the distribution of heart rates in the CareVue part of the database.

So MIMIC, for those of you who don't know, has intensive care data from about 60-something thousand admissions to intensive care units at the Beth Israel Deaconess Medical Center over a period of about 12 years. And one of the technical difficulties that we encountered is that in the middle of that time period. The hospitals shifted from one information system that they used in their intensive care unit to another. CareVue is the old one. MetaVision is the new one. And of course, they're not exactly compatible. So we'll see some examples of that.

So this is the old data. So this is from CareVue. And you look at that and say, well, heart rates range from 40 to 200 roughly, which is OK. But then there's this funny thing. There are two peaks. So where, if ever, do you see two peaks in physiological data? Not typical. And so my initial reaction was--

[LAUGHTER]

So then I looked a little closer, and I said, hmm, what do the heart rates look like from these two systems? And if you look in CareVue, you see the picture that I just showed you. And if you look in MetaVision, you see this other picture, which looks more like what you would normally expect.

And so I'm sitting there scratching my head going, OK, there must be some difference between these. It's not that simultaneous with the switchover of the hospital from one information system to another. Physiology of people changed, and all of a sudden some subset of people started having faster heart rates. Right? But if you think about that what subset of people have faster heart rates?

AUDIENCE: Athletes.

PETER SZOLOVITS: Hmm?

AUDIENCE: Babies?

AUDIENCE: If you're in a stress test.

PETER SZOLOVITS: Unh-hmm.

AUDIENCE: Is it children?

PETER SZOLOVITS: Yeah, kids. So I said, hmm, interesting. So anyway, if you look at the statistics, you see that the mean heart rate in CareVue is 108, and the mean heart rate in MetaVision is 87. But of course, means are not that meaningful when you look at these bimodal distributions.

So then I said, well, what if we just look at adults? So we look at people from age greater than 1 up to age 90. And I'll say a word about that in a minute. And I look at those two distributions. They look pretty close. They look pretty similar. So that means that the number of patients of different ages in the adult group is similar in the two data sets.

But if I don't exclude the very young or the very old, then I see this funny distribution where I have suppressed ages greater than 90 but not the young. And what you see is that in CareVue there's this giant spike at age 0.

So what happened at the hospital is that under the old system it was also being used in the NICU, the Neonatal Intensive Care Unit. And the new system was not being used in the NICU. And therefore, they didn't capture data about babies.

And in fact, if you look at age versus heart rate of the entire population, you see two very peculiar things. So here are the adults that we've been talking about, and here are the babies. And sure enough, they have higher heart rates. And then here are these 300-year-old people.

[LAUGHTER]

You go, wow, I don't think I'm going to have a heart rate when I'm 300 years old. So who are those people? Anybody have a clue? Yeah?

AUDIENCE: Entry errors.

PETER SZOLOVITS: Sorry?

AUDIENCE: Entry errors?

PETER SZOLOVITS: There are too many of them. Yeah, entry errors is always a possibility, but there's quite a few data points there. Yeah?

AUDIENCE: [INAUDIBLE]

PETER SZOLOVITS: Close. It's not quite missing data. So HIPAA, the Health Insurance Portability and Accountability Act, defines a set of criteria about protecting personal health information. And one of the things you are not allowed to do is to specify the age of somebody who is 90 years old or older. And the reason is because the number of 97-year-olds is pretty small.

And so if I tell you that Willy is 97 years old, then you're going to be able to pick him out of a population relatively easily, and so it's prohibited to say that. So as a result, everybody who's 90 or older gets labeled as being 300 years old in the database.

It's an artifact. It's like back in my youth, I worked as a computer programmer at a health sciences computing facility at UCLA. And we used to have a convention that missing data was represented by 99999. And of course, if you average that into a real data set, you get garbage, which people did regularly. So there are problems with this, and we're running into one of those.

If you look at just the adults, the two systems actually look very similar. So the blue and red dots, or the two systems, and I've drawn the trend lines between them, and you can see that they're very similar. So it looks like as you get older, your heart rate declines very slightly. But it does so equally in the two data sets. Yeah?

AUDIENCE: On the previous slide, beyond 300, it looks like they're older than 300?

PETER SZOLOVITS: Well that's because the ages there are computed at the time that the heart rate is measured. And so if you are 300 years old when you're admitted to the hospital, if you stay in the hospital for six months, then you're 300 and 1/2 years old by the time of that measurement. [LAUGHS] So that's why there are data points to the right of 300. Yeah, good catch.

OK, and then this is what the babies look like. And of course, they do have higher heart rates. And here here are the oldsters. So actually, there are people out to 310 years old because maybe they were discharged from the hospital. And then at age 100, they came back. You know, maybe they were 90 years old at the time they were initially admitted 10 years later. They came back, and we recorded more data about them, and so this is all relative to that 300.

OK, so that's just one example. And the lesson there is be careful when you look at data because it can really easily fool you 'cause there are all kinds of funny things about the way it's collected, about these artifactual things like 300-year-old patients and so on.

So here's a catalog of the types of data that are available to us. So we have the typical kind of electronic health record data from hospitals-- demographics, age, sex, socioeconomic status, insurance type, language, religion, living situation, family structure, location, work, et cetera.

We have vital signs-- your weight, your height, your pulse, respiration rate, body temperature, et cetera. So these are typically the things that if you ever go to a doctor's office, or you go into a hospital, the nurse will take you aside and weigh you and measure your height and check your blood pressure and take your temperature and stuff like that. These are standard vital signs, and so we have lots of those recorded.

Medications-- prescription medications, over-the-counter drugs, illegal drugs if you're willing not to lie to your health care provider, alcohol. Again, one of my earliest days, I was hanging out with a cardiologist at Tufts Medical Center, and we see this elderly lady who looks kind of terrible. And we're talking to her-- well, the doctor is talking to her. I'm trying to stay out of the way.

And he says, so do you drink alcohol? And she says, oh, no, never touch the stuff. And then we talk some more, and we go out of the patient's room. And the doctor turns to me out of earshot of the patient and says, oh, she's a chronic drunk.

