CHoRUS Bridge2AI Standards Office Hours-20250109_190419-Meeting Recording

January 9, 2025, 5:58PM

1h 5m 51s

Talapova, Polina 2:06
Hello.

Alvarez, Marta 2:09
Hello.

Rosenthal, Eric S.,MD 2:09
Everyone.
Jared Marty I.
I did see Andrew's message that he's gonna join late today, but I was curious if you had an agenda for today.
Specifically.

Alvarez, Marta 2:36
We do.
It's troubleshooting critical data quality issues after recent data.

Rosenthal, Eric S.,MD 2:44
We are then aligned, yeah, so.
If it's OK.
If I just gave a few earmarks, I was.
I did have the chance to catch up.
Couple times last week with Andrew, but I think we're we're looking to do is enter a phase in the next couple of weeks to months where we really.
Reduce the sites distractions to focus really on the issue of data quality and so.
That's exactly what I wanted to say, and I'm glad that it's where you're headed with the agenda.

Alvarez, Marta 3:32
Awesome. Thanks Eric.
Wait another minute or two to get people.
Sometime to get here before we get started.

Rosenthal, Eric S.,MD 3:46
Maybe while we're waiting, I I wanted to just talk a little bit about I I understand that you know we have the benefit of the data quality dashboard.
But one of the things we're also gonna work on, and I think Sierra's gonna gonna take this up, is really.
Refining the data acquisition.
Sop so that it sort of it.
It's as clear as possible and sort of in a level readable by engineers. In other words, you know it's fully specified as we can.
And that may also involve.
Supplying some metadata about one site.
Like what types of IC U.S. data is collected from and how data shifting was done and.
You know, think things like this, but.
Say that some of the aspects of focusing on data quality are.
Are also just about making sure that the data that we provide is fully contextualized so that people understand where it was derived from, what it what its provenance was, and what its inclusion criteria were.

McCrary, Ciera 5:09
Yeah, I have an initial draft almost done.
I'm finishing it up now so hopefully I will have something.
For at least like Jared, Polina and like some of you guys to get a high level view first before I send it out to everyone.

Alvarez, Marta 5:34
OK.
Thank you.
It looks like everyone that's on the call has been here before, so I think we can get started. If you'd like, Jared.

Houghtaling, Jared 5:48
That sounds great. I will.
I will dive right into it.
Maybe just first a bit of a note.
I'm coming down for a bit of a cold.
Might be coughing. Apologies.
For that, throughout the next little bit, but I'll go ahead and share my screen.
Because I think the big topic of today is data quality. But I'm gonna kind of touch on two aspects.
So like I think the first one is the kind of the the data or the formatting and the encoding and everything of the.
The delivered files is one piece.
I'm just gonna touch briefly on that just to to simplify ingestion.
And I've been working with some sites and thank you all for your your patience and going back and forth on this.
In that regard, and then the other piece is the actual like data, the quality of the data contents and what that looks like and how we can improve it and where it needs to be headed.
So.
Bit of a status update even at the very beginning.
This is a pretty exciting number I think for me to see.
14 sources in here.
It's thirteen sources plus plus the merge data set, which is awesome.
It's just about everybody has some extent of omop data, so congratulations to all of you for for some awesome work there.
I think.
This this is also a very up to date aeries instance so.
I ran.
I ingested Emory and Columbia yesterday, so for those of you from those sites on the call, your data is really.
Like very fresh and I was able to then run the merge process also yesterday.
So now you'll see that our merge data set has over 30,000 individuals.
Currently I think a a big chunk of those are from the MGH.
It's about what about 1/4 of that is from MGH.
But still.
It's a pretty nice diverse set.
One thing that was particularly apparent is if you take a look at these start observed and end observed the dates actually and this is I think where I had started to have some discussions with Eric last week and Andrew together.
There, there's a lot of diversity. Let's say that you have dates anywhere from the last 10 years to last 30 years, all the way back to 1900, which is gonna be fixed.
But all that to say, there's there's some discrepancies and diversity happening in this process.
Maybe also just a very quick side note. Why merge would have a different observed and observed the observation periods get recalculated following the merge.
I don't use the ones that were calculated at the sites I I calculate in the merge sets.
Essentially an observation period from the first observed event until the the final observed events in our merge set.
So that's just why there's some some discrepancy there.
Then you'll see that there is a list of data quality issues and.
They range in.
In what they are going to talk a little bit about those data quality issues in a bit and I'll go into like why they're relevant and how you can investigate them.
But before I dive into all of that, I did want to plug very briefly.
So we have given sites access or we've created guest accounts on the central cloud.
For representatives from all sites.
Of like the 20 people who received guest accounts, I think only a small handful have actually accepted their invite.
If you accept that invite, then you are entered into the MCB Azure tenant and I can give you access to this site. This this area site that I am showing right now and that will allow you to actually go and browse through your site's data quality data.
As we do on here and then you can run queries and try to investigate why things are.
Looking the way they are, let's say.
Or Sierra, go for it.

McCrary, Ciera 10:21
Oh no. I was just gonna say I, for those who haven't accepted it.
And it might get lost.
I am currently working to get a new push like a new e-mail push to make sure you can accept again.

