Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

GA4GH Computable Cohort Representation Agenda & Minutes 2021-2022

These are the minutes for the GA4GH **Computable Cohort Representation Subgroup**. For further information, please visit the GA4GH Work Stream page at www.ga4gh.org, or contact info@ga4gh.org.

The Clin/Pheno WS minutes hub can be found here.
Meeting recordings can be found here.

Subgroup Links

- Use Case Collection Form
- Landscape Analysis Spreadsheet (incl. form responses)

Meeting Protocols

- Please note that by participating in meetings, attendees agree to adhere to the GA4GH Standards for Professional Conduct.
- Meetings may be recorded for note-taking purposes. Recordings will be deleted within three months of the meeting taking place.

Table of Contents

Subgroup Links

Meeting Protocols

Table of Contents

Meeting Reports

- 2022-08-23 - Cohort Summary Statistics
- 2022-04-22 - Hackathon
- 2022-04-05 - Pre-Connect Hackathon Planning
- 2022-03-24 - Pre-Connect Hackathon Planning
- 2022-03-15 - Full Group Meeting - Hackathon Planning
- 2021-12-14 - Report Back Meeting
- 2021-12-02 - PHENO Team
- 2021-12-02 - MIACC Team
- 2021-11-18 - PHENO Team

2021-11-12 - MIACC Team
2021-11-02 - PHENO Team
2021-10-19
2021-10-05
2021-09-21
2021-07-27
2021-07-13
2021-06-29
2021-06-15
2021-05-13: Computable Cohort Representation Kick-Off

Meeting Reports

2022-08-23 - Cohort Summary Statistics

Please find the agenda/minutes for the summary statistics meeting here.

2022-04-22 - Hackathon

Please find hackathon minutes & materials here.

2022-04-05 - Pre-Connect Hackathon Planning

Chair: Melanie Courtot, Susheel Varma, Lindsay Smith
Attendees "Name (Project Affiliation)": Susan Fairley (GA4GH), Jeremy Adams (GA4GH), Ian Fore (NCI), Brian Walsh (OSHU), Jordi Rambla (CRG), Susheel Varma (HDR UK), Núria Queralt Rosinach (EJP RD), Soichi Ogishima (GEM-J), Philip Van Damme, Karen Haase, Francis Jeanson

	Actions Arising	Assigned To	Deadline
1.	Inquire about dbgap data (noting the caveats Karen raised)	Ian	
2.	Get a hold a few synthetic datasets		
3.	Set up servers:		

	- Data Connect - Including finishing up putting in the CINECA datasets - FHIR	- Ian, Francis - Susheel	
4.	Set up CINECA Beacon	Jordi	
5.	Do some pre-work on cohort definitions we have and some pseudocode for Asthma.	Susheel, Philip	
6.	Clean up the hackathon planning doc and start to put the agenda document together	Lindsay	

	Agenda Item	Speaker	Time

1.	Welcome, Housekeeping
	- GA4GH Connect Registration - April 20-22
	- GA4GH-wide Strategic Roadmap feedback opportunities
	  - Town Hall Meeting Registration
	  - Community Survey
		Lindsay Smith	5 min

2.	Planning a Hackathon at GA4GH Connect
	Planning Document
	Hackathon Roles
	~~1) Confirming the goal~~ Done
	2) Date/Time
	  - Thursday, April 21, 2022
	  8:15 AM - 9:45 AM (ET)
	3) Assigning roles and tasks
	  a) Session Chairs
	4) What preparation is needed?
	  ~~a) Selecting a phenotype/definitions~~
	  i) Done - Asthma
	  b) Selecting datasets
	  i) See previous meeting notes
	  c) Servers
	  i) HDR UK FHIR Server, AGHA?
	  d) How many hackathon meetings should we schedule? What will the kick-off look like?
		ALL	50 min

	e) Github Space 5) Agenda - Due April 14		
3.	A.O.B.	ALL	5 min

Minutes:

IF: Might be doing more planning at the Kickoff rather than actually doing it.
Set of asthma datasets we can possible use, but they are not harmonized: https://github.com/ga4gh/fasp-scripts/blob/master/notebooks/search/fhir/dbGaP%20studies%20with%20asthma%20in%20title.ipynb All come from dbgap, would be good to find another.
Question of synthetic vs real data.
In the CINECA dataset, looked at how they generated synthetic data. Looking at the European dataset it uses a tool called data synthesizer. Could put the data in data connect, less sure about how to do it in FHIR.
SV: Should pick one real world, on synthetic datasets. Get familiar with the servers and write down what the asthma phenotype looks like. Don't need to be exhaustive.
IF: Talking about a representation that will have to hit upon diverse data. Establishing the framework for FHIR and Data Connect where we can do it.
MC: With the dbgap data, that is controlled access right? We cannot use the data itself.
IF: Two routes: one is generated synthetic data, and the other is getting access to the data. Generating synthetic data is not trivial either. For general research use we can get access, check if we can get access.
KH: I'm on the data access committee, so I can help write a request.
IF: We've done that with the NCI DAC. The thing is it might not get us any asthma datasets.
MC: For CINECA it was a lot of back and forth with the cohorts. Not straightforward.
IF: For these dbgap datasets, can go to any of these and see summary reports of the data. Publicly available at a summary level.
SV: We have an explicit point on synthetic data. For dummy data, which is from statistical profiles, if synthetic using a toolkit, we cannot distribute that, has to go through access committee. There are existing datasets out there, can't we just use that.https://github.com/synthetichealth/synthea. https://simulacrum.healthdatainsight.org.uk/.
Purpose of the hackathon is representation, not generation.
IF: Plan that for the kickoff, put the data in.
FJ: SIngle data connect server? COnnect server to multiple sources?
IF: Hopefully do that with FHIR as well. When we talked about this before, one of the things we can do with this representation is that it is something we can aggregate data using. Let's focus on asthma, identify a number of sources of asthma, and can this representation be used against the real sources out there. Question then became - if you take those two sources, what is the cohort I want to pull together from those two sources.
SV: We already have a list of asthma phenotypes from phekb and others. With philip's help, define the representation, we have the synthetic data in FHIR and in connect if possible, and

see if we can extract those phenotypes. For the hackathon: We take existing synthetic data, explore how to represent the conditional cohort representation.
SF: Scalable vcf has encountered many challenges over access to data. Agree with what SV is saying.
IF: The DAC that KH has referred to, will turn this around in 48 hrs.
SV: If we can turn it around, lets do it. That doesn't change our plans.
IF: Just cautioning about being over conservative.
SV: Could be a session in the hackathon, what were the concerns of the DAC, what swayed them to allow the access for the purpose of the hackathon.
IF: Did nothing to change the existing access process, what was changed was the people on the receiving end.
BW [from chat]:Applying for access to the data is just as valuable as another technical aspect. We should just document to the participants or future readers how to get access. It reflects a reality.
SV: Suggest some concrete steps. If you can submit that request on behalf of us and get some of the learnings that's one thing we can do. Simultaneously, have some of the synthetic data pre-prepared in case we need it. Another action is to set up FHIR servers and data connect servers. I'm happy to host the FHIR servers and try to spin it up some near Montreal. If someone can spin up Data Connect that would be good. Need to load in these datasets into both servers.
IF: Data Connect will likely be spinning up an instance in a Docker container on your local machine.
SV: Those joining in would also want to have a publicly accessible endpoint. Spinning up in the cloud will be better so all can join in the fun.
IF: Would like to see the two synthetic datasets.
SV: Will spin up the datasets into FHIR.
IF: I can spin up Data Connect. If there's tab-delimited file for the datasets it will be pretty quick.

IF: What about Beacon?
SV: Maybe start thinking about an adaptor
JR: I was believing that we will use the Beacon interface, not SQL interface. If you have a good backend, put SQL infront of that, and use it. What we are testing is the feasibility of generating cohorts using a set queries. If this is the case, would like to use CINECA synthetic data behind the Beacon we already have and test the queries.
SV: So, we would have a Beacon backed by a CINECA datasets we could use to test asthma phenotypes. And then, another option is to have a synthetic population within a FHIR server, and in a Data Connect server. Another option is a real-world dbgap dataset in a FHIR server and/or data connect. Is it possible to spin up Beacon in front of a dbgap dataset?
IF: That's why we need a hackathon. My question to Jordi is how do I stand up a Beacon.
JR: Lighting a beacon is one thing, populating is another thing, so might be too much for a hackathon.

SV: Have one Beacon server with three datasets, one real, one synthetic, and maybe one other. Then have a Data Connect server and a FHIR server.
KH: For real-world dataset, are the Beacon servers secure?
JR: Depends on how you implement it. Can make it open or private. Can make it have boolean answers or counting answers. The query seems more precise with what we want from the cohort.
IF: What we're going to have to do with this is make it up as we go along.
JR: I see the value of real-world, but see more problems. I'm not demonstrating the quality of the data I'm demonstrating the query.
IF: A reality that we're never going to get away from, is that we may not be able to pull individuals out of the cohort environment. That's a separate question from how to send the query to each of those to find out if they have patients.
SV: What we're testing the latter piece, how to define a cohort, how to set up infrastructure to accept that query, process it and provide results. Not the underlying how you access data, get security, etc. That's another question for when we get there. 1) What does it mean to define a computable cohort definition, 2) how do you submit that to a service that will accept and respond appropriately.

