So it's a pleasure to introduce Jeremy Warner, who actually I first met this year when when I was chairing a panel at ESMO and he was one of the speakers who we invited. Jeremy is the director of the Brown Lifespan Center for Cancer Bioinformatics and data science and associate professor technically pending I guess, at Ed Brown. His clinical focus is hematology and he received his medical degree from Boston University and also in a Masters in
Photonics and Electrical and Computer Engineering from UC San Diego.

In addition to his focus on malignant hematology, Dr. Warner is a leading expert in the clinical and translational cancer informatics research, including high dimensional data analysis and visualization, natural language processing of narrative oncology texts, and the creation and implementation of health data standards.

Before coming to Brown, Jeremy was at Vanderbilt.
University Medical Center, where he was an associate professor of medicine and biomedical informatics. And I should also note that he is the deputy director of Escos Clinical Cancer Informatics Journal and a founding director of the New Brown University Center for Cancer Bioinformatics and Data Science. So without further ado, you're going to speak to us about using and improving real world ecosystem in cancer. Hey there. Ecosystem in cancer. Thanks. Look forward to it. Thank you. Thank you so much for having me.
And if anybody wants to come up to Providence, just one stop away on the Acela, so. Really nice that we're so close here.

In New England, so I just have a few disclosures first before I get started. So I have some grant funding, so I have some consulting.

I do have ownership in hemlock.org LLC, but has no monetary value unless one of you wants to be an Angel investor and we can talk after the presentation.

So what I'm going to talk about here is why do we need real world data and real world evidence in oncology and I'm going
to focus on electronic health records.

There are other sources of real world data of course, but you know most of this talk will really focus on the ER, the HR.

I will spend some time on talking about in particular interest in mine which is standardizing systemic anti cancer treatment representations and then I’ll spend some time talking about our COVID and COVID-19 and cancer consortium.

Which is a bit of a culmination, if you will, of some of these thoughts.
So there are some learning objectives here. If hopefully this is a CME, so we'll cover, you know, some aspects of natural language processing and how it can be used to get information out of EHR’s, why we need formal representations for complex concepts. Such as systemic anti cancer therapy and then learning about how these these ideas went and propelled the COVID, the Covid registry.

OK, so first of all, probably everybody might be already
familiar with these definitions,
but I think it’s always helpful to go over,
you know, what is real world data,
you know, it’s nebulous a little bit and depending on where you,
you’ll get a different definition.

So the idea here is that it’s a pyramid where you’re climbing a
The next step is knowledge.

The next step is wisdom.

You’ll know there’s 5 levels here.

There’s a little tiny level at the top which some people use you for understanding.
whereas real world data is really that base.

On the right here you see the sort of traditional pyramid of evidence based medicine.

So if you look at this from another dimension kind of looking from above, when we think about cancer in particular, you know I think about sort of three big aspects of cancer, there’s the genotype, the phenotype and the environment and sort of for each of these you have these layers.

So, so if you think about the data level for genotype that might just the somatic tumor sequence.
For phenotype, it might be just a histologic type, a cell, you know what, what is that? And for environment, it might be pollutant levels. Now this is data, but it’s not really telling you anything, right. So we need to kind of walk up this pyramid. The next level for these three buckets would be for genotype and environment. You might talk about pathogenicity. What does that change mean in terms of is it, is it a driver mutation? That’s sort of the next level information.
cancer behavior on the phenotype.

Side so is it, is it aggressive,

is it a high grade malignancy or is it something indolent kind of stepping up further?

For genotype knowledge, the knowledge level is actionability.

What can you do with this information? Can you actually prescribe a medication that will change the outcome for a patient?

Phenotype, same, you know, just generally speaking what are the treatment options and then the
environment are there risk modifications?

that can be taken and then really getting to that top level wisdom,

you know this is this is really complicated now.

So in in phenotype you're thinking about what are patient values and preferences and how do those influence what treatment options you might consider for genotype, what's the tumor going to do once it gets exposed to treatments,

how is it going to evolve under treatment pressure and an environment?

you've got issues about social justice.
And structural racism.

So those are all kind of like.

The ideas of climbing this pyramid alright,

Hopefully I’ve convinced you

the difference between data and

evidence as we kind of step up.

Now, why do we need this real-world evidence?

Well, clinical trials are wonderful,

but they’re also expensive,

slow to conduct, and they don’t

always represent the full population.

At risk also trials,

prospective trials,

collect some but not all potentially

pertinent information.

And our space is huge.
Oncology, the treatment space and oncology is huge. And then lastly, I think last but not least is that we’ve got this enormous data source, which is the electronic medical record. So just a few words about each of these items. So when you think about trials and disparities, this is a paper we just published very recently and this one was earlier this year and we just published another one in Jim Oncology looking at prostate cancer.
This one looks at immune checkpoint inhibitors across cancers. And basically the take home message here is that when you look across Childs, there is really a lot of disparity in who enrolls in trials. And it can be different by cancer type, but it’s pretty consistent across the board. And it’s not always underrepresentation. Sometimes it’s over representation, as you can see from the bottom row. But you know, essentially the yellow ones are intersections of, in this case, gender, age, race and ethnicity, and a cancer type where the enrollment is as you’d expect.
If it’s green, it’s sort of more than you’d expect, and if it’s red, it’s less than you’d expect. So you know this, this gets that generalizability and there might be statistical ways around this, but you know, essentially. Our knowledge from clinical trials is primarily coming from younger. White men, OK, so. How about the information that gets left out? So this is, this was amazing. This is the recovery group that really geared up during the early days of the COVID pandemic and found,
you know,

pragmatic trials that they ran in the UK and they found some really important treatment options for COVID. This is one of their papers. This is probably the most impactful look, showing that dexamethasone could help hospitalize patients with COVID. And I've excerpted a table from that paper. Take a minute and let.

So what's missing from this table? So this is a table of previous coexisting diseases in the patients who have COVID. Is there something missing? From this table.
Something that’s the topic of this talk.