Houghtaling, Jared 10:36
Thank you.
Yeah, so I went.
I went through just add add you all to this and I was not able to add many of you because you were not yet there to to be added let's say.
So just a bit of bit of a heads up there.
The other then piece before I also get into. This is is to ingest these data.
These are all they're coming into the cloud as CSV files and.
In the beginning, it was all individual CSV files per patient, which was opened up like a whole can of worms, and that was a challenge. We did some ingestion like that.
Now they're just CSV file per table, which has made things considerably easier.
But there's a few kind of key points that I want to highlight in those CSV file deliveries.
So what I would strongly recommend that you do.
In terms of how you coordinate this.
It's not correct.
Just go to this site.
So this GitHub repository is the odysseycommondata model.
Click on inst. EDL 5.4.
I'll put this in the chat.
If you're delivering data, please, please, please, please go to this site.
Pick the DDL for your.
For your database of of of choice, whatever you're using at your sites, run that DDL in a separate temp schema that you have access to that you can do whatever you want with.
And use the tables there as kind of your your extract tables.
So you'll maybe you get an extract from some other team at your university or hospital and they do their own stuff with it.
That's fine, but please make sure that what?
The the the data that you will provide centrally is sitting in tables that have this exact specification.
From their till. And so in each one of these directories, let's say let's we pop into.
Spark, let's say of data bricks.
And you all you need. There's gonna be 4 files here.
You can create the tables and then you can set indexes on the tables and constrain them and keys.
But you're only actually concerned with the the DDL itself. If you wanna do the other stuff, go for it.
But it's not.
It's it's, it's not a requirement for for delivery. And what you'll then see is you'll see this this kind of set of.
Table definitions and baked into these are not only the the data types.
But also in many I guess spark is maybe a bad example, but if you look for something like Postgres you'll see.
Where things cannot be null.
So you have not null constraints which can be very useful.
Because of course you can remove those and.
And violate convention and not know about it.
So if they're there, it's gonna force you to make sure you have these columns filled appropriately.
So please run the DDL and then insert your data into the tables that are created. If you're if you'd like, I can send you the set of insert statements, but more or less it should be like insert from.
I don't know.
Sorry.
All the columns from person that you need and insert them into the actual person table and you can use this pretty much to structure your your insertion statements.
If you are curious or wondering what that means, feel free to reach out to me offline and we can go through how to actually do that insertion.
But that will save a lot of time and also we'll make sure that the data that you want to deliver is ingested appropriately.
There have been a couple cases already where I didn't know the contents of a couple different columns and they were not in the order that I expected, so I had to kind of guess and we don't have to do that.
So please make sure you you follow along with this DBL and you don't have to use this for all of your work at your site, but for the extract please insert it here so.
Just one plug there, but.
As as we are progressing now with more data deliveries in Omop, it's it's getting significantly better.
So thank you all and it will continue, of course to get better over time, but just one, one last plug for using the DDL.
And then your CSV files themselves.
I would recommend using like an all caps naming convention. I think we have. There's a there's a post actually on the.
I'll maybe also make better or I don't. If Marty, if you wanna link that discussion thread.
With the data delivery requirements or something, but I I think I cover most of it there.
So I don't need to.
I don't need to go through that again.
One thing I would like though is.
If possible for sites to calculate your condition era and drug era.
Table values, so I think almost all of you have and and this kind of maybe gets to what Eric was talking about with a more specific or detailed.
Data acquisition SOP or like a data delivery SOP but.
You all are delivering an observation period and in theory we can calculate the the the condition error and and drug error.
But it would be I think.
Good for you to do that.
On your side, especially if you're gonna run any sort of quality checks prior to delivery, you'll need to have those tables filled.
So I would assume that we'll place them in the required table set coming up, but I think some sites have delivered them, some have not.
So just a heads up there and then dose era has to be calculated at your site and I don't think anybody maybe only one site so far one or two sites I should say.
Has has calculated and delivered that table.
So that is currently missing, and that's a very important table. When you talk about pharmacokinetics or any sort of drug based studies, you wanna know how much drug patients are receiving at what time and for how long they received it?
So please please if you don't know what I'm talking about, maybe Marty as well.
Another link if you could post a link to the dose era discussion.
How to calculate that what what it means?
Sorry, Eric. Go for it. Yep. Yep.

Rosenthal, Eric S.,MD 17:50
Jared, can I ask a question?
I think I think medication dosing data will be one of the I'll I would predict the lasting contributions to this data set if we can, if we can accomplish it, there's so many models about someone's heart rate variability and this and that, but not as much about.
The inputs about what we actually do to the patients to get them in a certain state and so.
We're we always look at phenomenon phenomenology.
But we we're not necessarily linking practice, so if we can do that, it's just a huge deliverable to the scientific community. That being said, I do have a question for you, Jared, which is that my understanding is that omop has some gaps in regards to infusions and can.
You tell us if there's a way in which we can overcome that here or.
What's the state-of-the-art?

Houghtaling, Jared 18:49
Yes. So there's a also a discussion thread that I started about a proposal for how we should capture these and I think.
The let's say the the short answer is we would add several columns to the drug exposure table that would capture the information needed to like understand an infusion essentially.
The question came up also, how does this relate to the dose era and and I want to capture infusion but my infusion data is in.
Like I don't volume per.
Time per wait or something like this and I need to convert that into an actual dosage for the patient in terms of a quantity of of drug that they're receiving.
And so yeah, short answer, we're going to add some columns in order to capture this and we'll provide guidance and examples of like what this practically looks like, but we have a discussion thread and we're going to hopefully try to push it forward here in the next in.
The in the coming weeks.
And jill, also in the chat.
So hopefully did that answer your question.

Rosenthal, Eric S.,MD 20:00
Jared. Yeah, and that's amazing.