I said, well, how do you know? And he says, well, from lab tests, from the appearance of her skin, from her general demeanor, from various sort of ineffable factors. And so patients lie. They really do because they don't want to tell you things.

Medications, by the way, is a big deal. So there is this whole field called med red, medication reconciliation, which is the hospitals or the doctors' offices attempt to figure out what medications you're actually taking. So I'm a member of the MIT health plan, and if I sign into my health plan account, it tells me that I'm taking some pills that I got 12 years ago as part of a laboratory test, where I took two pills which were supposed to have some physiological effect, and then they measured that. And I've never gotten another pill and never taken one since then, nor would it be particularly good for me. But it's still on my record, and there's no notice of it ever having been discontinued. And that's a real problem because if you're taking care of a patient, you'd like to understand what drugs they're actually taking, and it's hard to know.

Then lab tests-- so this is the things that you imagine that we do a lot of, and these are components of the blood and the urine mainly, but also of the stool, saliva, spinal fluid, fluid taken off the belly, joint fluid, bone marrow, stuff coming out of your lungs. It's anything and any place where you can produce some specimen, they can send it to a lab and measure things in it, and they measure lots and lots of different kinds of things. And these are often useful.

Pathology, qualitative and quantitative examination of any body tissue, for example, biopsy samples or surgical scraps. You know, if they do an operation, they cut something out of you, that typically winds up on a pathologist's bench, who then tries to figure out what its characteristics are and that's, again, useful information.

Microbiology-- ever since Pasteur, we know that organisms cause disease. And so we're quite interested in knowing what organisms are growing inside your body. And typically, testing is not only to identify the organism but also to figure out which antibiotics it's sensitive to and insensitive to.

And so you'll see things like reports of sensitivity testing at various dilutions. In other words, they try to give a strong dose of an antibiotic a week weaker dose a week or dose a weaker dose a week or dose to see which is the minimum level of dosing that's enough to kill the bacteria.

There's a comma missing there, but input, output of fluids is another important thing because people, especially in the hospital, often get either dehydrated or over hydrated. And neither of those is good for you, and so trying to keep track of what's going into you and what's coming out of you is important.

Then there are tons of notes. So an important one that we're going to look at in this class is discharge summaries. So these are the typically long notes that are written at the end of a hospitalization. So this is a summary of why you came in, what they did to you, the main things they discovered about you, and then plans for what to do after your discharge. Where are you going to go? What drugs are you going to be taking? When are you supposed to come back for follow up, et cetera. I'll show you an excruciatingly long one of those later in the lecture today.

But we also have notes from attendings and/or residents, nurses, various specialties, consultants. The referring physician-- if somebody sends you to the hospital, that doctor will usually write a note saying this is what I'm interested in. Here's why I'm sending in the patient. There are letters back to the referring physician saying, OK, this is what we found out. Here's the answer to the question you were asking. There are emergency department notes. So that's often the first contact between the patient and the health care system. So these are all important.

And then there's tons and tons of billing data. So remember the EHR systems were initially designed by accountants. And they were designed for the purpose of billing. And so we capture a lot of data about formalized ways of describing the condition of the patient and what was done to the patient in order to submit the right bills.

You obviously want to bill through it as much as possible. But you have to be able to justify the bills that you submit because insurance companies and Medicare and Medicaid don't have a good sense of humor. And if you submit bills for things that you can't justify, then you get penalized.

And then there are administrative data like, which service are you on? So this this is occasionally a confusing thing. You can go into the hospital and have heart problems, but it turns out that the heart intensive care unit, the cardiac intensive care unit, is full up with patients. But there's an extra bed in the pulmonary intensive care unit, and so they stick you in that unit, but you're still on the cardiology service. And so there are these sort of mixture kinds of cases that you still have to take care of. Transfers are when you get transferred from one place to another in the hospital.

Imaging data-- so I'm not going to talk about that much today, but there are X-rays, ultrasound, CT, MRI, PET scans, retinal scans, endoscopy, photographs of your skin and stuff like that. So this is all imaging data, and there's been a tremendous amount of progress recently in applying machine learning techniques to try to interpret the contents of these data.

So these are also very important.

And then there's the whole quantified self movement. I mean, how many of you where an activity tracker? Only about 1/3? I'm surprised at a place like MIT.

[LAUGHTER]

So you know, we measure steps and elevation change and workouts. And you can record vital signs and diet and your blood sugar, especially if you're diabetic; allergies, allergic incidents. There's all this mindfulness, mood, sleep, pain, sexual activity.

And then people have developed this idea of N of 1 experiments. For example, I had a student some years ago who suffered from psoriasis. It's a grody condition of the skin. And the problem is there are no good cures for it. And so people who suffer from psoriasis try all kinds of things. You know, they stop eating nonce for a while, or they douse themselves with vinegar. Or they do whatever crazy thing comes to mind. And we don't have a good theory for how to treat this disease.

But on the other hand, some things work for some people. And so there's a whole methodology that has been developed that says, when you try these things, act like a scientist. Have hypotheses. Take good notes. Collect good data. Be cognizant of things like onset periods, where you know you may have to drip vinegar on yourself for a week before you see any effect. So if that doesn't do a thing after one day, don't stop. And furthermore, if you stop then don't start something new immediately because you will then be confused about whether this is the effect of the thing you were on before or the new thing that you're trying. So there's all sorts of ideas like that.

So this is a slide from our paper on MIMIC-III. And it gives you a kind of overview of what's going on with the patient. So if you look at this-- I'm going to point with my hands-- at the top is something very important. This patient starts off at full code. That means that if something bad happens to him, he wants everything to be done to try to save him. And he winds up in comfort measures only, which means that if something bad happens to him, he wants to die-- or his family does if he's unconscious.

So what else do we know about this guy? Well GCS is the Glasgow Coma Score. And it's a way of quantifying people's level of consciousness. And you see that at the beginning this patient is oriented, and then gets confused. And finally, is only making incomprehensible words or sounds. Motor, he's able to obey commands. Eventually, he's only able to flex when you stimulate his muscles. So he's no longer conscious.