Cancer, right.

There’s no cancer in the stable.

They did not collect cancer and or they didn’t report it.

Well, we actually, we actually went and you know got their case report forms, they didn’t record cancer.

So here they enrolled 10s of thousands of patients in these trials. And they don’t know if these patients had cancer or not. And so I mean, amazing work, but we’re missing a key piece of information.
that I met that I mentioned before is

that this idea that our treatment space

is huge but head-to-head comparisons

of important drugs are mostly absent.

And I’ll just give you one example.

So this is the space of PD1 inhibitors

which have changed our obviously changed

our fields from our hemac knowledge

base which I’ll talk about a bit later.

We have 137 trials that have been published

using 64 different regimens of various.

This includes XUS by the way.

If you’re like they’re not 13 you

want to have actually there are,

but many of those are only approved in China.
So 83 of those are phase three trials. Take home point is one of those 83 actually compared to PD1 inhibitor to a PDL 1 inhibitor kind of it actually compared to Kobe matanov and atezolizumab and that’s grand total of zero of these trials compared 1PD1 inhibitor to another PD1 inhibitor so. You know, maybe I’m missing some trials that are ongoing now that have yet to be published. But at this point in time, we don’t have any data at all on whether 1PD1 inhibitor is better than another except for indirect treatment comparisons,
so. Hopefully I’ve convinced you that.

We should at least think about using real-world data.

But. They are messy, ambiguous and unpredictable.

So let me talk about some some challenges that we have once we start delving into the real world.

So first of all. This is real-world data from the Medline. OK, so did you know that there were 21 clinical trial institutions in New Haven? Did you know that?
That's amazing, right? Here they are.

Smilow Cancer Center, Smilow Cancer Hospital, Smilow Cancer Hospital at Yale University, Yale Cancer Center and Smilow Cancer Hospital, Yale Medical school.

Alright, I think you get the idea, right?

So I mean this is real world.

I mean you have to do something.

I mean, a computer is not going to know,

right? I mean,

so if you want to use this data in some way,

someone’s got to do some work

to actually fix this, right?
That is a big part of working with real world data. Yale New Haven hospital.

There’s the 21st, OK? OK. So how about, so that’s bibliometrics to some degree, how about treatments, how many tyrosine kinase inhibitors are there? And so this is a little project that a student of mine undertook where they mapped out how many letters you’d have to switch around or basically misspell.

So that one tyrosine kinase inhibitor would actually be another one.

And so it’s it’s fewer letters
00:12:25.440 --> 00:12:27.085 than you think and.

00:12:27.085 --> 00:12:29.635 You know, these drugs get misspelled

00:12:29.635 --> 00:12:33.550 all the time in a pretty amazing ways.

00:12:34.410 --> 00:12:35.586 I see that the net,

00:12:35.586 --> 00:12:38.480 issue with the next slide, but.

00:12:38.480 --> 00:12:42.968 So this this is real data.

00:12:42.970 --> 00:12:45.114 From the Vanderbilt University

00:12:45.114 --> 00:12:46.186 Medical Center.

00:12:46.190 --> 00:12:49.234 So this is from our text list of medications.

00:12:49.234 --> 00:12:50.824 Now you might say, oh,

00:12:50.824 --> 00:12:51.920 let’s just, you know,

00:12:51.920 --> 00:12:53.248 we’ve got to be able to get these

00:12:53.248 --> 00:12:54.149 medications from structured data.

00:12:54.150 --> 00:12:55.949 That may or may not be true.
It depends. We can talk more about that.

But these are real misspellings of the drug or Latino BI.

Think you can tell looking at this that all of these are or lot in him.

But again, I mean if you don’t have some sort of system to harmonize all those misspellings.

You’re not going to know which patient got what drug.

So that’s you know, that’s a real world issue with real world data.

This is work that we did some years ago on staging,

so cancer staging.
Here is what I call manageable ambiguity. All right, so. And again and maybe you know during sort of discussion we can talk about the value of structured versus unstructured data, we're just trying to go for the big stages. And you know, the problem that we knew ahead of
time is that these things are going to be recorded variably in different notes by different types of doctors. But you know we did a pilot with about 1000 patients with lung cancer with over 460,000 clinical documents across them. Now if you pause for a minute and you think about chart review. Think about how long it would take you to go through 460,000 documents, right so. Here’s my pitch for natural language processing. You can actually automate this kind of thing and do this kind of work at scale. So cutting to the chase a little bit here.
First of all, we found that out of those 964 patients, 99% had some kind of stage freeze in their note. At least one. And we also had a gold standard which was the tumor registry data. So we were able to compare our system to the subset. You’ll notice only 790 out of those 964 had tumor registry data, but we were able to do a comparison and you know our system worked really pretty well. The green, you know, basically the matches are in the green, the big numbers and we got some things wrong.
but we didn’t usually get things really wrong most of the time.
So if it was stage one, we called stage four, that was a big mistake. Only happened once.
This, this shows actually, so again 460,000 documents. And you have to look at all those or can you just look right after a patient was diagnosed you know with if you think of some of this inspiration for this project came from the copi measures.
And if any of you have done that work you'll remember I believe and they may have changed but at one point the coping measure was recorded in one of the first two progress notes written after diagnosis. So it kind of makes sense that you would look for stage early on but if you look at this black line here at the bottom. So does this. Are you seeing my? You don’t see the arrow, are you? I don’t think you’re seeing the error, OK. If you look at the black line towards the bottom, you’ll see that.
If you look at the notes in the first five weeks from diagnosis, actually there’s a pretty high rate of unknown stage. Like we couldn’t determine it. So we saw this kind of inflection point. And so that’s another thing just to note when you’re working with real-world data is that, you know, time matters.
So I mentioned we found stage in 99% of the records. What I didn’t mention is that most of those are 84% had more than one stage in their records, OK and some degree of discordance. So one note might say they have stage one, another note might say they have stage two. Actually when we constructed a network graph on the right here you see like every possible combination was present, every possible combination including you know more terms that are more generic like early stage,
advanced stage.