Houghtaling, Jared 20:11
Yeah, absolutely.

Rosenthal, Eric S.,MD 20:12
We're very motivated.

Houghtaling, Jared 20:14
Awesome. Thanks.
But yeah, so to take a look at the discussion thread and again any please feel free to interact there.
And ask questions or do whatever we'll try.
Also keep that updated as we go.
Maybe just quickly while I'm on the topic.
I do have an e-mail digest every week that gets sent out.
For folks who are interested that that contain, like any updates to the discussions, if any of you on the call, if this is the first time you're hearing about it or if you would like to be added to this, throw your e-mail in the chat and I can.

Add it to you.
It comes out on Friday mornings.
But it's more or less it. It compiles all of the updates in the standards discussion area.
And it sends off a quick e-mail that says here's what happened in the last week, so you can take a look.
Sorry to respond to your question in the chat.
I'm not gonna have time on this call to go through every data quality issue from every site.
What we will do is we will set up one on ones we've done this already with sites, but we will set up one on ones.
With sites, it's typically about an hour meeting to go through explicitly that sites data quality.
Checks in in in detail.
Maybe this is sorry.
There's a lot of. I'm talking about a lot of different things now, but.
So I guess short answer, we're not gonna go through all of the site data quality, but what I will try to do at least is go through ones that were overlapping between sites like things that I saw on a lot of sites.
That's that were were missing essentially that are potentially large flaws for trying to do. The type of research so.
As I say now and you had your hand up.

Pirani, Cenan N. 22:17
Yeah, sorry, this might be going back to what you said earlier, but this is we're still trying to get our data ready kind of to deliver.
We were under the impression that.
That each cmap kind of set needed to be by patient.
You're telling me that's no longer the case, that we can just deliver a whole table of all the patients?

Houghtaling, Jared 22:39
For Homebase.
For the other data types.
The other data types should be per patient organized, but cmop is a separate and I can maybe pull up quickly.

Rosenthal, Eric S.,MD 22:52
Play.

Houghtaling, Jared 22:54
I can point you in the right direction on on that SOP.

Pirani, Cenan N. 22:58
And when did that change?
Did that change recently?

Houghtaling, Jared 23:01
I don't know when the SOP went out last summer, maybe.

Rosenthal, Eric S.,MD 23:08
Yeah, I don't think that represents a change.

Houghtaling, Jared 23:15
Yeah, I'll, I'll. I'll send this to you.

Pirani, Cenan N. 23:16
Then we flipped it around.

Houghtaling, Jared 23:18
I don't remember where the SOP is, but maybe Gill.
Yeah, I, I don't when that that officially happened, but we can take a look or we I can touch base offline.

Clermont, Gilles 23:30
Yeah, that was a that was abundantly discussed and at office hour and it was several months ago.

Pirani, Cenan N. 23:30
OK.

Clermont, Gilles 23:39
So yeah, I I I agree with you.
Probably around Midsummer. Something like this.

Pirani, Cenan N. 23:48
OK. Thanks.

