CHoRUS Bridge2AI Standards Module Office Hours continued-20240502_180022-Meeting Recording

May 2, 2024, 4:55PM

1h 4m 31s

Alvarez, Marta 4:28
Hi there.

Williams, Andrew E 4:30
Hey.

Schmidt, Heidi E 4:36
Hello.

Williams, Andrew E 4:38
Hey, Heidi.

Alvarez, Marta 6:13
You know, there's still people coming on, but we might want to give, I think, more Teslas on the call and he's not been on these calls as far as I know before.
And I know he was invited specifically, so hopefully we can give him a quick minute to introduce himself when more people are on the call.

Williams, Andrew E 6:35
Yes.
Great idea.
So I know.
Eric was also invited. Think.

Alvarez, Marta 6:50
I put a list of who was invited in the I don't.

Williams, Andrew E 6:57
Umm.

Alvarez, Marta 7:00
I don't have any confirmation. Umm.
Except for. I believe, well, one person accepted it on their calendar, but it wasn't Eric.

Williams, Andrew E 7:14
Wondering if I did or more Tessa.
Knows whether Eric is going to be able to join.

Schmidt, Heidi E 7:24
I can take a quick look at the calendar.
Hold on.

Alvarez, Marta 7:31
Thank you.

Schmidt, Heidi E 7:58
In his calendar, he has.
Umm.
Another meeting.
It's a a chair call.
So I don't think you'll be.
I'll be able to attend here and it looks like it goes on for about an hour and a half.

Williams, Andrew E 8:19
OK.
The reason to be particularly interested in in Eric and I think we we're going to get a lot out of this meeting.
I was very exciting.
It's going to be sort of a one of the very first times we've been able to look at what has been uploaded and looked at it using some of the tooling that's been designed to inform both data quality as well as I think the presence of data most immediately that we might be able to.
Uh bless for use in the data challenges that are coming up.
So at some point in the relatively near future, we're not sure of the exact date we're going to be.
Releasing some data of some type from Chorus Bridge to AI for the purposes of challenges that there's been some effort to define.
What those challenges are, and so now that there are data in the cloud, we can start to understand which data seem potentially ready.
I think want to do, you know, put her best foot forward and make sure that people are developing and training and testing models on.
Uh on data that is suitable for AI, since that is the whole goal and so we want to be a little careful about releasing things and understand their quality and suitability.
And so all of the great efforts everybody's bent on getting their data ready for contribution to this is going to start to play out for the first time.
And in real tests of stuff.
So that's very exciting.
The timing of it is such that we may just want to release a subset of the data, and so starting to look at what's there and assess its suitability is going to benefit greatly from people who both have a the clinical expertise to understand what data might be required and how we might use alternate versions of things as we start to look at things, we'll kind of see why we could kind of answer that question with these data or maybe with that or maybe this combination as well as some authority to say these are.
Written in stone, these versions of the of the challenge questions are written in stone or not.
Other kinds of things like that.
So Eric's got a couple of different excellent vantage points on those questions that not many other folks involved in the project do.
But since he is not here, we're hoping that other people, I guess what we might try and do, Marty.
Umm.
There's not gonna be here as have the most definitive list of the challenge questions that we can get our hands on during the call available.
I don't have, but I think is the most definitive list, so maybe we could try and get that from Sierra or someone else.
Umm.
And bring it up at the appropriate time.
So I think regardless of the challenge, we're also just excited to be able to look across the data for the first time and for Jared to kind of show what's what's visible in the cloud.
And I maybe discuss some of the.
Well, celebrate the folks that have of uploaded their data.
Nudge folks who haven't yet try and understand what might be the current obstacles they're facing in a little more depth, as well as some of the issues around conformance of data that you've noticed so far.
So I think that's the that's the prelude and it'll be using this Aries tool that you've seen a couple times, but now it'll be showing chorus data.
Alright Jared.