Everything you know happens and you

know and and and on the bottom left

here you can see a histogram of of Co

occurrences of various stage information.

But I do think that so that

really potentially ambiguous.

One take home point from this

though is that we we use a

really simple decision rule on,

you know, what is the actual stage?

We just chose the phrase that showed up

the most OK and that and that seems to work.

So if stage three shows up

in the notes 100 times,

in stage one shows up twice.
Chances are at stage three now, just sort of a practical rule and it worked.

Now getting back to that, you know, the whole idea of.

You know unknown or sort of lack of information and missingness which is a major issue with real world data.

This is another mini project we did looking at colon cancer and we wanted to say could you find patients with stage 3 colon cancer and this was for the OCM project, the oncology care model.

So you know really important as a metric to know if these patients
got appropriate treatment within a proper period of time. But again what we saw here is this sort of crossover at about seven weeks, at which point? You know, the stage was changing or it was or is missing in the records and it wasn’t until about seven weeks after diagnosis that you get to a kind of steady state where you can definitively say a patient has stage three or we don’t know the stage so.

Here’s some really interesting work from here, actually from Yale, from the Radiation Oncology department, where they they actually looked at
missingness as a variable, if you will. So they took the National Cancer database, the NCDB data, and they split patients into whether they had complete records or had some missing data from their record. Now the NCDB is not EHR data, right? But it is based on EHR data. And you know the punch line here is that missing this is an independent prognostic factor for survival which is really an interesting thing to think about, right. And and it kind of depends on what kind
of what kind of cancer you have as well. So they found for instance on the left. If you have non small cell lung cancer, it’s the non metastatic patient who had a real difference in their prognosis if they were missing data. Whereas with prostate cancer it was the metastatic group that sort of split apart. But either way, I mean this is. Whereas with prostate cancer it was the metastatic group that sort of split apart. But either way, I mean this is. It’s certainly not something that we conventionally use as a metric. Certainly not in a clinical trial.
because there's it's not an issue, right?

Case report forms are complete, but missing this itself can be informative as in real world data.

So what I wanted to do now is actually take us down a little different path briefly, which is a brief diversion into the history of medical records. Anybody know what this is? It's a local local. It's kind of cool if you never been there. It's still there. Doesn’t exactly look like this anymore, but you’ll see why I’m showing.
this is also a real thing.

OK, so this is one of my favorite vehicles from the Lane Motor Museum in Nashville. Which is Doctor Weiner mentioned, I was there for about a decade and so this is a real vehicle. There's they actually have a collection of these and it makes me think of electronic medical records because it it works, right? It it actually this person’s actually driving this car. But we don’t exactly see propeller driven cars on the roads these days, right? So our ER, but it works.
So MR’s are functional, but are they fit for the purpose that we want to use them for?

I think many of us have some ideas about that, but when you think about medical records, this is obviously a little bit before the computer, you know, medical records have been around for. Almost, you know, 3500 years in one form or another. What’s interesting to me? Is that they were primarily used for teaching or didactics. Until very recently,
that was the only purpose of medical records.

And then sort of the second purpose that arose, if you will, didn’t arise until the 1880s. It’s not that long ago if you think about it.

And, you know, essentially to have a written record of what happens in case there was a lawsuit around medical malpractice. And we’ll skip that and sorry.

So it’s a little bit hard to read some of this,
but you know how about billing that that’s billing is the major driver rate of how our medical records look like today. But that only really happened in 1960s is really not long ago and until you know not so long ago physicians were paid with food and lodging. If they were lucky. This is a picture from the Confucian medical system where there’s at least some cases where the the court physician was basically executed if the if the emperor did not get better. So that’s a pretty harsh payment
or penalty if you will.

But you know what really changed things was the Medicare Act of 1965, which basically established this profile. You know, quote usual customary and reasonable fees which. Drive so much of what we do.

That's actual quote.
It certainly changed things a lot. And then what I’d argue also changed was really more recent was in the 90s when the physician fee schedule was introduced and then something called the evaluation and management guidelines, which I think. A lot of us know more than we ever wanted to know about, but those really changed how medical records were written. Noticed that haven’t yet used the word electronic, right? So now what about patient care, which I think all of us want that to
be the primary purpose of medical records.

This kind of dates back to the 1800s in some ways.

The case records of the Massachusetts General Hospital introduced some ideas like history of presenting illness, past medical history and so forth, medical record numbers. The whole idea that you would track a patient by a number was introduced at the Mayo Clinic in the early 1900s, where they also introduced the chief complaint and the review of systems. And then the American College of Surgeons,
this is amazing bit of history
if you didn’t know in 1918, they.
There was no federal mandate of any.
They basically mandated as
They mandated that hospitals had to
They mandated that hospitals had to keep records including a discharge
summary that basically said was the,
you know, patient,
you know alive or dead at the time they left.
And at that time fewer than 20% of
physicians kept any kind of record at all,
which is like. Amazing, right?
Now this is tying back to that Eli Whitney.
So this is, you know,
for those that did take records, this is kind of what they looked like as these are called case books. I'm not sure where this one is from, but it’s basically a handwritten. This is that it’s physician centered, right. This is not, this was written as. The doctor saw patients, so if you ever wanted to go back and say OK, Mr. Smith or whoever, like put their case together, good luck.
So really the most recent innovation if you will in medical records was this one and that from Austin from the mid 1960s which is the problem oriented medical record which which was conceived as a quote medical record that guides and teaches. So kind of back to that idea of didactics in a way and I'm sure everybody's familiar with this, this idea, this soap notes, right. What I like from the paper when doctor we'd introduced this idea. This is a quote which I think actually forecasts the ER right so,
It's worth reading it. It can be readily seen that all narrative data presently in the medical record can be structured, and in the future all narrative data may be entered through a series of displays guaranteeing a thoroughness, retrievability, efficiency and economy important to the scientific analysis of a type of datum that has hitherto, been handled in a very unrigorous manner. It's an amazing quote. This is essentially before any
But he basically saw it, saw it coming.