Houghtaling, Jared 23:54
OK, I was gonna say one more thing, but I have.
I have forgotten it now.
Yes, so so going back to what Dell had asked.
What I've also done in the last couple weeks is I updated the report package.
I'm sorry, that's just not the right place. I'll show you quickly.
So there's now a new version of our reports.
So 2.1, I'll send this in the chat.
These reports get executed.
After you deliver data, they kinda go through your data set.
They compare it with what's in the merge.
They give you kind of a breakdown of quality and content and characterization and everything.
And I'm now able to run this.
Actually it's kinda interesting.
This interfaces with now the data delivery method like the kind of file manifest that gets that gets created for this.
This dashboard to kind of speed things up, but I will create one of these for all of the sites it does.
It's dependent on an up to date, So what I've had to wait, I think we're kind of at the point at the end now of this big wave of deliveries. And so now that I have them.
More or less finalized, I think it sounds like there's maybe gonna be one or two more deliveries this week from maybe mgh and and I'm not sure if somebody else is gonna put one in.
But once I have a merge that's fairly stable, then I'll run this process.
And one of the things that it does is it's awesome.
Pit will as well. Cool.
But it it runs cohort creation logic against the merge set.
And because I've augmented the merge set so I know for every row in the merge set, I can trace it to a specific sites.
We can run on the merge, set a cohort generation, and then I can pick apart that cohort and stratify it by sight.
We don't quite yet have enough to stratify by ICU type, but I'll get into that in a little bit.
But what you will then see is like hey, here are the most common cohorts which you can think of as kind of like a.
I don't know a patient profile almost or somebody like a cohort of of diabetic patients or cohort of patients with sepsis admitted through the emergency room or whatever. It's going to be.
So those cohorts are shown that in the report you also get a detailed list of like all 500 of them.
With your count of patients represented versus the count of the merge patients.
So there's some interesting kind of insights hopefully to be gained.
By reviewing that.
This areas view doesn't have yet cohort views.
At some point I will upgrade to a newer version of Aries that can provide some of that functionality, but we haven't haven't.
I haven't been able to make that transition quite yet.
So for now this is this is what we're working with.
OK, but let me think just quickly here.
I'll first if if you haven't seen this view, this is this is Aries.
It's acronym for a research exploration system.
I'm showing all the sites that we have currently in our Postgres database on Azure.
So each one of these isn't is its own cmop instance.
The counts in this data source view are slightly muted.
Has some interesting logic where it looks at observation periods and events.
It also has some obfuscation logic if if the counts are too small.
So these are kind of approximate counts.
I would.
I would take them as once we go into the specific person report it's it's more explicit.
So at the network level, we have this type of a view. We can quickly look at overview of quality.
So here you'll see.
And this is it's already getting too noisy with so many sites, but the data quality dashboard. I've gone through it in some different office hours, but it has three categories of quality checks.
So it looks at are the tables complete?
Do they conform to?
Odyssey conventions and is the data actually plausible? Like the dates appear when they should and do people have events before they're born or after they die? Stuff like that.
And so those are the three categories and what you see is that we're like, I mean, there's not as many checks on completeness, but in general the we're doing better on completeness and conformance and plausibility are are kind of the the areas.
To to investigate.
And then if we come down here, we can actually look per cmop table and we can see which tables have the the biggest issues.
So what?
You'll see.
Is and this is not really a fair comparison in my opinion, this this should be normalized because there are different numbers of checks per table.
So it's not really fair to say all measurements are terrible.
There's like 500 measurement checks for relative to maybe only 150 drug exposure so.
But all that aside kind of gives you some insight into which sites have issues in which in which.
Domains.
And so like I said, those of you who accept this invite will have access and can, like, start to investigate this application.
This is the population overview and I'm I'm sorry, I'm just kinda going through quickly here areas and then I'll I'll dive into the quality checks, but I think it's relevant to show you where stuff is at in like a brief introduction.
So we have a population overview here.
You can already also see some issues, right? If we look at age at first observation, Mayo seems to have patients that are -25 at the 1st, so they have either a date shifting problem that's that's potentially not affecting their.
Date of birth but it is affecting their data.
I don't know. Or there's other issues that there's also other MGH also has some negative dates as well, I see.
And then we have patients that are very, very old.
And both mgh.
But maybe, maybe he's religious.
I don't know.
I think MGH is working on that as well.
And then we have cumulative observation.
And it seems like we have indeed this is one where we had like from 1900 was the 1st starting date.
And so it's like 180.
Out until like I wanted the future.
So there's a population overview that already shows you some insight into data quality.
Here you can actually see domain content at a kind of a high level.
Red is measurement, so you can see that that we're pretty measurement domain dominant.
Followed, probably closely by observation and drugs. This is pretty typical for cmop sources that I have worked with.
Procedures are are limited, but that makes to me some sense when we talk about having.
A both small patient cohorts and and in theory most many sites are limiting just to the visits that that.
That these patients are.
Encountering ICU.
And so you're maybe in a single visit, you're not having so many you're not undergoing so many procedures.
So yeah, that's your data strand reports.
And then there's an unmapped source codes.
This is what we've talked to you about. If you map things to 0.
What this does is it actually takes a look at.
Things that are mapped to zero and what the source value was that led to that mapping, and then it gives you the sources of the sites that actually are are implicated in this.
So like for instance we have these.
Sorry.
Three sites so mimic actually really two sites mimic and Duke.
Have these unit source values that that are just simply unmapped, like we know what they are, but they haven't been mapped.
So you can take a look at that as well as you go.
This does.
It is predicated on delivering source value fields though, so just a heads up that if you want to do that I would. I would definitely check the source value fields to make sure that they contain only what you expect.
If you if you deliver data but it is helpful to have a view on what things are not being mapped.
To some extent it, and we'll go through this in more detail at the individual one on ones.
And then the level below that, so not the network, but now we're at the level of individual sites, but across all the site releases.
So maybe the interesting one to show here is actually the merge.
So what you see this is actually a nice trend. We see that the population is growing.
The December deadline brought a lot of patients and then even in the last couple releases we are on a nice trajectory upward and the data quality is actually getting getting better. I should also say though.
The the data quality, at least on the merge.
I wouldn't.
This is this is slightly under selling the the number of issues the measurement table has gotten to be close to. I wanna say 400 million or 500 million rows. The data quality checks can take a very long time, so I do some of the plausible unit.
Checks for the measurement table I have.
I have omitted just to save time so.
But we can talk also more about that with your sites.
This middle category, sorry.
Does it contains this kind of longitudinal history of the release and like I think of course merge has the most releases, it has the most data.
It's it's interesting to see how it evolves.
Over time, and you can also look at data quality history where you can really see.
And you can see actually are the number of checks because we've removed the.
The checks related to plausible units have gone down considerably, like we've we've we've chopped them in half. but they make up a massive portion of the checks. They take a ton of time because each one is an A distinct query on a 500 million row database. So but.
I working to refactor that to improve it.
Also cool to see you see that we did update the vocabulary.
Between.
May and.
May and December, we also actually up updated the data quality.
Version since then, so this also isn't really apples to to apples here, but.
Still interesting to note and you have this in for each site, right?
You can go and take a look at all the the checks and per table and how they've evolved and what your logic changes do.
So there's some interesting information as you start to have more releases in this zone.
And then you get to where the meat of of the information is, and that's actually at the release level.
So right now we're looking at this is Mgha latest release, but I can take a quick look. I have not even looked at the merge release yet since I ran it this morning.
This is what it's looking at. I do know.
Maybe I can pause just quickly.
So first, we're gonna probably. This'll end up looking tighter this year of birth.
I think we're gonna do something with the mimic dates of birth.
I'm not exactly sure what.
But then the ethnicity.
I know, Eric, you had some concerns that ethnicity proportion of of Hispanic was was far lower.
Than what you would have expected.
I don't know.
I know that there are a few sites that are not mapping this.

Rosenthal, Eric S.,MD 36:18
Yeah.