Houghtaling, Jared 12:04
Alright, thanks for keeping it up.
Uh, and for the intro, so indeed I've spent the last couple days now that I have access to the data that has been uploaded on the cloud. I've been able to go through the process of ingesting it.
And I was able to do it successfully and I'll show today what that actually looks like.
Uh, but maybe.
And I'm not sure how we want to structure.
Maybe I'll show what's there and then I'll I'll discuss kind of by a findings or or my my take on how to improve or update the process for the remaining sites.
So I'll go that way.
So let me just share it and show you what I'm working with here.
OK.
And this is something for those of you who are on last week and I presented the posters from both standards and cloud.
This is the kind of cloud side of that so right now on and what I'm showing you is actually a portal on the MCH Azure base that flashy, it's just more or less a list of links to different services.
Andrew mentioned Aries.
We also have Atlas available and then this lab environments for those of you familiar with Jupiter Lab, you can go in.
Uh and I I this is the exact environment that I also used in the workshop.
Now, a couple weeks ago, so I think we sent the link around last week.
Those of you wanna watch it, you can see kind of how you can use this type of lab environment, but of that.
So what I did is I I took the files that were submitted, specifically the OMOP files and I process them and ingested them into Postgres on the cloud.
And to do that in that process I I actually ended up date shifting the data uh by ± 31 days.
This is again, it's a first pass at this.
I'm just kind of where we're going through and and kind of yeah, identifying how we can do this and what's going to work in the context of of creating this data set.
But these data are date shifted and there were three sites and I'll already I'll shift to here.
So there were three sites and thank you to those three sites that have data in OMOP uploaded on the cloud.
So it's nationwide. Seattle and UCLA.
Nationwide and Seattle both had a a relatively large number of patients, I guess in relation to what they will eventually upload.
UCLA, I think had something like 20 patients or so, which is why it shows up as zero here to obscure official count there ohm.
But I can go through today about what this actually looks like in areas after now I've ingested this and maybe more importantly as like in a researcher, what you might, uh, how might you use areas to investigate things like study feasibility and data quality and things like that across the network ohm.
And so the way that I see this is that area is kind of the first step toward actually putting together a study and creating cohorts and doing some more sophisticated analysis.
So this kind of gives you an entry way into that or a view on on what's available in your in your data set.
So let's dive in.
We have areas as I've I've gone through before has three different levels of detail that you confuse so you can view the network level.
You can view the data source level, so in this case for MIMIC or for nationwide and you can also view this source release level to.
You can actually zoom into the specific.
Data source, release of interest, and that's more or less an OMOP entity is the source release and.
And so if we start at the network level, we can actually do, and I'm gonna get to this piece last.
So there's the there's the status source feasibility where we can actually go in and say, hey, I want to know how many patients my I'm looking specifically for measurements I wanna know across my consortium whether data sources have patients with that.
And so you have that very high level, but then you can get more and more specific and you can look for a particular concept, a measurement concept that and whether or not it that exists within your network.
So maybe we can do that quickly.
Let's just pull up.
The standard blood pressure measurements, and we'll see if yes, if this exists.
But when it's not so then, and I guess if you were to do this, you would probably have a set of concepts of interest.
You could you could go through and check them, but that's getting very granular, right?
That's looking just for very specific concepts and up here you have domain requirements, desired domains, ranges, things like that.
So I'll I'll go through that.
Maybe a little bit.
Network performance is more or less how quickly things executed.
That's not so relevant for what we're doing on the central server, because we're doing everything in the same environment.
We have a quality assessment, so this is kind of an interesting view where you can actually compare side-by-side data quality performance across the different sites you'll see in this case, Seattle has very few issues related to plausibility, and you can actually go by actually CDM table and look at what so in this case, the measurement has the the most issues, which is consistent pretty much with what we see elsewhere.
And that kind of gives you again a network view.
You have a population overview where you can look at years of observation in your cohort.
Interestingly, Seattle and Nationwide are both children's hospitals and you can see that represented in ohm, I guess misses more age at first observation, but in in general when you start to live zoom into the the people, you see that this is a report.
I also like to see which is the Strand report which breaks down the data sets by domain.
Again, you see measurements are are the overwhelming majority of what's actually been submitted across all four sites, ohm.
And then you have a lot of drugs in both Seattle and Nationwide.
And then you have a smattering of the others, and then small amounts.
And then Usmap source codes.
This is something that I've talked about before, and that's gonna be pretty important to moving forward, which is if you don't have a mapping for a particular data elements, we would recommend including that elements and mapping it to 0.
So having some sort of source value and a source concept ID, but then having a 0 value in the concept ID field for the particular domain.
The nice thing about that is that it gives us a view and and you could do this locally or or on submission.
I guess we can discuss about what might be best in that regard, but it within the context of areas, it will give a view at a network level of potentially shared things that need to be mapped.
So if if a lot of people have, let's say a SNOMED code that is not currently in OMOP but they have this SNOMED source code, we can make a point to make sure that gets added to their central mapping repository and distribute that to sites.
So right now that is not the case, but this this is not being populated.
But yeah, moving forward that that is a potentially useful aspect of of this tool and I'll I can show some other examples where that's that's in place ohm.
But anyway, that's the last piece of the the network overview.
This the source overview I think is less interesting as of now because I first uploaded a couple days ago a small test set and then I uploaded the main set for both nationwide and for for Seattle.
Umm, the cool thing about this view eventually is that you have like a nice quality history and a domain continuity so you can view as you would get multiple releases submitted.
You can view how the quality has changed.
Ideally it's improving and then you can actually look and see in terms of continuity how your events are changing over time between source releases.
So like I said, we only have two releases here.
The first one was actually just a subset of data for testing.
The second one is more or less every all the home updated that was there and you can see the number of events and the change between those, those sources.
And like I said, as we get more releases this this view becomes more interesting.
As of now it's it's not so awesome, but the last bit I think is is pretty relevant for analytics and for kind of diving in to know what types of data you have available.
And again, this is not connecting to the database.
This is actually just generated using aggregated results and it's a static.
It's a static set of files that are being shown, so I'm not.
There's no database connections or anything.
There's no person level information.
UM, but what you can do so you can right now we're looking at nationwide and if you go to measurements, umm, you can see the number of people with the measurements and you can see heart rate is very high.
It's captured, of course many times ohm, but what's pretty interesting is you can actually view a comparison of that same measurement.
So in this case concept 3027018 across your network for those sites that actually capture that measurement.
If and I I talked to back in in, I think December about harmonizing units and why units are important to measurements and hopefully this gives a bit of an idea as to why uh, in this case, both UCLA and nationwide use the same per minute units in our in our measurements.
And you can then compare those side by side.
Uh, it is interesting to see you have higher heart rates at nationwide than UCLA.
I'm guessing it's pediatric versus general population, but yeah.
So you have more or less this this view on any measurements that are preserved across data science, new versions of areas are expanding that functionality to other domains.
Currently this is this kind of side by side value view is only available in the measurement domain, at least within areas, but it gives you a feel for things that you can do.
I think systolic blood pressure might also have. Yeah.
So then indeed, you again have this ability to compare distributions of measurements across sites.
Ohm and that that's at like a value level, right?
But what I think is also interesting is let's say I know Eric was interested in in particular procedures and so you can look at things like ventilator, uh.
And you can see how many records of what's actual concepts you have for ventilators.
And this is in the nationwide data sets.
You can do the same thing in Seattle.
Uh, and then you can try to see if you have an overlap.
Like if we were to say right, let's take that concept ID ohm and let's see if we pop back to our feasibility, we can actually search on that concept.
Uh.
Signature and then you can see how many are actually available and get a view on.
So in this case, Seattle, Nationwide and MIMIC all have that that concept UCLA does not, but you would have been an idea of a min versus A Max population and this min versus Max is gonna if you added multiple concepts it'll go from like yeah it's kind of an org condition I guess if you want to say across all the concepts you might add and this is what also this is saying here.
So that is kind of a a quick view on how you might use this to look up conditions, measurements or procedures of interest, at least at a high level.
And like I said, this is really the first step to toward more complex analysis.
I don't know Andrew if if I can dive into any of that in more detail.