Eye movements-- he's able to follow you spontaneously. He's able to orient to speech. And eventually orientation at all. So this is clearly somebody who's going downhill quickly and, in fact, dies at the end of this episode.

Now, we then look at labs so we can see what is their level of platelets at about the time that they're measured, their creatinine level, their white blood cell count, the neutrophils percentage, et cetera. And there's not every possible data point on the slide. This is just illustrative.

The next section is medications. So the person is on morphine. They're on Vancomycin, which is an antibiotic. Piperacillin-- I don't know what that is. Does somebody know?

AUDIENCE: Antibiotic.

PETER SZOLOVITS: It's what?

AUDIENCE: It's antibiotic.

PETER SZOLOVITS: OK. Sodium chloride 9%, So that's just keeping him hydrated. Amiodarone and dextrose. So dextrose is giving him some energy.

And then these are the various measurements. So you see the heart rate, for example, is up pretty high and is going up near the end. The oxygen saturation starts off pretty good. But here we're down to 60% or 50% O2 sat, which is supposed to be above about 92 in order to be considered reasonable. So again, this is a very consistent picture of things going very badly wrong for this particular patient. So this is all the data in the database.

Now, if you want to try to analyze some of this stuff, you can say, well, let's look at the ages at the time of the last lab measurement in the database. So we have the times of all the lab measurements.

So we can see that many of the ICU population are fairly old. There's a relatively small number of young people and then a growing number of older people in both females and males. If we look at age at admission by gender-- so this is age at admission not age at the time the last lab measurement was done-- it's a pretty similar curve. So we see that females were 64.21 at time of last lab measurement; 63.5 at the time of admission.

So we can look at demographics, and demographics typically includes these kinds of factors, which I've mentioned before. And again, if we're interested in the relationship between this and, for example, the age distribution, we see that if you look at the different admission types-- so you can be either admitted for an emergency for some urgent care or electively. And it doesn't seem to make a whole lot of difference, at least in the means of the population age distribution.

On the other hand, if you look at insurance type and, say, who's paying the bills, there is a big difference in the age distributions. Now, why do you think that private insurance drops way off at about 65?

AUDIENCE: Isn't insurance always covered for everyone by the state health?

PETER SZOLOVITS: It's because of Medicare. So Medicare covers people who are 65 years old. There's a terrible story I have to tell you. I was talking to somebody at an insurance company who's a bit cynical, and he said suppose that you see a 63-year-old patient who's developing type 2 diabetes, what should you do for him?

Well, there are standard things you should do for somebody developing type 2 diabetes, like get him to eat better, get him to lose weight, get him to exercise more, et cetera, et cetera. But his cynical answer was absolutely nothing.

Why? Well it's very cheap to do nothing. Most people who develop type 2 diabetes don't get real sick in the next two years. And by the time this patient is 65, he'll be the government's responsibility, not the insurance company's. Nice.

AUDIENCE: Yeah.

PETER SZOLOVITS: So of course a lot of the elderly are insured by Medicare or Medicaid, not that surprising. Self-pay is a pretty small number because it's insanely expensive to pay for your own health care.

What about where you came from? Were you referred from a clinic, or were you an emergency room admit? Or were you referred from an HMO or et cetera? And other than a transfer from a skilled nursing facility or transfer within the facility, within the hospital, it doesn't make much difference. The averages there and the distributions look moderately similar.

If you're coming from a skilled nursing facility, if you are in a skilled nursing facility, you're probably old because younger people don't typically need skilled nursing care. And I'm not sure why transfers within the facility are significantly younger ages, but that's true from the MIMIC data.

What about age at admission by language? So some people speak English. Some people speak not available. Some people speak Spanish, et cetera. So it turns out the Russians are the oldest. And that may have to do with immigration patterns, or I don't know exactly why. But that's what the data show.

If you do it by ethnicity, it turns out that African-Americans, on the whole, are somewhat younger than whites. And Hispanics are somewhat younger yet. So that means that those subpopulations apparently need intensive care earlier in life than whites. So this is a topic that's very hot right now, discussions about how bias might play into health care. Yeah?

AUDIENCE: What does unable to obtain mean?

PETER SZOLOVITS: It just means that somebody refused to say what their ethnicity was.

AUDIENCE: When they were asked this?

PETER SZOLOVITS: Yeah. I think. I'm not positive.

AUDIENCE: So just to confirm. This also represents Boston's population dynamics too, right?

PETER SZOLOVITS: It's the catchment basin of the Beth Israel Deaconess Hospital, which is Boston clearly. But there are-- it turns out that a lot of North Shore people go to Mass General, and so different hospitals have different catchment basins.

AUDIENCE: Does it have anything to do with like, is this just the ICU? Or is this everybody who goes to the hospital or the ER?

PETER SZOLOVITS: These are all people who at some point were in the ICU. So these are the sicker patients. Yeah?

AUDIENCE: So just want to double-check there's a higher proportion of black, African American people in the population here as well because the red is higher than the others?

PETER SZOLOVITS: No, actually-- I don't remember if I have that graph-- I think this is cumulative.

AUDIENCE: Oh, OK.

PETER SZOLOVITS: So most people are white for whatever definition of white we're using. And I think it's only the increment that you see on top.

All right, how about marital status? Well, according to this, it's bad to be single. So I could sort of see that for hospitalization. I'm not sure why it's true for the ICU because if you don't have anybody at home to take care of you when you get sick, it seems reasonable that you'd be more likely to wind up in the hospital. But I don't know why you'd wind up in intensive care. Yeah?

AUDIENCE: Isn't it possible that those are also single people are probably younger than married people, and those are probably younger than--

PETER SZOLOVITS: Yes, yeah.

AUDIENCE: [INAUDIBLE] people.

PETER SZOLOVITS: Yeah, that's probably also right.