I think the most important part of this quote is this to be concluded.

We’re living through the evolution of these electronic medical records.

This is actually a two-part paper, that’s why it says this.

But I think, you know,

he could have been like OK,

we don’t know what’s going to happen.

It’s worth taking a step back

and saying what you know what.

So now I’m going to say electron what

is the electronic health record for?
And it's got primary uses and secondary uses. The primary uses are patient care and delivery, financial billing. But it’s this, when you talk about real-world data and real-world evidence, that’s a secondary use, as it’s conceived here in this model, which the Institute of Medicine put forward. At that time there were issues around funding to rollout electronic medical records and. What I like on the bottom here is in...
2003 the mass medical society did a survey where 89% of physicians wanted EHR data, but 48% refused to use an ER. So little bit of a disconnect there and by 2004 hardly anybody was using medical records. So what changed? Arguably this is, you know, one of the events that really changed things. Is everybody familiar with Katrina and what happened in New Orleans? Does everybody know why the? Record so there’s a picture there on the right.
00:28:24.120 --> 00:28:25.880 were in the basement.
NOTE Confidence: 0.878462154285714
NOTE Confidence: 0.878462154285714
00:28:26.580 --> 00:28:29.380 It’s the they are so heavy that the
NOTE Confidence: 0.878462154285714
00:28:29.451 --> 00:28:31.701 building literally would have collapsed
NOTE Confidence: 0.878462154285714
00:28:31.701 --> 00:28:34.641 under the weight of the paper if
NOTE Confidence: 0.878462154285714
00:28:34.641 --> 00:28:36.566 they’d been up on higher floors.
NOTE Confidence: 0.878462154285714
00:28:36.566 --> 00:28:38.624 So that’s why they have their
NOTE Confidence: 0.878462154285714
00:28:38.624 --> 00:28:40.592 medical records in the basement and
NOTE Confidence: 0.878462154285714
00:28:40.592 --> 00:28:42.320 and they were all destroyed, right?
NOTE Confidence: 0.878462154285714
00:28:42.320 --> 00:28:43.320 They were all just lost.
NOTE Confidence: 0.878462154285714
00:28:43.320 --> 00:28:45.498 So, so fast.
NOTE Confidence: 0.878462154285714
00:28:45.498 --> 00:28:48.078 We’re a little bit the High Tech Act in 2009,
NOTE Confidence: 0.878462154285714
00:28:48.078 --> 00:28:51.422 which Obama signed this this is what really.
NOTE Confidence: 0.878462154285714
00:28:51.430 --> 00:28:52.004 You know,
NOTE Confidence: 0.878462154285714
00:28:52.004 --> 00:28:54.013 gave a lot of money for institutions
NOTE Confidence: 0.878462154285714
00:28:54.013 --> 00:28:56.111 to really start putting any Mrs.
but what is interesting is if you look at sort of the adoption curve and there’s a couple, I won’t get into the details here. There’s a couple ways of like what is an EHR basic versus complete and so forth, but you actually see them starting, so here’s E&M coming out in the mid 90s. Here’s Katrina in 2005. There’s the High Tech act. By the time the High Tech Act comes out, actually we’re like well on the adoption curve and so, you know, definitely help things along, but you know,
the process is already starting.

And then you know where.

So this is already five years old,

but I think, you know it's it's.

So, you know, it's problematic.

But you know already by

five years ago people were

reporting that EHR's were

a major driver of burnout.

So, so you know, it's problematic.