Williams, Andrew E 24:42
Wanna pause a little bit?
I know you've presented areas a few times.
You presented some of the things that Aries uses.
The data quality dashboard and so on.
OK, it's got a lot of functionality and I just want to stop and rehearse some of the main things we're using this for as more sites get their data in, we're going to provide feedback probably on a quarterly basis about things that seem like they need to be improved on the ETL side and you'll be acting on those those report that's going to go out and it's going to reflect stuff that can be seen here.
As well as what you might see locally if you stand up data quality dashboard in Aries, so you'll have a network view that adds information that you won't have locally if your data.
Seem like they ought to be similar to another institution's data because the kind of institution you are, but it looks very different.
Just might be something to investigate like if all the other institutions that seem are similar, then that's that's a cue that there might be something to investigate.
Umm.
And there are other issues that that only come to view in a network.
So I mean, so there's one whole set of the functionality for this is just as an aide to your own efforts to produce the data that covers the concepts that chorus has defined as being high priority and would want you to make sure you're covering and getting as well as all the quality issues associated with them.
And then the other thing Jared been talking about is after that data is there and you've got a network of sites that have all of their data in there.
What can you use it for?
So the feasibility aspect is directly relevant to how we started off the call talking about the release of data for the challenge.
So that's a specific set of things that we want to understand.
It's not the global picture and it's data quality issues.
It's what?
What data are present that might be relevant to questions around intubation and prediction of the things related to ventilation and so forth.
And so there's kind of separate this whole set of functionality roughly divides into stuff for data quality and ETL improvement and stuff for actually conducting a study.
It's got both of those built-in, and he's been covering them, and so that's probably stuff a lot of folks know, but it's just worth stopping because it's a lot of information and I just want to make sure I'm helping to reinforce some of the main things we want to make sure people understand.
But what we're going to show next, actually, what do you think, Jared?
I want to at some point before it's too long.
Since we're seeing now that essentially 2 of the chorus sites have uploaded a very substantial amount of their EHR data at some point before we end the call, I want to have give sites a chance to let us know kind of where they are with respect to being able to upload the EHR data that they have and the waveform data that they have in particular.
Uh, so I'll let you decide when you think the best time is that.
But I wanna I wanna make sure we don't have too much time elapse before we have an opportunity to discuss that.
And I also I see that Cleira has joined us.
One of the things we wanted to do was begin to start the conversations around which data elements might be present that are relevant to the challenges.
And Sara, you might have sort of a most definitive list of what those challenges are.

So we could at least start that.
I know we're not be able to finish it, but kind of do a little bit more of what what Jared was just doing in terms of showing some of the concepts that are relevant in the feasibility view here.

McCrary, Ciera 28:39
Yeah.
Umm yeah.
I don't wanna take up everyone's time because I know it's probably busy, but I can at least put in the side a picture of the definitive challenges.
We have 4 right now of the challenge list that we have right now, so we can at least people come all add over while Jared who keep going and we can go from there.

Williams, Andrew E 29:06
Great.