So here's an interesting question, a little bit related to something you'll see on the next problem set. So could we predict in-hospital mortality from just these demographic features? So I'm using a tool in language called R. This is a general linear model, and I've set it up to do basically logistic regression. And it says I'm predicting whether you die in the hospital based on these demographic factors.

And it turns out that the only ones that are highly significant are age. So that's not surprising, that older people are more likely to die than younger people. It's generally true. And if I'm unable to obtain your ethnicity, or I don't know your ethnicity, then you're more likely to die. I have no clue why that might be the case.

And other things are not as significant. So if you speak Spanish or English, you're slightly less likely to die. You see a negative contribution here. And if you speak Russian, you're slightly less likely to die. But it's significant not at the p equal 0.05 level, but it is at the p equal 0.06 level. And marriage doesn't seem to make much difference in predicting whether you're going to die or not.

Now, remember, this is ICU patients. And we're looking at in-hospital mortality.

AUDIENCE: For ethnicity, can they learn that at any point in this study, or just right at the beginning? Or do you know? Because I don't know.

PROFESSOR: I don't know.

AUDIENCE: Because it could be that unable to obtain means that they died before we can ask them.

PROFESSOR: No, because there wouldn't be that many of those people, I think. There are not that many people who don't live past the intake interview. And they do ask them.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, that would be an example. But I don't think you'd see enough such people to show up statistically.

OK. Well, so I've already mentioned that there is this problem of having moved from CareVue view to MetaVision just in the MIMIC database. But of course, this is a much bigger problem around the country and around the world, because every hospital has its own way of keeping records. And wouldn't it be nice if we had standards?

And of course, there's this funny phrase, the wonderful thing about standards is that there's so many to choose from. So for example, if you look at prescriptions in the MIMIC database, here are two particular prescriptions for subject number 57139 admitted on admission ID 155470. And so they have the same start date but different end dates. One is a prescription for Tylenol, acetaminophen, and the other is for clobetasol propionate 0.05% cream. That's a skin lotion thing for-- I think it's a steroid skin cream.

So if you look in the BI's database, they have their own private formulary code where this thing is acet325 and this thing is clob.05C30, right? And if you look, there there's also something called a GSN, which is some commercial coding system for drugs. Maybe having to do with who their drug supplier is at the hospital. And these have different codes.

There's the National Drug Code, which is an FDA assigned nine digit code that specifies who made the drug, what form it's in, and what's its strength. And so you get these. Then there's a human readable description that says Tylenol comes in 325 milligram tablets. And the clobetasol comes in 30 gram tubes. And the dose is supposed to be 325 to 650, i.e. one to two tablets measured in milligrams. The dose here is one application, whatever that is.

I don't know what the 0.01 means. And this is a tablet and that's a tube. And this is taken orally. That's administered on the skin, right? So this is a local database.

AUDIENCE: For a doctor, they just [INAUDIBLE]

PROFESSOR: At most hospitals, that's true now. It wasn't true when the MIMIC database started being collected. And the BI was relatively late in moving to that compared to some of the other hospitals in the Boston area. Each hospital has its own digitorata for what it thinks is most important. And I think the BI just didn't prioritize it as much as some of the other hospitals.

OK, so then I said, well, if you look at prescriptions, how often are they given? So remember, we have about 60,000 ICU stays. And so iso-osmotic dextrose was given 87,000 times to various people. Sodium chloride 0.9 percent flush. Do you know what that is? Have you ever had an IV? So periodically, the nurse comes by and squirts a little bit of stuff in the IV to make sure that it hasn't clogged up. That's what that is.

Insulin, SW. I don't know. Salt water? I don't know what SW is. Magnesium sulfate, dextrose five in water. Furosimide is a diuretic. Potassium chloride replenishes potassium that people are often low on. And then you go, so why is there this D5W and that D5W? And that's probably some data in the system, OK? One of them has an NDC code associated with it and the other one doesn't but probably should. Yeah.

AUDIENCE: I was actually going to ask, does yours mean that they're standard across hospitals or just that we don't have the data?

PROFESSOR: The NDC code should be standard across the country, because those are FDA assigned codes. But not every hospital uses them, OK? And for the ones that say zero, I'm not sure why they're not associated with a code in this hospital's database.

OK, next most common, you see normal saline, 0.9 percent sodium chloride. So that was the same stuff as the flush solution but this time not being used for flush. Metoprolol is a beta blocker. Here's another insulin this time with an NDC code, et cetera.

I love bag and vial, OK? So these are not exactly medications. A bag is literally like a baggy that they put something into, and a vial is literally something that they put pills in. And why is that in the database? Because they get to charge for it, OK? And I don't know what the charge is, but it wouldn't surprise me if you're paying $5 for a plastic bag to put something in.

OK, so if we say, well, how many pharmacy orders are there per admission at this hospital, and the answer is a lot. So if you look at-- it's a very long tailed distribution, goes out to about 2,500. But you see, if I blow up just the numbers up to about 200, there's a very large number of people with two prescriptions filled, and then a fairly declining number with more. And then it's a very long tail. So can you imagine 2,500 things prescribed for you during a hospital stay?

Well, a little more about standards, so NDC is probably the best of the coding systems. And it's developed by the FDA. The picture up on the top right shows that the first four digits are the so-called labeler. That's usually the person who produced the drugs, or at least the person who distributes them. The second four digit number is the form of the drug, so whether it's capsules, or tablets, or liquid, or whatever and the dose. And then the last two digits are a package code which translates into the total number of doses that are in a package, right?

So this is a godsend. And all of the robotic pharmacies and so on rely on using this kind of information nowadays. Unfortunately, they ran out of four digit numbers, and so there's now a-- they added an extra digit, but they didn't do it systematically, and so sometimes they added an extra digit to the labeler and sometimes to the product code. And so there is a nightmare of translations between the old codes and the new codes. And you have to have a code dictionary in order to do it properly and so on.

OK, well, if that weren't good enough, the International Council for the Harmonization of Technical requirements for Pharmaceuticals for Human Use developed another coding system called MedDRA, which is also used in various places. And this is an international standard, which is, of course, incompatible with the NDC.