IF: Jordi, where is Beacon with access control?
JR: Yes, you can have a secure access Beacon. Reference implementation has that. OAuth token.

SV: List of actions: 1) Ian to inquire about access to dbgap data
IF: Subject to Karen's message in chat [just throwing this out there: another complicated part of getting the controlled access data is that all collaborators have to submit a Data Access Request (the Research Use Statement, etc. can be the same, but all listed collaborators that will have access to the data will have to individually submit a DAR). Again, not sure within this Hackathon framework how many people will be accessing the data, but is something to think about.]
SV: Drop us a note in email with the application draft.
2) Me or others to get a hold of a synthetic data set. 3) is set up data connect (IAN, FRANCIS), FHIR (ME) and Beacon servers. Loading data can be part of the hackathon.
Action on Jordi to set up CINECA Beacon.
Action on Philip and I to do some pre-work on cohort definitions we have and some pseudocode for Asthma.
Finish putting the CINECA datasets in Data Connect. JR can put one dataset in for sure.

SV: Pick one datasets for sure.

IF: Dbgap on FHIR?
SV: Push out for the next hackathon.

IF: We have one hackathon extending over roughly a few weeks.We have one beginning session planned for Connect, need to plan the follow up.

Brian can help with the server, but not setting it up.

2022-03-24 - Pre-Connect Hackathon Planning

Chair: Melanie Courtot, Susheel Varma, Lindsay Smith
Attendees "Name (Project Affiliation)": Brain Walsh, Subhashini, Marion Shadbolt, Karen Haase, Ian Fore, Philip Van Damme, Francis Jeanson, David Hansen, Jordi Rambla, Peter Robinson

	Actions Arising	Assigned To	Deadline
1.	Check with Susheel about the HDR UK FHIR Server	Lindsay	
2.	Checking that the servers and datasets are usable	Planning team next meeting	
3.	Set up a Github Space	Lindsay	
4.			
5.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping - GA4GH Connect Registration - April 20-22 - GA4GH-wide Strategic Roadmap feedback opportunities - Town Hall Meeting Registration - Community Survey	Lindsay Smith	5 min
2.	Planning a Hackathon at GA4GH Connect Planning Document Hackathon Roles 6) Confirming the goal	ALL	50 min

	7) Date/Time - I did not select a specific day or time block yet, but please let me know if there are any preferences - How many hackathon meetings should we schedule? 8) Assigning roles and tasks a) Session Chairs? 9) What preparation is needed? a) Selecting a phenotype/definitions b) Selecting datasets c) Would setting up a Github space be useful?		
3.	A.O.B.	ALL	5 min

Minutes:

IF: If we need to query across different sources, in terms of data models of specific attributes, that actual data model is going to be different across the different sources
FJ: Interesting during the hackathon to explore that further
PvD: CQL is data model independent. Can we make this definition using several data models
IF: How does CQL let you discover the model to do a federated query. Same thing with data connect. If we're looking at different tools or representations, but working with the same substrate (sets of data). Question of real vs synthetic but maybe we can do real. A key part is having a diversity of data sources, in Phenopackets, data connect, FHIR/CQL. To understand how those different technologies help us deal with that problem.
JR: We were talking about trying to gather a cohort consisting of records from different sources. A believe a hackathon on this is enough. If we need to transfer into Beacon, that could be part of the hackathon. Get something back for research.
IF: To keep it grounded, the thing we start with, can we find 3 or 4 different sources from which you'd want to assemble a cohort. Each source is a pre-defined cohort. Work bottom up. How would you do the query across.
JR: Once you know one data source is having people with asthma and other different features, how do you go farther and pick the individuals. If the return is in phenopackets at that point, great.
PvD: Everything we are doing should contribute to a new standard?
IF: Not a guarantee
More to understand the problem, and how things relate to that problem.

JR: We have asthma samples in CINECA.
IF: We can take some dbGAP data and Beacon data and create
Susheel has some information

Need to check if datasets up and running
IF: If there are 4 different datasets, what does a computable cohort need.

IF: Can bring some datasets up through data connect, and beacon, what can we bring up through FHIR. There is not a FHIR server to use for those studies.
BW: Can follow up on the dataset we have. Been working on expressing this in ANVIL as FHIR. It is dbgap, or should be. Phs000920 DS-IRB-COL.
IF: Any of those data dictionaries we can bring up through data connect as well. If you have a FHIR server, you can search through the rest interface. What else do you need to bring up?
PvD: I haven't used CQL extensively, but what you can do is define a query or expression which can be translated to FHIR rest API calls. Can completely integrate that in FHIR, use the library resource to store that library, which you can exchange to another server.
IF: What we need to set up is a CQL, need a working server somewhere. Who is going to set it up, how is everyone going to access it.
PvD: CQL would not be a server, but a way to access to logic. Can have a FHIR server that support CQL.
If we don't have a server, if we have data in Phenopackets format, there is an IG for Phenopackets in FHIR. Could potentially map that to FHIR data as a proof of concept.
PR: The IG itself is finished. Been trying to implement a HAPI FHIR app.