Houghtaling, Jared 29:08
Sounds great.
Thank you.
Umm OK.
So I guess at A at a high level, I think Andrew summarized it nicely.
You indeed have this ability to track the changes in a single source over time.
Look at how the quality changes.
Look at how the.
Yeah, the domain quantity of events changes and that will then inform how you can update and improve your transformation.
And then in the second context, you're actually looking at and comparing your your OMOP data against the rest of the network. Uh.

McCrary, Ciera 29:39
Feeling inside out?
Yeah, from this like I'm but.

Houghtaling, Jared 29:52
And I think like I said, the next step to this and we start start talking about evaluating feasibility for a data challenge.
Umm you start to need to be able to create or let's say isolate numbers of patients of interest or I I understand how many particular patients fit within a particular category that you've defined or that can answer research question.
This is something that I demoed at the workshop and this is now actually an Atlas instance running on.
Running on the the MCH cloud and it's referencing the data from these sites that haven't been adjusted.
So these four sites are also available to, to peruse and Atlas here and then it nationwide, sealing UCLA.
And and what I've actually done in this instance is I've gone through and imported all of the phenotype library definitions.
Ohm into the cohort definitions that can actually be generated against these data sets.
So if we step into one of these really quickly, a cohort definition is and you can see this gets actually very complex, but it's more or less capturing 20 a set of medical events that are related to a patient and as some sort of temporal sequence potentially kind of all these different logic elements about what type of group of patient do you wanna you wanna investigate.
In this case, it's a type 2 diabetes patient.
UM and Atlas is a tool that allows you to kind of define that in a graphical user interface.
And the cool thing about Atlas is it like on the back end of this, there's a standardized representation of these definitions and those standardized representations are available in the phenotype library, which I got to talk about a little bit a couple weeks ago.
For those of you who on the call before the workshop. Umm, it's.
And so this is a very useful tool to grab these cohort definitions or these phenotypes, patient phenotypes and actually you can import.
And what I've done is I've imported them from this location into Atlas and we can then actually generate them against the data that's been that's been ingested and we can use that as kind of a metric to see which datasets support which phenotypes effectively.
And when you're looking at it in in the context of running the challenge with, say, we wanna have a challenge related to intubation and a challenge related to I stroke or neurological issues with an ICU.
Uh.
Then you can actually define what you would expect that to look like using this tool or using an existing cohort definition.
That's been validated by a community.
So these phenotype phenotype library.
Kind of aggregates different definitions of phenotypes or cohort definitions, and others that are effectively clinically validated, and so there there are varying stages of clinical validation I should say.
And so then in order to answer the question that that Andrew is getting at, we could take a look and see and and I'm now just quickly looking.
So if you look at excavation with with intubation, comma recovery and then kind of a free for all uh, but if we were to look at comma, these are all not COMMER laters and Gibson is also not there.
So I think if we were to do, we have.

Walmwright, Mark 33:44
Jared, these are mostly Pediatrics, so it might be helpful to look at something and say like hepatic failure or cardiac arrest.

Houghtaling, Jared 33:54
Yeah, absolutely.
So in terms of actually to generate something on the fly here, we can do that.
So this is then more or less just saying I wanna have patients enter my cohorts that have some condition recorded of this hepatic failure and it uses this descendant concept capture which says hepatic failure plus more specific representations of that in OMAP UMM.
And what you can then do with that is you can go in and say let's generate a paddock failure for Seattle.
This is actually now running the query, so it's it's taking that definition, which is represented by this this SQL query, and it's actually running that against our Seattle data set.
We see, at least in this subset, that nine patients conform to this particular definition.
With these concepts in the Seattle dataset, just as a rough estimate, one place that is not included that I'm not showing here is that we will merge these data sets together into a single Corus data set, and that's another process that I'm kind of in the process of doing, which then instead of having nine patients from Seattle that fit this, we would actually have a broader set of patients that are kind of aggregated from their different sources, be it nationwide or other other sources as we ingest them.
And then we'll increase our counts in that merge data set to actually run bottles and have larger population sizes.
Uh, but so that was a cohort generation on the fly for Seattle.
We could do the same thing for the other ones, and once you have that generated then the process that I walk through at the workshop was actually taking those cohorts, building a prediction or estimation model around them.
And then training that model to kind of, yeah, predictor estimated outcome in the interest.
So again, we're getting now very detailed into the analytics side of things, but all that to say this is kind of where we're headed as kind of a preview of of the data that has now been adjusted.

Williams, Andrew E 36:00
And there is a connection between this work which is taking what is at least a good start at a definition for something like hepatic failure.
And using it to understand the data that you're producing at your own site, apart from any use in a study.
Without going through the work of developing a cohort definition, you've got concepts a lot of times you've got hundreds of concepts that all relate to one thing, and it's a little bit more of a challenge to see how some number of specific versions of something that get at creatinine or something are meaningful as discrete units.
Whereas if you have a a cohort definition that uses a concept set that gathers all of them together in some way, then you start to it's it's easier to interact with and say, well, this data looks good.
This looks like kind of what I would expect.
If you're somebody who understands what to expect at your own institution, you've got the clinical and or informatics knowledge needed to say this is.
This is a meaningful thing.