CPT is the Common Procedural Terminology, which we'll talk about in a little bit. And they have a subrange of their codes which also correspond to medication administration. And so this is yet another way of coding giving medicines. And then the HCPCS is yet another set of codes for specifying what medicines you've given to somebody.

And then I had mentioned this GSN number, which apparently the Beth Israel uses. This as a commercial coding system from a company called First Databank that is in the business of trying to produce standards. But in this case, they're producing ones that are pretty redundant with other existing standards. But nevertheless, for historical reasons, or for whatever reasons, people are using these.

OK, enough of drugs. So what procedures were done to a patient? If you look in MIMIC, there are three tables. There's procedures ICD, which has ICD-9 codes for about a quarter million procedures. There's CPT events, which has about half a million, 600,000 events that are coded in the CPT terminology. And then MetaVision, the newer of the two systems, has about a quarter million procedure events that are coded in that system.

So some examples, here's the most common ICD-9 procedure codes. So ICD-9 code 3893 of which there are 14,000 instances is venous catheterization, not elsewhere classified. So what's venous catheterization? It's when somebody sticks an IV in your vein, OK?

Very common. You show up at a hospital. Before they ask you your name, they stick an IV in your arm. That's a billable event, too. Then insertion of an endotracheal tube, you know, if you're having any problems like that, they stick something down your throat. Ventral infusion of concentrated nutritional substances, so if you're not able to eat, then they feed you through a stomach tube, OK? So that's what that is.

Continuous invasive mechanical ventilation for less than 96 consecutive hours, so this is being put on a ventilator that's breathing for you, et cetera. So you see that there is a very long tail of these. So those are the ICD-9 codes.

Now, CPT has its own procedure codes that go into a tremendous amount of detail. So for example, this is the medicine subsection, and it shows you the kinds of drugs that you're being administered that are involved in dialysis, or psychiatry, or vaccines, or whatever. And then here are the surgical and the radiological codes. And there's tons and tons of detail on these. Yeah.

AUDIENCE: So how can they put these codes as 1,000 to 1,022? This is really annoying for anyone--

PROFESSOR: No, these are categories. So if you drill down, there's a fanout of that tree and you get down to individual codes. Just as a nasty surprise, CPT is owned by the American College of Physicians, and they could sue me if I showed you the actual codes because they're copyrighted. And you have to pay them if you use those codes. It's crazy.

OK, so if you look at the number of all of these codes per admission, you see a distribution like this. Or if I separate them out, you see that there are more ICD-9 codes and fewer of the CPT and the codes that are in MetaVision. But they look somewhat similar in their distributions.

OK, lab measurements. So you send off a sputum sample, blood, urine, piece of your brain, something. They stick it in some goo and measure something about it. So what is it that they're measuring?

Well, it turns out that hematocrit is the most common measurement. So this is how much hemoglobin is in your blood, or what fraction in your blood, and is very important for sick people. And the second most important is potassium, then sodium creatinine, chloride, urea nitrogen, bicarbonate, et cetera. So this is a long, long list of different things that can be measured, and all the stuff is in the database.

So for example, here's patient number two in the database. And on July 17 of 2138, this is part of the deidentification process to make it difficult to figure out who the patient actually is. This person got a test for their blood and they reported atypical lymphocytes.

So there are a couple of interesting things to note here. One is that some things have a value and others don't. So this is a qualitative measure, so there's no value associated with it. Just the fact of the label tells you what the result of the test was. The other thing that's interesting is this last column, which is LOINK, and I'll say a word about that in a minute-- actually right now.

So LOINK is the Logical Observation Identifiers Names and Codes. It was developed by our colleagues at Regenstrief Clinic in Indiana about 15 years ago, maybe 20 years ago at this point. And the attempt was to say every different type of laboratory test ought to have a unique name, and they ought to be hierarchical so that if you have, for example, three different ways of measuring serum potassium, that they're related to each other but that they're distinct from each other, because there may be circumstances under which the errors that you get from one measurement versus another are different.

And so this is the standard way. If you send off your blood sample to a lab, they send back a string like this to the hospital or to your doctor's office that says, it's coded in this OBX coding system, and here is the LOINK code, and this is the SNOMED interpretation. And so this string is the way that your hospital's EHR or your doctor's office system figures out what the result of the test was.

HL7 is this 30-something year old organization that has been working on standardizing stuff like this. And LOINK is part of their standardization. So if you look at these, you say, well, again, how many tests per admission? Again, a huge, long tail up to about 15,000 for a very small number of patients.

If you look at lab tests per admission, you can do a log transform and get something that looks like a more reasonable distribution. By the way, that's a very generic lesson when we're going to do analyses of these data, is that, often, doing a transform of some sort, like in this case, a log, takes some funny looking distribution and turns it into something that looks plausibly normal, which is better for a lot of the techniques we use. Yeah.

AUDIENCE: [INAUDIBLE] means the same thing? Like, for instance--

PROFESSOR: Yes.

AUDIENCE: --hematocrit [INAUDIBLE]

PROFESSOR: Yes

AUDIENCE: --same?

PROFESSOR: Yes

AUDIENCE: Always same?

PROFESSOR: Yes, that's the whole idea of creating the standard. And that has been pretty successful, pretty successfully adopted.

OK, chart events. So these are the things that nurses typically enter at the bedside. And so there are 5.1, 5.2 million heart rates measured in the MIMIC database. And calprevslig is an artifact. It exists in every record. And it's some calibration something or other that doesn't mean anything. I've never been able to figure out exactly what it is.

SPO2 is the partial pressure of oxygen in your blood. If you use a pulse oximeter, that's what that's measuring. Respiratory rate, heart rhythm, ectopy type, dot, dot, dot.

Now, you might be troubled by the fact that here is heart rate again, right? But I've already shown you this, that heart rate in CareVue and heart rate in MetaVision were coded under different codes in the joint system that we created out of those two databases. And so you have to take care of figuring out what's what if you're trying to analyze this data. Not only do we have that problem of different age distributions across the two different data sets, but we also just have the mechanical problem that there will be things with the same label that may or may not represent the same measurement at different times in the system.

