Adrian H: 00:04 Hey, this is Adrian Hernandez, and welcome to the NIH Collaboratory Grand Rounds podcast. We're here to give you some extra time with our speaker, and ask them the tough and interesting questions you want to hear most. If you haven't already, we hope you'll watch the full Grand Rounds webinar recording to learn more. All of our Grand Rounds content can be found at RethinkingClinicalTrials.org. Thanks for joining.

Leslie Curtis: 00:27 Hi, this is Lesley Curtis from the NIH Collaboratory Coordinating Center and today we're here with Tom Carton and Keith Marzullo who will be reflecting on data linkage within, across and beyond PCORnet.

Leslie Curtis: 00:42 Tom and Keith, it's great to have you with us today. I think we're, we're aware of the issue of data linkage generally but Tom, I wonder if we could start by having you talk specifically about the problem that you're trying to solve, and how you've approached it.

Tom Carton: 00:57 Sure Leslie, thanks. So this is a problem that came up very early on in REACHnet. So I'm the Pi of the REACHnet clinical data research network, which is in Louisiana and Texas, with a good number of sites based in the Greater New Orleans area. One of the things that we wanted to deal with early was how we were going to handle the issue of patient data across multiple institutions, in the world of a limited data set as is the PCORnet common data model with only dates as the identifiers.

Tom Carton: 01:32 We needed some kind of a technology to be able to link, and to dedupe mostly for the deliverable that we had at the time to PCORI, which was longitudinal data capture. We investigated various methods, both internal and then external, and landed on a solution that allowed us to salt, hash, and then link in a four party system with the vendor providing the salt, the partners providing the hashes, and then the REACHnet Coordinating Center at the Louisiana Public Health Institute providing the linkage. We basically just created linkage keys that then we could then use for specific research use cases downstream.

Tom Carton: 02:20 So we did this once with a governance that allowed for this process, and then numerous times after for specific use cases to studies. I think from a PCORnet level we learned from the demonstration studies, specifically the antibiotics demonstration study, where we were linking clinical data with claims data from the health plans health plan research network, that local solutions are just that, local. We had a situation where we were able to match and link with Humana to fulfill our obligations for the ABX linkage and HealthCorps, and PEDSnet did the same, but our methods were different, our timelines were different, and the study team had to deal with two different approaches.

Tom Carton: 03:14 That really got a number of folks thinking, this was years ago, about how we might be able to approach a more global method across PCORnet, where we could learn from the networks and their distributed experiences.

Leslie Curtis: 03:31 Keith, I wonder if you could explain to our listeners what we mean when we talk about salt and hash.

Keith M: 03:38 So a hash is essentially, you take an input string and then it basically runs through a mathematical algorithm to output a jumble of letters and numbers of a fixed length. The idea is that the same input string, or different input strings should not end up with the same hash. It's essentially a way of encrypting the data, so that it's secure and it can be safely transmitted.

Keith M: 04:08 A salt is essentially a piece of text that gets added to the input string, so that if you have two names of John, for instance, they can end up being different hashes, and so that's a way of protecting the data if it's sitting there at rest so that somebody can't try to reverse engineer what the hash might be. When we do linkage across multiple databases, we actually want to start to use the same hash across patients, same salt across patients, excuse me, otherwise you end up with different hashes. and then there's no way to determine that the same patient exists in multiple databases.

Leslie Curtis: 04:51 Thanks, Keith. That was a really clear explanation. Let's continue, and I'd love to hear from you a little bit about the action plan that you came up with as a result of this, the work that you've done in PCORnet.

Keith M: 05:08 Yeah. The way that we were framing it, I think the first aspect was to kind of provide some of the rationale for why we want to do this. What are some of the use cases that we think we can accomplish? I think we proposed a very basic set of use cases initially, just to prove that we could do this. That would be allowing us to determine, what is the overlap across networks , what's the basic number of patients across PCORnet, and then can we use that to generate what you'd consider to be a table one, in terms of the demographics and comorbidities, and other aspects of the network?