So in other words, this concept set or cohort definition level is often useful when you're developing an ETL because it gives you a sense of when you put the data together into more meaningful units and you graph it out, or you get a tabular view of it and Jared hasn't had a chance yet to show how once you've defined a cohort, you can get, you can drill down into cohorts and show other attributes associated with them and so forth.
And how they're affected by changes in logic and all of that, some of those views, those graphical and tabular views at this more meaningful cohort level are really useful to say, oh, this looks right.
This looks like we would expect or something.
Looks off here.
They aren't leveraged so much in the Aries tool or the data quality dashboard tool which stays at this granular single concept level.
So this is an important supplementary activity that you can choose to engage in.
It's the reason that having example use cases is often important part of doing the development work.
You if you stay at just the individual concept level or in the aggregate over the whole OMOP data set level, it's often hard to see things that will show up immediately as soon as you start to use it in a real use case.
And so this is even though it's mostly about the analytics as as Jared saying it is a very useful adjunct to any ETL development stuff.
So it's just worth noting that.

Houghtaling, Jared 38:55
Thank you.
Alright.
And I also wanted to kind of give put some of this in context as well for those of you, I've seen demos from Dell. I think Dell is on the call, umm about the Integrated Viewer application and this would be the idea that you could actually go through and view like the images or the waveforms from a particular patient.
And in that application you'll be able to actually search for a type of patient that you're looking for.
Maybe you want to look for a diabetic patient or your somebody with heart failure, or somebody who came in.
I don't know with with some set of of, uh, preexisting conditions, I don't know.
And what we will do is actually use these cohort definitions to help with that search process.
So when you wanna, if you wanna view the the waveform signals from somebody that has, I don't know, some arrhythmia then we will, her application is going to refer to the cohort that's been generated for some general arrhythmia and it'll pull those patients that fit that particular phenotype of interest and then you can browse them as like a short list of possible patients.
So there this is not only just a standalone tool, this is integrated with a lot of the other tooling pieces and and applications that are gonna be happening downstream once we start to get this spun up into an analytical approach.
Umm, so just a bit of a shout out there to to that UM.
Alright, so I know Eric is not here.
I don't know, Andrew, if that we have.

Williams, Andrew E 40:31
And.

Houghtaling, Jared 40:32
I see still 25 minutes left, so I don't know how we wanna best go through.

Williams, Andrew E 40:34
Yeah, I'd like to shift a little bit.
I think this has been a fantastic overview.
Again, we're really excited as we said at the beginning, we see that two sites have uh submitted substantial amounts of their data, another set of sites have submitted some waveform data.
Also, we are feeling a little so we don't want to go too fast.
We are feeling some urgency around defining the.
Umm.
Data that is going to be available in time for release and use in the challenge.
A separate group not on this calls not an issue we're going to solve on the call.
A separate group is going to be reviewing the data that's been submitted and that gets submitted over the next number of days.
To make that judgment and so, uh, that's teeing up the question of, umm, where sites are at and where they expect to be at, I guess over the course of the next few days in terms of starting to upload data being make and we should take some time to clarify where we're at with the SOP for uploading data which is well on its way to being finalized.
It's already been covered.
The steps have already been covered in a couple of places, but the sort of officially released and blessed SOP for it is is making progress as a last connection to what was said here, specifically around cohort definitions.
If you start to try and use this either Atlas or the any other way of working with the cohort definitions that you can get from the phenotype library, that is a good idea.
It's good to know some of these have been validated, others are just been proposed.
And are, you know, in process umm and it's good to review what odyssey means by a cohort definition.
Your first intuition might be wrong.
Your first intuition might be this is how you define.
Just diabetes at all.
It's more than that.
You have to have an entry date and an exit date.
You may have other inclusion criteria that are specific to a particular question that all get wrapped up into a single cohort definition, so it's just worth reviewing that to make sure you understand what that is.
If you start to use it, but.
Now that we've covered that, I think if we could shift to a discussion of where sites.
Are and I guess nudge.
You don't need to upload, but also to update where you're at.
If you've got a blocker, one of you know the Google form for giving your current status is a place where you can say we're blocked by getting local permission from our institution to do X or something.
So there's a lot of sites currently that have an in progress.
Umm, which is correct if if there's sort of a timeline I think might understand blocked with respect to a timeline, she might not be blocked permanently, but you say I can't get it done in the next week or so because we're still waiting to hear from somebody.
You might update it with an understanding of what that blockage is for for uploading, but umm, I know Marty, if you can start calling on folks who might be willing to give us where they're at with the process of trying to upload at least their EHR data.
If you want to also comment on the waveform data that would be, that would be great.
To the extent that you, you know that so.

Alvarez, Marta 44:30
OK.
I'll start with a uh, nationwide.
Looks like Jenny and Caleb are on the phone and also OK.

Xu, Jinyu 44:40
Yeah, we have uploaded the data both the hour and the waveform.

Williams, Andrew E 44:48
It.

Alvarez, Marta 44:51
Hit uh, sorry.
I see Mike and Andrew are on from Pitt.