OK, this is the number of chart entries per admission, again, on a log scale. So you see that there are about 10 to the 3.5 chart entries per admission, so thousands of admissions, of chart events per admission. We also track outputs. So Foley catheter allows your bladder to drain without your having consciously to go to the bathroom, so they collect that information. There are 1.9 million recordings of how much fluid came out of your bladder.

Chest tubes will drain stuff out of your chest if you have congestion. Urine is if you pee regularly, stool out, et cetera. And again, I'm not sure I understand what the difference is between urine out Foley versus Foley. They may be the same thing but one from CareVue and one from MetaVision, so again, typical kinds of problems.

If you look at the number of output events per admission, you're seeing on the order of 100, roughly. Well, if you're tracking outputs, you should also track inputs, and so they do. And so D5W is this dextrose in water, 0.9 percent normal saline. Propofol is an anesthetic. Insulin, heparin, blood thinner, et cetera. Fentanyl is, I think, an opioid, if I remember right.

So these are various things that are given to people. And they affect the volume of the person. So this is an attempt to keep the person in balance and keep track of that. MetaVision inputs are classified somewhat differently but they have similar kinds of data. And if you combine them, you get, again, a distribution on a log scale that shows that there are on the order of 10 to the fifth input events, so quite a few input events, because this is recorded periodically.

Now, the paper that I-- yeah.

AUDIENCE: What's the input again? Is that when you come to the hospital and get admitted or--

PROFESSOR: No, no, no. It's an input into you. So it's like you drink a glass of water, the nurse is supposed to record it. Although, she doesn't always because she may not notice it. But if they hang an IV bag and pour a liter of liquid into you, they do record that, OK?

All right, so I had you read this interesting paper and a discussion prior to that paper, because one of the authors is a former student of mine. And I know one of the other guys pretty well. And the former student, Zak Kohane, came back some years ago from a conference in California and was explaining to me that he ran into a venture capitalist who discovered that there is an interesting physiological variation in the abnormality of lab tests that are done at night. And he suspected that there was a diurnal variation that lab tests actually become more abnormal at night than they do during the day.

And Zak, who is not only a computer science PhD but also a practicing doctor, turns to him and says, you're an idiot, right? Who has their blood drawn at 3 o'clock in the morning. It's typically not healthy people, right? So this is another of these nice confounding stories where, if you have a test done in the middle of the night, it probably indicates that you're sicker.

So he and Griffin recruited their third author and went off and did a very large scale study of this question, which is what the paper that I asked you to read reports on. And so I said, well, I wonder if I could reproduce that study in the MIMIC database. And the answer, just in case you get your hopes up, was no, in large part because we just don't have the right kind of data. So there are not that many white blood counts that were measured in the MIMIC database, for example.

But if you look at the-- this is MIMIC data. And if you say, what's the fraction of abnormal white blood count values by hour-- so this is midnight to midnight. And each hour, there's some fraction of these test results that are abnormal. And sure enough, what you see is that, at 5 o'clock in the morning, a much higher fraction of them is abnormal than at 3 o'clock in the afternoon, OK, which is consistent with Zak's peremptory comment about the guy being an idiot.

So once again, I said, well, can we build a really simple model that predicts who's going to die in the hospital in this case? That's the easiest one to predict because I have that data. We could get three-year survival data, which is what they were looking at. But it's harder and it runs into censoring problems of what happens if the person was hospitalized less than three years before the end of our data collection period and so on. And so I avoided that.

But what this is showing you is, for each of the hours, zero to 24, what is the number of measurements? And for each of those hours, what is the fraction of those measurements that's abnormal, OK? So I said, well, let's just throw it into a logistic regression model. And what comes out is something really weird, which is that a few particular hours are significant, but most of them are not.

And that looks like noise to me, right? Because you wouldn't expect that, at 8 o'clock in the morning, the fact that you had something measured matters. But at 9 o'clock in the morning, it doesn't. That doesn't seem sensible. So I don't think there's enough signal here.

And in fact, when I looked at the number of white blood count measurements at night and related to mortality-- so false means people lived and true means they died. But you see that there's not a whole lot of difference between the distributions. But you also see that the number of white blood counts is relatively small in this database. And so I think we just don't have enough data to do it.

On the other hand, if you look at a panel of different drugs, you look at mean values of blood urea nitrogen or calcium chloride, CO2, et cetera, you see that there is variation across time. So there is some sort of variance that's either caused by the diurnal physiology of the human body or by the routine practice of medicine, about when people choose to take lab measurements. And in fact, if you look at the fraction of high end low lab values, they do vary by hour. And in particular, if you look at white blood counts, you see that the fraction of high values goes up at night and the fraction of low values goes down at night, right, which is consistent with what they saw as well.

There is another way to measure it, which is, instead of using normal ranges, the lab actually gives you a call that says, is this value normal, low, or high? And we can use that. That's a little bit more subtle because it depends on calibration of the equipment and is updated as the calibration changes. So that's probably a little bit more accurate. But you see essentially the same phenomenon here.

But if you look at the distributions of when measurements are done that turn out to be normal versus when they turn out to be abnormal, there is a lot of similarity between the normal and the abnormal curves of when those measurements are taken. So we're not seeing that.

OK, let me race through to the end. This is my heartbeat from my watch. You can actually download the stuff and put it in your favorite analysis engine and take a look. So here I was running across the Harvard bridge. And if you look at my heart rate variability over the 30 seconds or so, you see that the interbeat interval ranges from about 550 to about 600 and whatever 20 milliseconds.

And so you could calculate my heart rate variability, which is thought to be an indicator of heart health and so on. You can calculate that I was running at a pace of about 100-- my heart was beating at a pace of about 100 beats per minute. So you know there's all sorts of information like that available.