Keith M: 05:50 But I think ultimately, really the purpose of this is to do targeted research studies in a much more rapid and efficient process. To do more things like the demonstration projects, to allow us to do surveillance and other types of research. That was kind of how we framed the discussion. These are the types of things that we wanted to accomplish,

Keith M: 06:13 and then in terms of just the specific steps, we laid it out basically in terms of governance and technology. The governance were related to things like, what parts of the common data model would we need to expand to encompass this additional work? What type of IRB protocols would we need to either establishe or amend? I think, based on previous experience it would be, most sites have a local IRB protocol that they use to govern their common data model activities. We anticipate being able to amend that protocol to cover that aspect of the work.

Keith M: 06:57 Then, just expectations around sort of what the study specific protocols would need to look like in order for people to do the work. Then there was a set of activities related to the PCORnet data sharing agreement, to just make sure that the agreement that's currently in place across the network would allow the exchange of this information to occur.

Keith M: 07:26 Then technology was just understanding what sort of vendors might be out there to provide this service, and then based on some landscape scans that had happened across the network previously, developing a number of attributes that could be put into an RFP, that would then be a sort of released, so that PCORnet could decide on a vendor to help provide this software and this service for the network.

Keith M: 07:56 Then finally there would be a set of activities related to the distributed query infrastructure, so that we could actually execute these queries using the existing query tools that we have for PCORnet.

Leslie Curtis: 08:10 Clearly this was really quite a significant undertaking for the team that worked on this. I wonder Tom, if you could touch on really some of the challenges that exist, or that you've encountered maybe beginning on the governance side.

Tom Carton: 08:31 Yeah, sure. I think Keith broadly outlined the approach and divided it nicely into governance and technology, and touched on the IRB and the data sharing processes. I think locally, when we first did this within ReachNet, it was difficult because we really didn't have any existing documentation or agreements.

We had an IRB that governed the common data model, as Keith had said, and really had to socialize our preferred approach for data linkage from scratch with all of our participating IRBs.

Tom Carton: 09:12 It was challenging to help IRBs understand, that in many ways they're not dealing with issues of computational data linkage to be able to speak the language that they could understand and to answer questions clearly and carefully to bring them on board. I think we have that same challenge across PCORnet and the level of experience of IRBs in dealing with these issues is heterogeneous across networks. Within our network we've kind of already had these conversations, so I'm not super concerned about our IRBs being comfortable with the PCORnet-wide method when they're okay with the REACHnet method.

Tom Carton: 09:54 However, for sites that have not tackled this issue yet, we're basically socializing these ideas from scratch. As you know, there are different levels of conservatism across different IRBs, and you're limited by the most conservative interpretations of what you're putting in front of them. That's a challenge, and we're dealing with that as a work group by basically bringing some experts across PCORnet together to create guidance documents that then individual networks can share with the IRBs.

Tom Carton: 10:28 In terms of the data use agreements that we have in place across PCORnet, they're flexible in that they allow for growth of the common data model and so we have conceptualized this as a growth of the current common data model, with an additional table, or additional elements within existing tables to accommodate for the hashes that then can be used to do linkage.

Tom Carton: 10:53 From a global PCORnet perspective, we have A, learned from the experience of the individual networks, but then B, worked within existing PCORnet level governance. That's allowed us to lessen the challenges, but nonetheless, we still have the concerns of the IRBs at individual institutions and so forth that we're currently working through, as well as just the members of the expert panel and work groups that are concerned and represent the interests of their systems, and want to make sure that this is done in a way that is safe, and secure, and translatable and so forth.

Leslie Curtis: 11:33 Right. That clearly is a priority for everyone, I'm sure. So, I want to really conclude by thanking both of you for joining me today. Really exciting work that you've done. I have no doubt that it'll be both critical for PCORnet, but for many others as well.

Clearly, you've worked through many of the thorny issues that exist in this area. So thanks for joining us today.

Leslie Curtis: 12:01 Our next podcast will be validating a computable phenotype: should results change a trial's prespecified primary outcome, with Gregory Simon and Susan Shortreed, and that'll be posted the week of December 3rd.

Adrian H: 12:19 Thanks for joining today's NIH Collaboratory Grand Rounds podcast. Let us know what you think by rating this interview on our website. We hope to see you again on our next Grand Rounds, Fridays at 1:00 PM Eastern time.