Kirley, Mike 45:02
I can speak to that.
So as far as wave form data, we still are developing the process to export it.
Umm, but for EHR data, we're.
We're supposed to submit it in the next couple of weeks.
I'm exporting it for review by next Tuesday so.
Yeah.

Williams, Andrew E 45:37
Thanks.

Alvarez, Marta 45:39
And I'm sorry.
Did his did you say EHR as well?
Yes. you did.

Williams, Andrew E 45:43
Yeah, yeah.

Alvarez, Marta 45:44
Sorry.
OK.
And UM and jate.

Schmidt, Heidi E 45:51
You must have known that I wee just typing or EHR data is almost finalized.
There's something to look at code wise in the observation table.
Aries and DQD and Achilles have been run on what's there.
We won't have specimen note or device exposure.
Waveforms are being a deidentified by my colleague and.
Some of that data has already been uh put in Azure BLOB storage and that's where we're at.
Once I've got our data after the ETL sorted for observation, then it's.
Need to be written out.

Alvarez, Marta 46:48
OK.
Umm Mayo.
So OK.

Abramovich, Sergej, M.D. 46:54
Ohh everyone, so our dates.
As I said, we are currently bottle back with this risk management process and what I see it's not a deep understanding but but it's see during the conversations with Alex Ruiz and our software engineers who are responsible for that process, the things moving forward and our engineers, they talked me that it's not gonna be a big deal, it's it's OK, but it requires some feeling of forms, answers for questions.
So this is our current bottleneck, but I hope it will be resolved soon.
Thank you.

Alvarez, Marta 47:37
OK.
Thanks, UCSF.
I think Eddie is on the call.
Umm.

Amorim De Cerqueira Filho, Edilberto 47:48
Hey guys.

Alvarez, Marta 47:48
Maybe he dropped off OK.

Amorim De Cerqueira Filho, Edilberto 47:48
Sorry. I was.
I just had a vibrate and lost what the conversation is, but I heard my name there.

Williams, Andrew E 47:57
We are asking where sites are with respect to uploading their EHR waveform data.
Just Jared, in the earlier part of the know, if you were on so it was showing what was up and.

Amorim De Cerqueira Filho, Edilberto 48:04
Yes. I'm going to be downloading hopefully the waveform data for all the cases in 20 minutes, they said.

It's the plan when it comes to the waveform, so see if that's gonna work because the there is no mechanism for large scale and waveform download, so we are just setting that up now.
But then, hopefully he's gonna work so more to report next week.

Williams, Andrew E 48:31
How about that EHR data there, Eddie?

Alvarez, Marta 48:32
And umm.

Williams, Andrew E 48:34
What do you think?

Amorim De Cerqueira Filho, Edilberto 48:34
So I thought you xeen mention waveform.
So for EHR 20 minutes 40 minutes ago, we had a meeting on also the same Michael scaling the extract.
So we are setting up a dublb to be able to run that tagger.
We got all the other EHR data, but the tokenization that you see as F is it takes so getting mad tagger to run or something that has been a blocker for us.
But now it looks like that's gonna work, but more to report on that and it coming weeks.
The problem is you SF is that we didn't get yet approval to put the data at Azure.
So we are waiting for MCB on some of the sign offs and forms on that.
So that's the bigger, bigger blocker we have right now.
So it's a data security assessment stuff.

Williams, Andrew E 49:29
Thanks if you could prompt somebody at your site to update any information in the umm, speed up using that Google form about those things, I think those are particularly interesting where we know there might be just hard to control the administrative processes that are gonna affect timelines for getting data released for the challenges.
It'll be particularly important to have some central place where we kind of know which sites are experiencing those sorts of things, so that that would be great.
But thank you for that update.
It's exciting about the waveform.
That's really cool.

Alvarez, Marta 50:04
Umm.
How about Colombia?

Nametz, Daniel C. 50:06
Yeah, I can.

Alvarez, Marta 50:06
There's a few people on.

Nametz, Daniel C. 50:08
I can go for Columbia.
So we've had this blockage with the AI working group here at Columbia or NYP. where the meeting just keeps getting delayed, delayed, delayed.
But finally, we have a meeting with that group on Monday, so hopefully we can get the approval to upload the data and then in the meantime for EHR.
With the assumption that Monday's meeting will go well, uh, I am downloading that data and then I need to put in the request for the wave forms, but I'd imagine we'd be at the top of the queue for the waveform download.
So yeah.

Alvarez, Marta 50:52
Thank you, Duke.
I see.

Ashok Bhatia 50:58
This is the show.

Alvarez, Marta 50:58
Yeah.

Ashok Bhatia 50:59
Yeah, I'm on our side.
We have got a umm I'm meeting scheduled with the.
With the Microsoft Alex Ruiz regarding the Azure pipeline setup that is the bottleneck for now.
But we also have got a problem.
Our CCN team is working with I think Jared, Jared regarding the Artah.
We have got some problem on that one as well.
Umm.
And yeah, those are the two things that we are still like working on.
And Regina will be updating the the form regarding this.