Now, as I said, I'm not going to get into this today, but this was a very successful recently published paper where they're able to take a look at images of the lung. So this is a transverse scan of the lung. And they have a deep learning machine that is able to identify these two yellow marked things as pulmonary emboli as opposed to these other things that are just random flecks in the tissue. And I can't do that by eyeball. Maybe a good radiologist might be able to, but this is claimed in the paper to outperform decent radiologists already.

This was one of the articles that led Geoff Hinton to make this rather stupid pronouncement that said, tell your children not to become radiologists because the profession will be over by the time they get fully trained, which I don't believe. They may do different things, but they won't go away.

This was a slide from Ron Kikinis at the Brigham, and they're using automated techniques of analyzing white matter in order to identify lupus lesions. So lupus is a bad disease that shows up in these magnetic resonance images in certain ways.

The last thing I want to talk about today is notes. So my students did a little exercise last semester where we tried to see how good is the average ape, namely member of my research group, at predicting mortality? And so we took a bunch of cases from the MIMIC data set, blinded to the question of whether the person lived or died. We gave the data to people in a kind of visualization tool, sort of like the one that I showed you earlier, that summarizes the case, and then also gave people access to the notes, the deidentified notes about those cases, to see whether people could predict, better than a coin flip, whether somebody was going to live or die.

And the answer is yes, slightly better, OK? Not immensely better but slightly better. And furthermore, it looks like, by giving them feedback, so as they're looking at these cases and trying to make the prediction, they make a prediction, you tell them if they were right or wrong, we learn. And so we get slightly better than slightly better than random, right? It's kind of interesting.

OK, so one of the things I discovered is that, at least when I was playing the monkey in this exercise, I found the notes to be immensely useful, much more useful than the trend lines of laboratory data. Partly, it's because I'm used to reading English. I'm not so used to reading graphs of laboratory data. But part of it is that there is a level of human understanding that is transmitted in the nursing notes and in the discharge summaries and so on that you don't get from just looking at raw data.

And so there is very much the sense, which we're going to talk about in a couple of weeks, of how can we take advantage of that information, extract it, and use it in the kinds of modeling that we want to do? So in MIMIC, if you look, we have nursing notes, and radiology reports, and more nursing notes, and electrocardiogram reports, and doctor's notes, and discharge summaries, and echocardiograms, respiratory, et cetera. And if you look at the distribution of the lengths of these, these are, unfortunately, not on the same scale.

But the discharge summary is the thing that's written at the time you leave the hospital. So this is sort of the summary of everything that happened to you during your hospitalization. And it's long. So, you know, it goes up to like 30,000 characters. You know, it's a short story, not so short short story.

Nursing notes tend to be shorter. They run up to about 3,000 characters. This other set of nursing notes, which I think comes from the other system, is a little bit longer. It goes up to about 5,000.

Doctor's notes are a little bit longer yet. They go up to about 10,000, 15,000 characters, typically. And there are various other kinds of notes. So I just wanted to show you a few of these.

Here's a brief nursing note. So this is a patient who is hypotensive but not in shock. Patient remains on this drug drip at 0.75 micrograms per kilogram per minute, no titration needed at this time. Their blood pressure is stable at more than 100. Their mean arterial pressure is 65, greater than 65. Wean them from this drug presumably if it's tolerated.

A wound infection, so anterior groin area open and oozing moderate amounts of thin, pink-tinged serous fluid. Patient's stooling with small amounts of stool on something and dangerously close to the open wound, et cetera. So this is sort of the nurse's snapshot. She just went in, saw the patient-- by the way, I say she, but probably a vast majority of nurses in Boston area hospitals really are women, but there are some male nurses-- and will record sort of a snapshot of what's going on with the patient.

What are the concerns? In principle, this is going to be useful not only as a part of the medical record, but also when this nurse goes off shift and the next nurse comes on shift. Then this is a recording of what the state of the patient was the last time they were seen by the nurse. In reality, the nurses tend to tell each other verbally rather than relying on the written version.

I remember one time talking to a nurse in an intensive care unit in another part of the country, and I said, so whoever reads your notes, and she says, quality assurance officers, so the hospital has people responsible for trying to assess the quality of care they're giving, and lawyers when there's a lawsuit. And she was very happy because she had saved the hospital 10 million dollars by having carefully recorded that some procedure had been done to a patient who then had a bad outcome and was suing the hospital for their neglect in not having done this. But because it was in the note, that was proof that it actually had been done, and therefore the hospital wasn't liable. But there is a lot of information in here.

Now, I'm going to show you many pages of a typical discharge summary. So this is somebody on the surgery service who came in complaining of leg pain, redness, and swelling secondary to infection of the left femoral popliteal bypass. So she had surgery-- I think she. Yeah, female. She had surgery which didn't heal well, so major surgical or invasive procedure, incision and drainage and pulse irrigation of the left groin, and left above-knee popliteal site incisions with exploration of bypass graft, and excision of the entire left common femoral artery to above-knee blah, blah, blah, blah blah, blah. So this is what they did.

History of the present illness-- she's a 45-year-old woman who underwent the left femoral, a.k.a. doctor something or other with PTFE, whatever that is, over a month ago on a certain date. By the way, these bracketed asterisked things are where we've taken out identifying information from the record. She had been doing well post-operatively and was seen in the clinic six days prior to presentation. At this time, she acutely developed nausea, vomiting, fevers, and progressive redness, swelling, pain of her left thigh, et cetera, OK? So that's just page one of many pages. Yeah.

AUDIENCE: Just a question. Is this completely [INAUDIBLE] information [INAUDIBLE] patient's name or date?

PROFESSOR: Not in this system. There are people-- Henry Chueh at Mass General spent 10 years building a system that had autocomplete and so on. And some doctors liked it and some doctors hated it. And the MGH threw out all of their old systems in order to buy Epic, and so it's gone. It was like 10 years of work down the drain. But it was not a spectacular success.

Because whenever you have auto complete, you have to anticipate every possible answer. And people are very creative, and they always want to type something that you didn't anticipate. So it's hard to support it.