But OK, here's a here's a

few other challenges. So.
And I’m sure everybody who’s clinical knows these things already. But carry forward a copy pasting is ubiquitous in medical records and there’s just a ton of redundancy. Here’s a paper that basically shows that. Umm. You know, large, large portions of any note particularly look at have been copied forward from previous notes. More than half of progress note material is copied forward from previous notes.
00:30:30.614 --> 00:30:32.566 know how many progress notes have a manually
NOTE Confidence: 0.861171268571429
00:30:32.566 --> 00:30:34.207 entered text versus copied in any kind.
NOTE Confidence: 0.861171268571429
00:30:34.210 --> 00:30:37.720 And you can see again like very few progress
NOTE Confidence: 0.861171268571429
00:30:37.720 --> 00:30:40.448 notes have have fully written text.
NOTE Confidence: 0.861171268571429
00:30:40.450 --> 00:30:43.510 Which you would say is fully original, but.
NOTE Confidence: 0.861171268571429
00:30:43.510 --> 00:30:44.973 So I think it’s a legitimate question
NOTE Confidence: 0.861171268571429
00:30:44.973 --> 00:30:46.708 to say what are we dealing with here?
NOTE Confidence: 0.861171268571429
00:30:46.710 --> 00:30:47.851 Is it a giant pile of paper
NOTE Confidence: 0.861171268571429
00:30:47.851 --> 00:30:49.099 or is there actually meaning.
NOTE Confidence: 0.861171268571429
00:30:49.100 --> 00:30:51.500 So this is a little little tiny project
NOTE Confidence: 0.861171268571429
00:30:51.500 --> 00:30:54.032 I did and when during fellowship where
NOTE Confidence: 0.861171268571429
00:30:54.032 --> 00:30:56.699 I basically took one of my patients
NOTE Confidence: 0.861171268571429
00:30:56.700 --> 00:30:59.150 charts and I counted up like how
NOTE Confidence: 0.861171268571429
00:30:59.150 --> 00:31:01.569 many data points are in that chart.
NOTE Confidence: 0.861171268571429
00:31:01.570 --> 00:31:04.810 And you can see the blue bars are all the
NOTE Confidence: 0.861171268571429
structured data elements like billing
codes or vital signs or lab values.
And then these red bars are the words in the clinical documents and you see that that just drowns out right, the structured data.
So there’s a lot of data there but. It’s awesome.
There’s even more than that, right?
So this was more than 10 years ago, there was another 277 pages of scanned documents with 69,000 words in them that were basically inaccessible,
but and the take home point here.
Is that this is what it all boils down to,
OK?
Patient with diffuse large B cell lymphoma.
It was a complete remission after
getting 6 cycles of our chop.
I think that’s enough for most research.
OK now how can we,
how can we boil things down like
that because that’s kind of
maybe what we’re talking about here.
So and of course there’s more to it right.
But you know when you think about what’s
in ER’s or EHR’s and and what is not.
Umm.
You have to know what you’re,
you have to know what you’re going to find, right. So, let’s say you know you’ve unlocked this medical record, but it’s not necessarily going to have what you want. Basically some big buckets, right. So you’re going to find the person’s date of birth, no problem, right. But you’re not going to find probably where they were born, the circumstances of their birth, where their complications. Very unlikely, because they won’t have lived.
their whole system with their life within the electronic air, and they won’t have all that data. You might find their biologic sex, no problem, but are you going to find their gender orientation, that sexual identity? You’ll find race in this city, but are you going to find other social determinants of health? You’ll find the medications that they are prescribed, but will you find what they actually took, the medication that they took and the
00:33:03.529 --> 00:33:05.432 regimens and we’re going to get into
NOTE Confidence: 0.861171268571429
00:33:05.432 --> 00:33:07.562 that next you’ll find laboratory tests,
NOTE Confidence: 0.861171268571429
00:33:07.562 --> 00:33:10.040 but you want necessarily find images.
NOTE Confidence: 0.861171268571429
00:33:10.040 --> 00:33:11.528 And so forth.
NOTE Confidence: 0.861171268571429
00:33:11.528 --> 00:33:15.000 So kind of you know as moving
NOTE Confidence: 0.898628412727273
00:33:15.117 --> 00:33:17.040 forward. Thinking about what you know,
NOTE Confidence: 0.898628412727273
00:33:17.040 --> 00:33:19.760 the low hanging fruit.
NOTE Confidence: 0.898628412727273
00:33:19.760 --> 00:33:21.848 You know it’s the cancer type,
NOTE Confidence: 0.898628412727273
00:33:21.850 --> 00:33:23.482 it’s easy like we don’t need to create
NOTE Confidence: 0.898628412727273
00:33:23.482 --> 00:33:25.147 a new system to get cancer type.
NOTE Confidence: 0.898628412727273
00:33:25.150 --> 00:33:27.376 You can get that from billing codes,
NOTE Confidence: 0.898628412727273
00:33:27.380 --> 00:33:28.426 registry data.
NOTE Confidence: 0.898628412727273
00:33:28.426 --> 00:33:32.087 The treatments are hard like our chop
NOTE Confidence: 0.898628412727273
00:33:32.087 --> 00:33:34.452 times 6, that’s hard determining that
NOTE Confidence: 0.898628412727273
00:33:34.452 --> 00:33:36.810 the patients in a complete remission,
NOTE Confidence: 0.898628412727273
00:33:36.810 --> 00:33:37.479 that’s really hard.
So what I go for the middle, I don’t go for the middle ground, right, I’m going to tackle the thing in the middle. So now I’m going to switch gears here for a bit and talk about our work on standardizing systemic anti cancer treatment. And before I get into that, If you’ve not seen this XKCD cartoon, it’s a classic. And this is a challenge, right? Whenever you decide to create a new standard or you actually just you know. Just creating more complexity or not. Hopefully we’re not. Well, what we did in this space,
there really weren’t 14 existing standards.

There were none.

And so as everybody here knows,

Chemotherapy regimens are complicated and given in cyclic fashion combinations.

This was the standard when we got started on our work.

This is, you know, one example of these things called cancer chemotherapy handbooks, kind of recipe books, physical books, right.

With some some details here, but maybe not enough.

Here’s another example from 2005.
Which if you kind of look in detail about what’s there. There’s a lot of optionality here, some of the references. Here’s a little excerpt from the Adenoma. I don’t know carcinoma of unknown primary section, but the references are to non small cell lung cancer so there’s sort of a mismatch there in the evidence base. So what we did is we. Really basically tried to collect all this information and put it into a computable format, which is our hemlock.org website and
the ontology that comes from it.

So he might.org is a website with the goal to collect all standard of care systemic anti cancer treatment. That’s the goal. It’s a big goal and at the website has grown over more than a decade now. Of almost 1000 primary content pages, over 7000 references, and a large editorial board, actually members of which are from Yale. And and many page views, so 1.4 million page views last year, we do get visitors from all over the world were primarily US based. I always like to throw in that we’ve
00:35:57.510 --> 00:35:59.590 had one visitor from North Korea.

00:35:59.590 --> 00:36:00.808 I don’t know who it is,

00:36:00.810 --> 00:36:03.426 but I don’t think I want to know.

00:36:03.430 --> 00:36:05.806 So what what can we do with this website?

00:36:05.810 --> 00:36:08.330 So what we did over time,

00:36:08.330 --> 00:36:10.360 over the past 11 years is create

00:36:10.360 --> 00:36:12.231 a structure such that we could

00:36:12.231 --> 00:36:13.796 actually take the content and

00:36:13.796 --> 00:36:15.260 develop a formal model.

00:36:15.260 --> 00:36:17.120 And so this is the model?

00:36:17.120 --> 00:36:19.507 Or this is part of the model?

00:36:19.510 --> 00:36:21.070 And I don’t have time right

00:36:21.070 --> 00:36:22.110 now obviously to kind

00:36:22.169 --> 00:36:23.867 of go through all these details,

00:36:23.870 --> 00:36:27.188 but it’s somewhat complex and enlarge

67
we have over 100,000 concepts and over 300,000 ways in which those are interrelated in the latest version. This is showing that each regimen on the website is in such a way that we can take all those pieces and put them into the data model. And then we can start to do cool things with real world data.
with some folks in South Korea who basically had access to.