Houghtaling, Jared 36:19
You see this orange bin is.
Is unknown or no matching concept, which means it's it's either mapped to 0.
I guess 0 is unknown.
Sorry, so the green is is 0 and then.
Orange is just you map it to something that doesn't exist.
Yeah. I don't know if you have any other thoughts or wanna explain further on that, Eric.
Sorry, I forgot to ask.

Rosenthal, Eric S.,MD 36:43
You turn on major.
Yeah, the you know, our hospitals are in generally urban locations and I I I know that some are in areas that you know have different demographics than others.
But I I just thought that we would expect to see 10%.
Maybe you know 15% Hispanic population, so to see?
What I what was the number that you saw five or was it greater than that?

Houghtaling, Jared 37:12
This is like this is is now even lower.

I think this is like, yeah, what 5% ish?
Yep.

Rosenthal, Eric S.,MD 37:18
Yeah, I I just thought, you know, there's, like you said, plausibility.
Like, is it plausible?
And maybe if we looked by site, for example, that would give us more information, like if it's some sites at 0, then you know that would be, you know helpful. Hey Andrew.

Houghtaling, Jared 37:31
Yeah.

RS

Williams, Andrew E 37:35
OK.
Let's just comment.
The Convention that's for recording race and ethnicity data in a different way, which recognizes that they're not fundamentally distinct ideas and.
That they're complex in a bunch of ways and enables sites to record multiple race and ethnicities is almost ratified and honestly and will be pushing for that. I think the.
Standard reporting categories that NIH requires are necessary that we have a mapping to those.
They are much less useful outside of that strict reporting purpose for understanding all the issues around the complexities of people's identity, racially and ethnically, and use of that information, and whether it's HealthEquity or other kinds of research.
So I think we're going to be able to do that to a much more.

Rosenthal, Eric S.,MD 38:32
Here we go.

Williams, Andrew E 38:33
Satisfying degree we we helped to author that whole approach.
And yeah, it's pretty close to being settled.
And so in the next few weeks, we'll we'll have a more interesting set of race and ethnicity data to look at and guidance on how to use that convention.

RS

Rosenthal, Eric S.,MD 38:56
The other comment I wanted to make is you know.
As.
You have experience with this.
You can guide us, but as we go through these.
Data quality issues, some of them in in my mind, seem to rise higher than others.
And I don't know if there is a general pecking order or like for which issues sites should should solve.
You know, first I mean one one is you know general formatting issues.
Is that it?
You know, is the data readable and so forth.
But I'll give an example that we were working with at our site, which was.
And so you showed ages that didn't seem plausible.
'Cause they were say negative numbers or we we hope that there's a lot of patients who are 110 years old, but we recognize they're not. So that led us to think about whether we failed for example to date shift our birth dates.
And so I I wonder if you might comment on whether there are there's a sort of order that you recommend sites work through, either because it could reflect an issue that sensitive like privacy or rather that it's sort of fundamental to, you know some.
Some other aspect you know the other issue we had with age is that unknown patients in Epic are often given a birth date of a 1910 January one or something like that.
And.
So that may not be as important as an issue like, well, the patient's birth date wasn't wasn't time shifter or something like that.
So can you guide us there where sites should start and whether you give them an order?

RS

Houghtaling, Jared 40:42
Yeah. So it's a great point.
I.
I mean, I think severity of the issue is is really fundamental to to being able to address them appropriately.
I think that's one of the big things that we discuss in the one-on-one meetings where we try to at least discuss is like hey, here are the 70 issues that we see of those severity. I think that there's maybe four or five that like really should.
Should take the bulk of your attention because they're the most impactful, just maybe generally to your data set for being comprehensible or
They will have big impacts on on downstream research.
So I think the fact like year of birth, there's there's a lot to be said about date shifting and this whole kind of realm of of dates, but something like for instance missing date times not having date times in your date set. That is a pretty major issue.
When you talk about wanting to do things on scales that are smaller than one day.
And I think for for maybe half of the sites here, I think that there are date time fields missing or many date time fields missing.
That's a.
Were there set to to be midnight or something?
Which, which is is is going to be a problem for tying all the data together appropriately.
I wanted to point out, maybe just quickly while we're on this.
View for Columbia. I don't know if anybody from Columbia is on the call, but this is also an age issue or it seems to me like an age issue.
Maybe it's a a a year of birth shifting thing, but to have so many people bend on 1944.
It seems.
Odd to me, I don't know if if that's expected, but maybe something to review for the Columbia team.
But they have a much larger proportion of Hispanic patients than than otherwise.
Sorry, go for it Andrew.

Williams, Andrew E 42:32
No, no, no, sorry. I just wanted to expand on the idea that there not all the data quality issues are created equal. We've drafted ontology.
That we are going to be building some tooling around.
For identifying whether a data quality issue is critical or not. So it's critically important to address or not to help prioritize as well as its state of investigation.
So if you've looked into it and you found that, well, that's the way we get the data.
We don't have any other way of finding out whether it's true or not.
That's all we can do about it.
We can only just flag that it's looks funny or something, so there's this process of saying this has been noted, hasn't been investigated or it's been investigated.
Undergoing investigation or we found out that it is and then the kind of data quality issue it is. Is it something that related to an ETL problem that's potentially fixable or is it a characteristic of the data itself?
And so we've got some terms that can be attached to individual rows in your own data.
That will help you understand both how important it is and where you are in the process of investigating it. And.
That's coming.
It's just, you know, we're we're aware because we go through these same kinds of operations ourselves that there's a need to sort through all the many things that show up, figure out which ones are important, and to keep track of where they're at. And I think we're going.
To be.
Talking to you, you know in the next.
Couple of months as we go through more iterations on data quality, about how to use that ontology and looking for your feedback, hopefully it's gonna be very helpful, but it's just amplifying the theme that there are some things like these birth dates that are really important and affect.
Everything else so on. So just wanted to amplify a little bit.
Set the stage.