AUDIENCE: What is Epic? That's like the new--

PROFESSOR: Epic is a big company that has been winning all the recent contests for installing electronic medical record systems. Remember in my last lecture, I showed that we're reaching about 100% saturation? So they've been winning a lot of the installation deals. And they're getting a lot of the subsidy. The estimate I heard was that Partners Healthcare, which is MGH at the Brigham and a couple of other hospitals, spent somewhere on the order of two billion dollars installing the system. So that included all the customizations and all the training and all the administrative stuff that went with it. But that's a huge amount of money.

AUDIENCE: I agree.

PROFESSOR: OK, so we have past medical history-- pack a day smoker, abused cocaine but says she stopped six months ago, has asthma, type 2 diabetes. Social history, family history. These are of the physical exam results. So it's giving you a lot of information about the person. Description of the wound down at the bottom. Pertinent lab results. So these are copied out of the laboratory tables. Yeah.

AUDIENCE: Just to double check with the drug results--

PROFESSOR: Sorry?

AUDIENCE: Just to double check with the drug results two slides back--

PROFESSOR: Yeah

AUDIENCE: It said-- so it has the fake dates of 2190 up there.

PROFESSOR: Yep.

AUDIENCE: So the fact that there was a positive test in 2187 would mean a year ago.

PROFESSOR: Yeah.

AUDIENCE: So that's the medication.

PROFESSOR: Yeah, the deindenfication technology here maintains the relative dates but not the absolute dates. So these are results, again, copied out of the laboratory database into the discharge summary. Brief hospital course, and then a review of systems, so what's going on neurologically, cardiovascular, pulmonary, GI, GU, et cetera.

Infectious disease, endocrine, hematology, prophylaxis. And at the time was discharged, the patient was doing well, no fever and stable vital signs, tolerating a regular diet, ambulating, voiding without assistance, and pain was well controlled. Medications on admission, so this was the medication reconciliation. Discharge medication, so this is what she's being sent home on.

Discharge disposition is to the home with some follow up service, and she's going home. And the discharge diagnosis is infected left femoral popliteal bypass graft and the condition. And these are the instructions to the patient that say, you know, here's what you can do, here is when you should come back and tell us if something is going wrong, et cetera. And here's what you should report if it happens.

You know, if you have sudden severe bleeding or swelling, do this. Follow up with doctor somebody or other. Call his clinic at this number to schedule an appointment and then follow up with doctor somebody else in two weeks.

I think this is the same one. So just a couple of final words about standards. So you saw in David's introductory lecture a reference to Odyssey, which is a standard method of encoding the kind of data that we're talking about today. There is a likelihood that the next release of the MIMIC database will adopt the Odyssey formats rather than the-- yeah. David's shaking his head, wondering why. Me, too.

AUDIENCE: Odyssey hasn't handled clinical notes very well yet.

PROFESSOR: Well, so, you know, what always happens, as you say, I'm going to adopt the standard asterisk with the following extensions. And that's probably what's going to happen. But it means that the central tables, you know, the ICD-9 code tables and the drug tables, some things like that, are likely to wind up adopting the formats of the Odyssey database.

You should also know about this thing called FHIR, F-H-I-R, the Fast Health Interoperability

Resources. So HL7 is the standards organization that had a tremendous success in the early 1990s in solving the problem of how to allow laboratories to report lab data back to the hospitals or the clinics that ordered the labs. And that character string with the up arrows and the vertical bars and so on that I showed you before that had LOINK encoded in it is that standard. That's called HL7 Version 2.

It's still in use very widely, they then got ambitious and suffered second system syndrome, which is they decided to build HL7 Version 3, which I used to teach in a class here 10 years ago. But one of my friends who works for a company that helps hospitals implement that sent me a 38 megabyte PDF file that describes what you need to know in order to implement that system. And as a result, nobody was doing it.

So FHIR is a gross simplification of that that starts off and says, if a doctor refers a new patient to you, what is the minimum set of data that you need to know in order to take care of that person? And FHIR tries to provide just that subset of all of the data. It has become a standard mainly because, after Congress spent $42 billion dollars or so bribing people into buying these information systems, they got mad that the information systems they bought couldn't talk to each other.

And so they called in, on the carpet, the heads of these IT companies, health IT companies, and they yelled at them and they made them promise that there would be interoperability. They promised. And out of that came FHIR. It was probably simultaneously developed but they adopted it. And so now, in principle, it's possible to exchange data between different hospitals, at least to the level of that degree of harmonization of the data.

In reality, the companies don't want you to do that because they like there to be friction in not being able to take all your data to a different hospital, because it is more likely to leave you at the one that you're at. So there is complicated socioeconomic kinds of issues in all this. But at least the standard exists and is becoming more and more widely deployed as long as Congress pays attention.

It's ugly. So here is what a patient looks like, right? It's the usual unreadable XML garbage. But fortunately, there are parsers that can turn it into JSON and simpler representations. And so that's pretty common.

So the terminologies that exist are LOINK, NBC, ICD-9 and 10. SNOMED I didn't talk about today. DSM-5 is the Diagnostic and Statistical Manual for Psychiatrists. That's used as a common coding method for describing psychiatric disease. And there are many more of these.

There's something called the Unified Medical Language Systems Metathesaurus from the National Library of Medicine that integrates about 180 of these different terminologies. And so there is a nice one-stop shop where you can get all these things from them.

So takeaway lessons, know your data. Remember that first example of the heart rates, that comes up over and over again. And doing machine learning and analysis on data that you don't understand is likely to lead you to false conclusions.

Harmonization is difficult and time consuming. And there are lots of things for which we just don't have standards, and so everybody develops their own representations. I had a PhD student about a decade ago who, in his thesis, wrote that he spent about half his time cleaning data. And I gave that thesis to another student who started a few years later who read it, and he comes to me just awestruck and he says, what? He only spent half his time cleaning?

Unfortunately, that's roughly where we are in this field. So sorry to be a downer, but that's the current state of the art. And next time, David will start by looking at actually building some models with these kinds of data and showing you what we can accomplish. Thank you.