Essentially medication level database.

And remember I mentioned you know way back when that we might get medications, but to actually understand regimens we have to do something extra.

And So what they did is they applied our model and they mapped medications through regimens and they were able to look basically over a decade of time 2008 to 18.

And you can see here that you know the changing pattern of care in that country.

So you see that for example of you know
bevacizumab wasn’t used really until 2014 and then it started getting popular.
And by the year 2018, it’s, you know, full F ox and Bevacizumab Kappa,
a good chunk of of the treatment regimens,
whereas something like fluorouracil monotherapy essentially disappears
off the scene by the time you get there.
This is much more recent so that
now we’re working with folks at the University of California System have a really cool combined database across all the UC’s and California is kind of a.
You know, country unto itself, once you start putting all this data together, this is just from UCSF and again we’re taking medication exposure data including time stamps and we’re mapping that to regimens. And and you see that. At least nowadays, full fernox is the most popular regimen there. And so that’s that alone is an interesting thing, right? You also see some funny things, right? Like so I didn’t know Leuprolide was a treatment for pancreatic cancer, did you?
But these are real patients, right.
And they actually have second malignancies.
So these are people who have also have prostate cancer and they’re also getting leuprolide.
So you, you kind of have to you know, it’s not enough to get that data out.
You’ve got to, you’ve got to determine, does it make sense, is it? Is it relevant and and?
And so that that’s why we’re seeing things like that so.
Here’s another.

Gives you a taste of what we can look at.

So that is looking at folfirinox and then looking at cycle by cycle.

So one of my long-term interests, as well as Doctor Zach here, is to understand treatment delays, dose reductions, removals of medications from a regimen drop, you know, dropping a drug and this starts to get at that and you can kind of see, each of those bars represents cycle to cycle, the cycle you see.
00:39:54.200 --> 00:39:54.680 People.
NOTE Confidence: 0.8194428005
00:39:54.680 --> 00:39:56.032 People dropping out, right.
NOTE Confidence: 0.8194428005
00:39:56.032 --> 00:40:01.639 And and so and then you can actually see
NOTE Confidence: 0.8194428005
00:39:58.982 --> 00:40:01.639 why and you can see on the top here.
NOTE Confidence: 0.8194428005
00:40:01.640 --> 00:40:03.536 These, these bands at the top are showing.
NOTE Confidence: 0.8194428005
00:40:03.540 --> 00:40:05.220 You know, these these are folks.
NOTE Confidence: 0.8194428005
00:40:05.220 --> 00:40:06.318 You don’t think that you have
NOTE Confidence: 0.8194428005
00:40:06.318 --> 00:40:07.050 a pointer or something.
NOTE Confidence: 0.8194428005
00:40:07.050 --> 00:40:10.235 Oh, actually, let’s see if this will.
NOTE Confidence: 0.8194428005
00:40:10.240 --> 00:40:10.510 Yeah.
NOTE Confidence: 0.8194428005
00:40:10.510 --> 00:40:12.670 So,
NOTE Confidence: 0.8194428005
00:40:12.670 --> 00:40:16.927 so you see these bands coming across,
NOTE Confidence: 0.8194428005
00:40:14.662 --> 00:40:14.662 those are basically patients that are
NOTE Confidence: 0.8194428005
00:40:14.662 --> 00:40:16.927 progressing and going on to a second
NOTE Confidence: 0.8194428005
00:40:16.927 --> 00:40:18.679 line treatment what’s not shown here.
NOTE Confidence: 0.8194428005
00:40:18.680 --> 00:40:20.198 Just to spare you a little
bit on the visual side, our patients who are or stopping therapy and essentially transitioning to Hospice or some sort of end of life care and that’s this big bar here. And then some patients these little little ones they’re going to a deescalated regimen. So they’re dropping the Arena taken or the oxaliplatin and so you can really start to see these patterns of care in the real world data so. This is my little advertisement for Humalog.
It’s available to you. You can download the whole thing and mess around with it if you’re an academic or non-commercial user and just Google Hemac dataverse and you’ll find it. Or you can use these links. It’s also available through something called the Odyssey Athena vocabulary. And yeah, we want more users. There’s a lot more that can be done with it. So along comes a pandemic. So, now I want to spend the last little bit here talking about the COVID-19 and cancer consortium.
Which yells a member and this is our mission statement, which has been the same since we were created in March 2020, which is our goal is to collect and disseminate prospective, granular, uniformly organized information on people with cancer who are diagnosed with COVID-19 at scale and as rapidly as possible. But what I want to talk about here for a minute is sort of what I call the ancillary goals of C19 or the unwritten goals.
can we build a consortium, can we build an airplane while also flying? Just, you know, can we do it? That was the question. Convening a group of stakeholders was really in, you know, a goal including patients, really engaging patients and then. Pertinent to the talk today, we knew that there were other efforts kind of getting rolling that were based on what was in that structured data. If you remember that’s the.
The tiny little blue bars right on the graph I showed you. So we wanted to, you know, get more than that. So this is. This is back in back in Rhode Island. Alright. Showed you Eli Whitney earlier. This is the this is Slater Mill in Pawtucket which I think I pronounced correctly but I'm getting my New England shops. And what’s interesting to me about this story is that he earned this name, Samuel Shredder Slater and the reason he was branded as a traitor.
is that he was accused of stealing
the ideas for industrialization.
From England where he was born and grew up,
and then replicating it in America.
So this is really the beginning of
the American Industrial Revolution.
But what’s interesting about that is
that he didn’t exactly steal the ideas.
Like he didn’t steal blueprints
or things like that.
He just like memorized them and
brought the knowledge with him.
So it’s, you know, that’s what he did.
So. I think that that’s great actually.
And so you know when we think
00:43:47.719 --> 00:43:49.493 about C 19 and I certainly don’t
00:43:49.493 --> 00:43:51.420 have time to go through all this,
00:43:51.420 --> 00:43:55.158 but we have many inspirations he,
00:43:55.160 --> 00:43:55.672 the hemlock,
00:43:55.672 --> 00:43:58.060 what I just spoke about is one of them.
00:43:58.060 --> 00:44:01.300 But in all the domains of C19,
00:44:01.300 --> 00:44:04.000 we are borrowing best ideas,
00:44:04.000 --> 00:44:06.536 modifying sometimes and putting
00:44:06.536 --> 00:44:09.012 together this consortium and and
00:44:09.012 --> 00:44:11.760 this is just sort of a list of that.
00:44:11.760 --> 00:44:13.140 The other thing I wanted to
00:44:13.140 --> 00:44:14.060 say about you know
00:44:14.110 --> 00:44:15.466 the C 19. Work and just,
00:44:15.466 --> 00:44:17.501 you know, research in general
00:44:17.501 --> 00:44:19.886 is that sponsors are critical.
00:44:19.886 --> 00:44:21.230
In a Samuel Slater’s case, he had a sponsor named Moses Brown who basically fronted him the money to build those mills. And our sponsor is Julie Klem at the NCI who didn’t front us any money but was very supportive and helped us kind of, you know, surface and socialize our ideas. This is our data schema, and what I want to emphasize here related to this talk is that everything in red is not available in structured data, so as we sort of built this up. You know, some of these things you can collect,
you know, in many different ways.