Houghtaling, Jared 44:32
All right.
Yeah. Thank you.
So couple couple things.
So one thing that actually is interesting is by default as well, data quality dashboard extracts do support notes, so you'll see you can actually you can show the note.
Column here.
We haven't added any notes, but it is a feature that we can effectively use and integrate here so that you could review notes like for instance the standards team or anybody else could make notes on these columns and say, hey, this is like with either severity or what.
So we have actually have some some options in the short term.
To do some annotations as well.
But that aside, what I've done this is Columbia still.
And Columbia's data set's relatively small in comparison with the others, but just to kinda go through quickly on some of these.
The initial ask for omop tables did not include care site provider location, any sort of health system information, and there are data quality checks that check those fields and what typically what has happened is sites of course have delivered their tables without removing the care site IDs that.
Something is associated to. In this case these are visit details associated with care site IDs, but the care site table is. Is is empty, doesn't have the associated ID.
So this this check will fail.
So things, anything that references care site, provider or location, you can pretty much disregard in this.
I can remove them actually from the quality checks. I should probably have done that.
They can be.
They can be essentially excluded from what gets checked.
Because until we start delivering and processing those tables, which will happen, they are not relevant.
This is one that's more concerning, actually quite concerning.
This is a device exposure failure, so half of the device exposure records in Colombia's device exposure table.
Are referencing concepts that we don't know about.
So either they're custom concepts for Columbia or Columbia is using a different version of the vocabulary, or not exactly. Sure, we can take we have to take a look and run the query.

But if you expand.
The the check itself.
It's you can actually run this kind of.
You'll see.
There's a little comment that says violated rows begin.
Violated rows end if you run the sequel in between those those comments, it'll give you the actual rows that are that are issues, but at least it it is identified as issues.
And so typically what also prior to the one-on-one sessions, I try to go through and run some of these and actually give the examples of what rows were.
Were or what rows it had picked up on.
And so this means and and of course devices. Maybe those are ventilators, maybe those are other devices, but that that's that can be very relevant in an ICU setting which device a patient is coming in contact with.
So that's an important point and I would say pretty high priority to fix.
These ones so discharge to concept ID.
These are.
Gonna be.
Let's see what's the differences that have a value that's not in the visit domain.
Yeah, these are more.
There's some overlap here, but this is more or less saying that in the visit occurrence you have this admitted to and discharge to concept ID that refer to where the patient came from and where they're headed. After this visit occurrence.
Those.
The visit domain if you're going to use those columns, these are not required columns.
They're relevant in in, in certain research approaches, or knowing kind of about the the movement of a patient.
I would say that both of these are probably lower priority, although still necessary to investigate.
Then the device exposure.
You want to talk about prioritization.
Again another issue here with conformance.
So we have an out of domain concept ID or a set of out of domain concept IDs. So something like 10% of our condition error table is not in the condition domain and I think once we go to the the the next page I think we'll see.
More of those pop up.
Yep, and this is essentially the same thing, so.
You'll have two failures.
One is the condition occurrence and one is the condition error which is calculated based off of the condition occurrence.
This is more or less saying you've you've mapped to things in the condition that that are in the wrong domain.
They're in the concept table, but they're not.
They're not.
Maybe their measurements are they should go in a different table.
And so we saw these kind of.
I guess when going through the if I will not go into too much more detail on these, we're kind of down to the nitty gritty.
But if we talk about what what has been shared across.
Across all sites or most sites, I think.
Domain and mapping issues. So things like either mappings to out of domain concepts or mappings to concepts that simply just don't exist.
Are fairly common across all sites.
With varying degrees of severity like some sites have maybe only a handful of of rows that are violating, and other sites have like half.
But that's those are semantic mapping issues typically.
And they could come from.
Mismatches and vocabulary. They can come from, like inappropriate loaded vocabulary.
There's there's lots of sources and we talk typically on the calls about how to approach it based on what we see.
So.
If I have to summarize, that's that's gonna be, let's say, semantic concepts, issues where you have issues with the concepts themselves. Mimic data specifically, had a lot of issues with data typing.
And so mimic uses identifiers that are big integers and the data quality tool actually was complaining about certain.
Integers that were very very negative, like, I don't know -900 zillion or something like this.
And that was throwing off a lot of and so if you go to actually the mimic, I'll show you that quickly. Just so you all are up to speed on that.
If I go to fail and I sort of records failed.
There are.
There's date issues, but then there's also date data typing issues.
See, let's see if I see it here.
Yeah. So So what you're seeing now is there's a lot of.
Issues that are related to things being in the future and of course mimic will have a lot of failures on that because they have explicitly shifted all of their stuff into the future.
I'm trying to find. Actually there's there are some that you'll also see pop up that are like things happening prior to birth, things happening after death.
Those are very important.
What's also very important is if you have end dates of either.
Conditions or drugs or devices that appear prior to their start dates.
Or that that appear far after those are more difficult. Typically in the EHR system, sometimes end dates can be.
Estimated or they're replaced based on some some default business logic.
So those those types of date related temporal issues are also many times very high priority.
I'm trying to find. Yeah, so it this is.
This is perfect.
So here again I I mentioned the data typing thing.
This actually I've only seen this with mimic data.
I've never even seen this check fail before, but this is just indicating because we're using big integers, so we're allowing sites to use big integers in their in their ID columns.
For whatever reason, mimics are are causing issues.
And so I don't.
I don't know that that well.
Yeah, it's not.
It's not a huge, huge problem, because we're able to map it in the merge set anyway.
OK, sorry go for Azure I I.