Houghtaling, Jared 51:45
And maybe a brief update there.
I met with Daniel two days ago to go through getting the quality checks run and it was, it seems like it was a matter of isolating the smaller cohort of patients from your very large OMOP data sets and then executing checks on those that.

Ashok Bhatia 52:01
Umm.

Houghtaling, Jared 52:03
But I haven't heard from him.
If that's been successful or not, but things are hopefully moving along there.

Ashok Bhatia 52:08
OK.
Yeah, I will check on with like, sure.

Houghtaling, Jared 52:11
Great. Thanks.

Alvarez, Marta 52:15
Uh Emery?

Bold, Del 52:21
EA for us, we have when is 600 patients right now with both waveforms and images.
And I guess our bottleneck is where we are, where we can decide whether we should update shift or not.
Yeah, that's where we are right now.
I think Tony is figuring out whether Emery has a any kind of restrictions handing out limited set.
So if not, then we probably will send limited set, so that's why we are right now and the original P ran.
That's one actually one.
I'm not sure if we should push it because it's giant.
Umm.
And whether we need to run it with like you know very targeted dictionary.
And yeah, that's that's where we are on the data quality dashboard.
We rent it and Achilles ran, but I haven't ran the errors indexer yet.
But I guess I wanna ask, do we need to set aside course set to run it or I?

Alvarez, Marta 53:26
OK.

Bold, Del 53:32
I thought we had, uh, you know, the dots, running the whole set was the good idea.
That was my understanding.
So I have access to the enterprise the whole Emory set.

Houghtaling, Jared 53:45
Yeah.

Bold, Del 53:46
So how?

Houghtaling, Jared 53:48
Yeah.
So I guess that's we we we had this discussion on one of the prior calls.

Bold, Del 53:53
OK.

Houghtaling, Jared 53:53
I think there's not a a single answer that's gonna cover everybody, I think.
In the case of Duke, they were running into time constraints effectively that were causing the the the runs to be much slower when they were doing it on the full set.
So running it on the subset of data was just a much quicker and more efficient way to do it.

Bold, Del 54:12
Ohh.

Houghtaling, Jared 54:14
It does limit, of course, the errors that you're catching to just any events related to this subset of patients, but it depends on kind of what your needs are and what your computational environment is to make that decision.

Bold, Del 54:20
Umm.
When you said run, uh, you're talking about data quality dashboard, right?

Houghtaling, Jared 54:30
Achilles DQD both.

Bold, Del 54:32
Both OK, each one for me.
Each one was like 28 minutes somewhere there, so I think we're OK with that.

Houghtaling, Jared 54:39
And.

Bold, Del 54:41
But what I'm what?
I'm confused is that if we get an instruction that hey, you mapped wrong.
If we had this our own course set.
Do we map it there or do you go back to the enterprise team A?
Hey, you mapped, you know, the chorus is one has to map it there and fix it there or what?
What is the workflow would be?

Houghtaling, Jared 55:05
So that's again, I think gonna be organization dependent probably.
I I guess you you have two options right?
You fixed the problem or you you update the data at the source?
Or you patch it and update your kind of representation of the subset of data and depending on how close you're working with that kind of enterprise team and what their needs are, might change that thing and.

Bold, Del 55:26
Oh yes, they they're actually on board.
They actually wanted to have quality data.
Uh, but the on the other hand, we might be doing more than we need to do.
You know what I mean?

Williams, Andrew E 55:39
I think the broad principle is and we have to probably move on to the next site, but the broad principle is the more you're using the data and feed it back to the central folks, the better their data is gonna be.

Bold, Del 55:43
Sure.
Uh-huh.

Williams, Andrew E 55:52
They are in the best position to judge if it's too idiosyncratic and issue for them to make any adjustments for, but it is going to be institution specific.

Bold, Del 56:00
OK.
Correct.
Thank you.

Alvarez, Marta 56:07
I'm in.
UFD.
Without that, Tecan was on.

OZKAZGAT BASLANTI,Tezcan 56:11
Yeah.
Hello ohm.

Alvarez, Marta 56:12
I.

OZKAZGAT BASLANTI,Tezcan 56:13
So we have the data about 1300 subjects with EHR and waveform data.
How, and Chester has been working on this and he is out for a while and we were thinking he would be back.
Uh, it's that is certain period of time, but that doesn't seem to be happening.
So we will need to kind of assign someone else to take that and complete it an upload one step that is left is kind of get our IDR integrated data repository approval in the deidentification process and making sure it is done as they approve and get that out.
So that's one block which we had there and but we are moving forward to assign it to someone else.

Alvarez, Marta 57:01
Thank you.
Umm will you want to talk for UVA?

Ashe, William (wadigo) 57:07
True thing.
So we've been working with related groups, not always the same people, but a lot of overlapping folks to get the tools set up to get the the, the UMM data back from them so that we can review it and make sure it has the right stuff and then ultimately get it verified.
So we have waveforms and I should say we're on this Omaha to essentially OMAP 5.3.
Is my understanding it's been built up from about 5.2, so it's a little rough in that definitely needs some work, but to to the extent we have 42 patients worth of EHR in the set to be verified, we have 16150 patients worth of wave forms in that set to be verified.
And we have a data request in place that has not been like I've not heard much recently about.
Unfortunately, the team that provides us the data and verifies that the data is the same and we've been similarly waiting for, I think over 2 weeks at this point to have the data to verify it.
Haven't heard anything, so I'll keep following up on that, but it's just slowly moving through.
It's to do shell processes.