But the red items. And you’ll see in a few slides that those turn out to be critical things like ECOG performance status, toxicity of cancer treatment pneumonitis. Items like that, that we really wanted to zero in on.

We’ve done really pretty well on capturing what I would call elusive variables. So these are kind of the
things that they’re in the ER, but they’re in that unstructured. Leak of data, but we we got a lot of them. So cancer status is the patient. Getting better, getting worse or staying this, you know the same as before, a stable disease. We have that in over 95% of the patients. Even smoking status is hard to get right. We have that. Did COVID affects the patients treatment plants that’s not going to be unstructured data necessarily. We have over 90% on that on the ECOG
which is a notorious difficult thing to get and all the various efforts such as flat iron and so forth have had had challenging and cancer link have had challenges with this. We have ECOG data on 88% although that does includes patients who just didn’t have any ECOG recorded but we that knowledge of no ECOG is still. Knowledge, right? And you know getting to our results again in
just focus on the red and what
we found is that these factors,
these elusive factors are really important.
And so this is unadjusted
just kind of descriptive.
If you had progressing cancer at
baseline you get COVID your 30
day mortality is 26% and if you
had an ECOG of two or higher your
mortality is extremely high.
And we also found that immunosuppression
which is a somewhat nebulous
definition here which is complex,
which includes a lot of things you
can’t easily get out of structured data.
So this is sort of the real world data is a huge driver of mortality.

And if you look at the right, the yellow table basically those are the patients who are immunosuppressed at baseline. And across the board, even younger patients have substantial mortality in our data set. Furthermore, if you add on top of that active cancer, So are they immunosuppressed and they have active cancer. Again, we have our definition for that.
Because if you’re not immunosuppressed and you have inactive cancer, in our data, at least you have a zero chance of dying in the 30 days, whereas if you’re older immunosuppressed, your chance goes all the way up to 30%. So really a huge spread here based on these data. And then if we start to look at. Multivariable adjusted analysis. Again, we see that these factors like ECOG or cancer status are highly associated with outcome, both mortality as well as severity, which means hospitalization, intubation and so forth.
We saw this as well more recently when we looked at vaccinated patients. So patients who are getting breakthrough COVID-19 after vaccine again we saw things like cancer status really you know being a, you know huge adjusted odds ratio there of six if you had an active and progressing cancer of of dying in 30 days. So I could talk about COVID-19 itself for an hour, but I’m going to pause and so I just want to share some parting thoughts. So, first of all, I think I’m, I’m, I’m, you know, I’m, I’m,
I'm a believer here that real world data has a great potential to yield real-world evidence if we approach it with an understanding about the completeness issues, the accuracy issues, and we anticipate them and we come up with either ways to adjust for them or or avoid certain data, certain variables in the first place. We need automated methods, right? Yes, we need automated methods, right? Like, it wouldn’t be great if NLP and a computer could do everything, but in reality a lot of real world data and real world evidence
depends on human curators going into EHR’s pulling out that data.

And to do that we need rigorous approaches.

We have a paper published earlier this year describing the approach we used in ACR Genie.

I encourage you to check that out.

It basically gets into.

You need two people to independently curate the same record at a certain rate so you can see if there’s comparability between their results.

And so forth.
If there’s widespread adoption of standards such as M code, hemac, omop and so forth, that will increase the usefulness of structured data marginally. I think NLP is having a moment. If you pick up the newspaper nowadays, like you’re going to see our other paper on chat, GPT, for example, which is generative NLP but sort of the other side of NLP, and then Umm. You know, really important though, and I didn’t get to touch on this at all except for the very beginning when I alluded to disparities and bias.
There’s a lot of concern that working with real-world data might actually make biases worse that are already present in that data. So we need new approaches to deal with that issue. New approaches to deal with that issue. Just have some acknowledgements here. So there’s two slides here.