Williams, Andrew E 53:03
Oh, this is great, Jared.
I I would love to have a brief conversation. maybe because you're winding up here about the note capacity that you just demonstrated and whether or not we should suggest.
That sites use it for at least some subset of the things that are in the cloud representation of their data to either ask questions or to explain the origin of the issue.
And so the question is, should we do that? And if we suggest that sites start to do that in the next couple of weeks, how they take it on as a task where there's a tremendous amount of annotation that could happen, right?
How do we make that a, you know, bite off reasonable chunks of annotation?
In ways that are going to be helpful both to sites and to our ability to guide sites as they address these issues.
So I think that's an appealing idea, but I'm gonna first throw it to you, Jared, to ask if you think.
Finding a manageable set of things to annotate in the cloud by sites.
Is potentially a useful thing to start working on in the near term for most sites.

Houghtaling, Jared 54:28
Yeah. So I think the first thing that we would need to do is come up with a workflow.
So Aries at least this version of Aries and I haven't checked on the the newer version, but it's not an interactive note display, so you would have to.
We'll have to augment the notes.
In another way, outside of Aries, which is also possible, but I would need to look up that, but I think let's say, technically speaking, if if that is a noninsue, which I don't think it is, I don't think it's gonna be a problem.
Then I think it's gonna kind of come down to how much sites are using Aries to investigate their data.
And I my hope is that once ste engineers and representatives get access to this space that you will go through in detail and like we'll have our meeting our one-on-one meeting but then you can go through offline and and kind of try to tease apart what.

Actually happening in your data because I haven't even shown all the cool stuff that you can see like per domain, right?
Like you can really look at all these measurements and you can sit inside the measurements and you can take a look at.
I don't know all the certain things there's station like temporal, temporal.
Elements that are interesting.
So pretty cool stuff is available for for. For this the using this tool and I would encourage any of you who get access to it.
To start poking around in it, and I think if sites are using it, I think the note thing could be a valuable way, at least in the short term, to.
Help prioritize and understand severity and then understand progress on certain issues.
So.

Williams, Andrew E 56:09
That's the. That's all right.

Houghtaling, Jared 56:09
But yeah.

Williams, Andrew E 56:10
That's great.
The the first thing you said was the most important that the functionality for interactive, you know, use of the note field isn't present in the cloud and it would take some development work to make it happen.
So that's in any event in the future and we can discuss it.
I'm sorry I don't usually throw out those sort of hypotheticals in the midst of a call, but I thought we might be able to just go forward.
Here with that.
But that's it's we'll take that into a Tuesday meeting and and see what, if anything, we do to follow up on it next time.
But you were starting to kind of round up with the conclusions about the kinds of things you'd seen.

Maybe you should just resume that, since the notes.
Conversation didn't pan.

Rosenthal, Eric S.,MD 57:04
I have a question that's maybe a little out of scope for this group, but just planning a thought, Jared, which is.
It would seem like.
The right thing to start looking at extracts that we can populate on publicly facing web pages.
Just about the progress of our of our work, we have to clear this, you know, with our RB of a. you know showing summary data, but we we've already shown summary data.
So I don't think there's some.
Allergies there.
Except that if we were connecting our, you know, cloud space to something external, so is is this the is Ares the kind of view that we would want to do that from or would we want to wait until the data quality issues are such that it's reasonable to?
Show you know the age distribution of patients and so forth.
I mean, even that the patient counts I think are meaningful and they they give a certain.
You know.
Level of information to the scientific community about where the project's at.

Haughtaling, Jared 58:13
I I think there's definitely interesting content to share in areas.
I don't know if areas is necessarily like.
The nicest presentation of that type of data, I think it's it's practical, it's useful, but it's not.
It's not the the prettiest thing, but I think.
It comes down to a little bit what what you're trying to show and I think as we get better in data quality, some of these plots will become nicer, like the population overview will start to look nicer as we have things.
Or maybe.
I don't know.
Maybe where it was but.
Kind of.
Some of these overall plots about about where sites are at, we'll we'll start to look cooler.
I think there's some interesting ones with like measurements.
So if you go into some measurements.
It'll show you.
How they compare across the consortium and I I I will try to maybe find one that does this.
I think those could be very cool if you have a measurement that's repeated a lot.
I always think that these are kind of fun, but if you go here and these are all like concentrations.

Rosenthal, Eric S.,MD 59:26
I was thinking if you have like Ras, Ras scores. RASS would be a really interesting one.

Haughtaling, Jared 59:27
Umm.

Rosenthal, Eric S.,MD 59:31
But yeah. heart rate is another.

Haughtaling, Jared 59:35
Just do heart rate for now.

Rosenthal, Eric S.,MD 59:35
Kentucky. Kentucky Bluegrass is always interesting.
That one.

Haughtaling, Jared 59:39
But this is not. Let's see.
Yeah. OK.
So you have stuff like this, right?
Where this is also an interesting view to to have on on this right and you see nationwide is a clear outlier and you can understand why, right.