Alvarez, Marta 58:19
OK, I think the only other is Seattle might be in the same boat as nationwide, right?
We already know what they have.
Yeah, I'm at.

Wainwright, Mark 58:29
No, wait, wait.

Marta, wait here for Seattle myself institution.
Yeah, this is Mark Wainwright.

Alvarez, Marta 58:37
Ohh no.
Yes, I'm sorry.
I I knew you were there, but I think we know the answer already based on the data that's been submitted.

Wainwright, Mark 58:42
OK, that's fine.

Alvarez, Marta 58:44
Yes, thank you though, OK.

Wainwright, Mark 58:44
But I don't have a comment about t the question actually we've been very conscious of data security for obvious reasons.
I'm wondering what's the approach sites are taking or if they feel the need for tracking which patients data have been submitted.
We see the need as we modify our data submission or submit bits and pieces of patients data as we acquire waveform or imaging for example to keep close track of which patients data for Corus have have left the institution and so we are setting up a system to do that.
I'm wondering if that's something that you see is something we should be discussing as a group for Corus or you're leaving it to the individuals institutions or you don't see it as much of an issue until.

Williams, Andrew E 59:35
It seems like a great idea to have a system like that.
I don't think we're requiring sites to do so, but to the extent that we can collaborate and share approaches as you're suggesting, I really I think that's a wonderful question to our conversation to have and umm and then come to a conclusion about and develop some method of disseminating and sharing what sites are doing.
I so that's that's my understanding.
We're not.
We're not requiring it at certainly seems like a best practice and a great thing to have raised.
So thanks for bringing it up.

McCrary, Ciera 1:00:13
Actually, umm Mark, if he once you guys have that, would you be willing to share that at one of the Tuesday meetings on that's with more of the tooling and standards and everyone that might be a perfect place for us to like show what you guys are working on and see if we can implement that a little bit larger.

Wainwright, Mark 1:00:34
Yeah, the glad to know it's a Tableau report that present and it's obviously it's a work in progress.
We learn as we go, but yeah, we'd be glad to. Yeah.

Alvarez, Marta 1:00:46
Umm, thank you.

Williams, Andrew E 1:00:47
So.

Alvarez, Marta 1:00:48
And Brian's also on the call for MIT, I don't know if there's anything that we wanna, I'm not sure where that stands since I know Jared is working with that a lot.

Brian Gow 1:00:57
Yeah.
Which Schmidt, Ted.
Umm, a version of the OMAP data, but not the the chorus formatted OMAP data to Jared and included some wave forms and images along with that.
So I think Terry can speak, but I think he's planning to integrate that and into its final form.

Houghtaling, Jared 1:01:18
Yep.
So the what I showed just earlier on this call was the demo set of MIMIC data that's publicly available and I have the OMOP tables from Brian for the the full set.
But I still need to do some conversion and some linking on it, which I'm hoping to do in the next week and I can upload them to ingest that into the the portal as well, so that should be one of the next sites that gets up in full I.

Williams, Andrew E 1:01:45
So I'm gonna reiterate that we're expecting an SOP to be blessed to kind of goes into the details of how to upload things.
Some of the outstanding issues are.
Whether raw data can be released centrally in order to do work and uh, Jared, you ran into some issues that were the DLL's are idiosyncratic in ways that require extra work.
We're in the future going to probably have a script that will yield a positive result whether the DDLS are in standard format and other kinds of validation that will probably precede the umm, upload, that'll be determined.
Uh, I would ask with a little more emphasis that I have in the past that everybody update use the Google form to update their status by the end of this week.
If they can, it's going to be really important for us.
There's a group.
As I said, it's going to be trying to understand what data are either currently available.
We're going to become available in time for a data release and to examine it in order to see which data can be used for the challenges, and we won't really be able to do that effectively unless we have an understanding of what to anticipate and we don't outside of the verbal descriptions that we get in meetings like this in in the data acquisition, we don't have a way to get that view and it's quite pressing that we at least understand the current blockers.
And so when you update that, please include not just the status indicator, but win your.
Your best guess is that you will be pushing the button to submit your EHR and your waveform data.
If you have not yet done so, C! update the status and include in there the date approximate date you expect to be pushing the button to submit the EHR and waveform data by the end of this week.
Please, we're going to really need that.
And so really thank you so much for all of your work.
I hope you're as excited as we are to be able to see the data using these tools and we look forward to working with you and and pushing forward and all that complicated and important work you're doing.
And thanks Jared for doing all that stuff and showing us everything.
And we will see you next week like.

Houghtaling, Jared 1:04:20
Thanks so much.
All talk to you soon.

Talapova, Polina 1:04:21
Good time.

Alvarez, Marta stopped transcription