I acknowledge the himanka.org editorial board. Others that have worked on it are funding and and Dolly.
which is the creator of some of those graphics you saw.
And here’s our acknowledgement for the C19, which is a huge endeavor that has more than 600 active investigators at this point.
And with that, I will pause for questions.
So I’ll start. And can you see the ones that are online or not?
I do see, yes, yes, I see.
I don’t for a second dispute the value of real world data in terms of being able to answer questions, but I’m struck by the fact that we have these two extremes we have.
we spend a fortune to collect every last bit of data and you know they cost $15,000 per patient or more. And we get lots of useless data as part of it. And then we then say, well, we can’t do get everything. So then we go to real world data where everything’s pretty messy and you have to make all these assumptions and clean up the data. And the the question is, is there a role for much simpler randomized trials done as part of standard practice?
I mean sure, yeah, I think the recovery trial, they showed that you can do these huge pragmatic trials in 10s of thousands of patients with off the shelf drugs, not a lot of money. Some of the drugs we won’t, we won’t say the words but you know and things like oxygen right. But when you get into the you know the expensive drugs that are not yet FDA approved, I think that’s a whole other area. I think that FDA has got to lead.
the way in some ways here because they and I didn’t get to talk about this, but you know, there’s a high profile rejection of real-world data within the last month or two. And I think they rightly looked at that set of data is not trustworthy and we’re not going to go for it. But I don’t think that that should shut down the whole endeavor.
I think that they need, we need guidance from them and you know about what components should and should not be, you know, collected routinely. I think that might simplify things a lot. Attempt to put together criteria that would allow you to say that this set of real world data is adequate to Brock inclusions from you know, in terms of how much it has to be cleaned up, how large the sample size has to be. It’s such an interesting question and I’m not aware of anything at this moment but I do you
NOTE Confidence: 0.90763503
00:54:02.934 --> 00:54:05.178 know we are there’s this great bias
NOTE Confidence: 0.90763503
00:54:05.178 --> 00:54:07.684 that I just learned about called the
NOTE Confidence: 0.90763503
00:54:07.684 --> 00:54:09.654 informed presence bias which I kind of
NOTE Confidence: 0.90763503
00:54:09.654 --> 00:54:12.197 knew I knew it but not by those words
NOTE Confidence: 0.90763503
00:54:12.197 --> 00:54:14.108 but that basically means that patients
NOTE Confidence: 0.90763503
00:54:14.108 --> 00:54:16.700 who spend a lot of time in the clinic
NOTE Confidence: 0.90763503
00:54:16.770 --> 00:54:19.362 or the medical system have a lot of data
NOTE Confidence: 0.90763503
00:54:19.362 --> 00:54:21.319 right whereas those that don’t don’t
NOTE Confidence: 0.90763503
00:54:21.319 --> 00:54:23.940 and and and it’s and it’s actually
NOTE Confidence: 0.90763503
00:54:23.940 --> 00:54:25.640 an incredibly important source of.
NOTE Confidence: 0.90763503
00:54:25.640 --> 00:54:30.050 The bias? That. That.
NOTE Confidence: 0.90763503
00:54:30.050 --> 00:54:31.162 You know, can you?
NOTE Confidence: 0.90763503
00:54:31.162 --> 00:54:33.231 So if a patient doesn’t spend enough
NOTE Confidence: 0.90763503
00:54:33.231 --> 00:54:35.265 time to get enough data generated,
NOTE Confidence: 0.90763503
00:54:35.270 --> 00:54:36.600 that’s something we should know.
NOTE Confidence: 0.90763503
That’s something we need to know, right?

But that’s almost that kind of, you know, descriptor is almost never available in any real world data study to my knowledge, so.

The online version.

Yea.

Schedule.

What is?

COVID-19.

Yeah. Yeah. So the question is, it seems to be the case that the patients with the pre-existing cancer having worse outcomes during the COVID era than before and why might that be? I can say from our consortium now we
only look at patients who had COVID.

So that’s a subset, right?

Well, as time goes on,

it’s going to be everybody maybe.

But what we do see is that you know at

least in our registry 40% of patients

have their treatment altered in

some way and usually that’s a delay.

But sometimes they can’t get the

same treatment that they were getting

before a surgery gets cancelled,

you know etcetera, etcetera.

And and we know from you know previous work,

obviously the treatment delays don’t

usually ever. Work out very well.
So we haven’t yet systematically evaluated that, but we have now several thousands of those patients. So we’re going to be looking at that probably in the upcoming year. As far as other patients, well, there were a lot of practice changes, right, especially in China, I think, but also with sort of substituting oral medications whenever possible, even if they were sort of known to be inferior or not, you know, not quite as good so that patients didn’t have to come into.
To the clinic.

So that’s been presented on in some settings, but you know I think what we think that those substitutions are generally OK, I know that. You know a lot of people went on neoadjuvant hormone therapy and instead of going direct to surgery for early stage breast cancer and you know so that they could push this you know during periods of time when elective surgeries were shut down. So all those things probably add up right.
But there’s absolutely a factor of psychology and patients being afraid to come into the clinic and you know potentially again skipping a treatment or. So, to answer your question is that it’s quite complex but I think we need to understand it better and of course new diagnosis coming in which we’re starting to get that information. There’s clearly a stage migration and and you know to later stage more more advanced, more metastatic disease. Because of delays in screening and so forth.
So. So I think we're going to face a challenging decade and I think Ned Sharpless forecast that at the very beginning of the pandemic. I think in the first month or two he wrote a paper and nature of science modeling out what that might look like and and that's probably going to come true but. Hopefully COVID ends really soon. So. Um, yeah.
Yeah so we're so we're overtime and and I think you know I mean there's there's many strategies to try to mitigate but you can't you can't eliminate bias right. So you can understand it you can try to mitigate it there's you know matching strategies to if you're doing sort of a case and control style approach where you try to make the controls as similar to the cases you know and everything but the closure. So you know and and and some of those are, some of those been around for decades, some of those are kind of emerging at this point.
But I don’t think we can forget that there’s bias in perspective trials as well, right.

So I mean I think either side of the coin.

Yeah, it’s just, it’s just, it’s just one more thing and it’s not the only, I mean it’s there’s also ascertainment, but I mean there’s a lot of biases, right and. You know, one thing we’ve worked on with our consortium is developing standardized language around limitations, which I think is critical because you know. The data is the data are the plural, right? But.
But the way it’s presented really does influence the reader, right?

So. So that’s something we’re thinking about and might have some. You know, thought pieces or something coming out about how to handle that.