Rosenthal, Eric S.,MD 59:47
Yeah. Perfect.
Perfect example.

Haughtaling, Jared 59:55
And so it's a, it's a very cool thing to see and this exists for all like you can go through the measurements and try to find these, but.
It's an interesting way to compare across sites essentially.
So these types of plots, I could say I could see being interesting to to showcase.

Rosenthal, Eric S.,MD 1:00:13
Can you just show us how you drill down just 'cause? There's just an array of what?

Haughtaling, Jared 1:00:13
Umm.

Rosenthal, Eric S.,MD 1:00:18
What can be done? But.

Haughtaling, Jared 1:00:19
Yeah. So I don't know why it's not showing on my view, but what you can do is anytime you have this measurement value distribution, you can come in and you can say check measurement value distributions across the network and then it's gonna show you for those measurements that.

Rosenthal, Eric S.,MD 1:00:34
Yeah.

Haughtaling, Jared 1:00:35
Have units.
It'll show you what the the units are and it'll compare them, and in this case.
Umm.
Yeah. you have, like Columbia has has. Yeah. did couple options.

Rosenthal, Eric S.,MD 1:00:47
Does this imply that those are the only sites that have measurements of heart rate?

Haughtaling, Jared 1:00:53
So it just means that it's only the only ones that have the measurement plus a value plus a unit.
So in order to show up in this plot, you have to have all three of those pieces, and if you don't, units associated with your with your measurements and so in theory, I need to look into this. But merge should also be in here as like some sort.
Of intermediate between all these.

Williams, Andrew E 1:01:16
Sadly, I'll say Eric, though that Aneies is not primarily meant as a major dashboard.
INC mechanism for displaying the major characteristics in a way that's easily digestible for people who are coming to see what what's in the data. I think if we looked at the N3C kind of dashboarding efforts, you'll see they're really targeting the kinds of questions most people.
Would have.
They're displaying them and easily understood ways that don't require understanding what domains mean and what other kinds of things that are assumed here you're kind of.
This is kind of presuming a knowledge of stuff about omop and Odyssey that isn't usually present most viewers, and so there's a lot of good dashboarding stuff that's built to work with omop that we could.

Rosenthal, Eric S.,MD 1:01:53
K yes.

Williams, Andrew E 1:02:04
You know, spend some time figuring out how to adapt.
And Aries would complement that for people who decide they like this and they want to use it, and then they want to see what is it good for this purpose or where are the data quality issues that might make it a problem to use or something like that.
I think Aries is great for that secondary.
Deeper dive for people who want to go into the nitty gritty, but it's not primarily meant as a major dashboarding approach.

Rosenthal, Eric S.,MD 1:02:33
Thank you.

Haughtaling, Jared 1:02:36
Alright. And I see we're at time.
But like I said, please accept your invites as soon as you do. I will add you give you access to this and you can start to browse and play around with it.
You'll also get access to the the dashboard itself if you wanna know how many files you've uploaded and and make sure that everything that you uploaded made it in here.
Yeah. So.
Thank you all for listening and I guess I'll touch base with with you next week.

Alvarez, Marta 1:03:05
How do you feel better, Jared?

Haughtaling, Jared 1:03:11
Thanks very much.

Williams, Andrew E 1:03:13
Sorry, you don't feel well, Jared and and thank you for the massive work that you've been doing on this over the last few weeks.

Rosenthal, Eric S.,MD 1:03:14
Thank you.

Williams, Andrew E 1:03:19
I know it's it's really great and pivotal, so thanks.

Haughtaling, Jared 1:03:25
Hopefully another merged version next week and we'll push out some of the other stuff so it's looking good.

Williams, Andrew E 1:03:30
It is great to see the trends you were showing today. It was really exciting.

Rosenthal, Eric S.,MD 1:03:35
Tared one thing for a future date.
I know there has been some discussions around linkages.
And.
What's what's the ideal state that we wanna aim for?
It would be maybe helpful to if you feel like there's some deficiency there that we could better specify in in the revisions to the.
That that Ciera is working on to just give them more of an engineering specification.
That would be really helpful to hear. Even in a meeting like this. What that means for data quality in terms of promoting linkages.

Haughtaling, Jared 1:04:10
Yeah.
Absolutely. I mean, I think I think.
Sites have not yet delivered.
There's another site has delivered the the linkage.
Like registry so like where you create a table that says this is my waveform file. This is the visit it's linked to. This is the person it's linked to.
So I've been able to do it, but like in a very approximate way trying to match date times and it's clear that some of them are not shifted and some are and and like people have named files differently and it's yeah.

Rosenthal, Eric S.,MD 1:04:47
Yeah, that.
Yeah, yeah, yeah.
Well, we'll spare you so that you can recover, but but we're I think we're interested in in hearing about the front door, not the back door, to try to get to that level of critical.

Williams, Andrew E 1:05:01
Yeah, Eric, that's that's a front burner for us is like getting sites to get visit detail information on ICU visit information. The procedure code that's gonna be a use to link to data and assuring that the data are stored in the right file structure. I think those are.
The key key things.
The ones I just listed once those are in place, we'll be able to query and do the linkages there and it's, you know.
The the different pieces are specified that sites have had their hands full doing other things, but I think it's front burner and we're not.
We're not neglecting.
Yet.

Rosenthal, Eric S.,MD 1:05:36
OK.
Just saying that I'm excited for it.
Thank you so much.

Haughtaling, Jared 1:05:40
You're taxing.

Alvarez, Marta stopped transcription