BW: Good points about the lack of research-oriented open source community. Been working on FHIR IG for research project with KF. Google, Amazon, Azure have a service that you can quickly throw up.
FJ: HAPI is the most widespread but not the most well supported.
IF: The KF FHIR server is up and available for testing.

Controlled access data so would have to create copied datasets from them.

IF: There is work to be done, but feasible to be done, set up a Data Connect server. For beacon that also sounds possible. It's less clear with FHIR.
PvD: We have some data, we take a data model, or two, and then put that data in the model, try to express it in using CQL, write a couple of queries based on the data that we have. Then we have the first cohort rep in CQL.
IF: For a fair test, data has to be kept federated.
PvD: We have different datasets of different cohorts. Want to find some way of representing those cohorts. Want a common way of doing that for different datasets.
If we take the data, see how data fits into the model, get that definition using CQL, I don't see how we could do it differently. Seems like we are combining datasets, but I don't think that would be what we would do. More creating the definition of that cohort in a machine readable way.
IF: From what you described in CQL, each of the servers has to translate that. We're not querying on core elements that are in every study.

IF: Creating a CQL library. Does this mean we execute these queries? We should.
Will try and make sure I can take what we've said and create synthetic data.
DH: Building out a pathling. Trying to see if my guy is available to contribute. He uses Synthea to generate synthetic FHIR data.

KH [from chat]: Another question for kickoff is how many people would you want to join in the hackathon, what is the ask for people and what kinds of expertise is missing from the group that would be important to have coming into the hackathon.

IF: Who are the people that can assemble the cohort we're talking about.
FJ: Definitely by the end of the hackathon.

IF: Pre-check that the datasets/servers are usable.

2022-03-15 - Full Group Meeting - Hackathon Planning

Chair: Melanie Courtot, Susheel Varma, Lindsay Smith
Attendees "Name (Project Affiliation)": Ian Fore, Francis Jeanson, Karen Plevock Haase, Michael Baudis, Philip Van Damme, Melissa Haendel, Laureen Fromont, Shubashini

	Actions Arising	Assigned To	Deadline
1.	Circulate Doodle poll for meeting times on March 24th and April 4th/5th	Lindsay	
2.	Submit form for Connect workshop	Lindsay	
3.	Search & gather additional datasets that can be used in the workshop	ALL	
4.			
5.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping - GA4GH Connect Registration - April 20-22 - Workshop proposals due **March 18th** - GA4GH-wide Strategic Roadmap feedback opportunities	Lindsay Smith	5 min

	- Town Hall Meeting Registration - Community Survey		
2.	Recap: - PHENO report back slides - MIACC report back slides		5 min
3.	Planning a Hackathon at GA4GH Connect Planning Document 10) Confirm that we want to host this hackathon 11) Defining the goal 12) Date/Time? - For the form, we need to select the preferred day(April 20,21 or 22) and time block (14:00-17:00 UTC or (20:00-23:00 UTC) 13) Assigning roles and tasks a) Session Chairs? 14) What preparation is needed?	ALL	45 min
4.	A.O.B.	ALL	5 min

Minutes:

SV: Next hackathon could be MIACC
IF: Question about whether the hackathon is a good fit for Connect, hackathons take place over a sustained amount of time. People often get together 2-3 times. Going to be virtual either way. Kick off at Connect?
MC: Advantage of Connect is that people tend to book time off.
IF: Plan the whole thing at the Connect meeting.
MC: Have a consolidation, kickoff meeting, dedicated meeting to sort out the plan, and outcome of that session is a firm plan
IF: Key word is dedicated. Kickoff at Connect, draw them in, but want them dedicated and committed to doing work
MH: I like the idea of a kickoff meeting at Connect. Use the meeting to define the goals and what the real work would be.
IF: The things that are in the Hackathon doc, the goals, move in that direction by saying these are the ways a cohort would be created out of a number of technologies. What is the representation that can be executed against those. Take some actual asthma datasets and say, how would you do the cross node asthma cohort query and what would it take to represent that.
MH: Basically, take know asthma computable phenotypes and compare, but the hackathon would be to execute those queries and have some criteria with how well they execute?

IF: Understand the problem of execution and look at what is involved in creating an assembled federated cohort.
SV: Refine the document and write out the list of options
MH: I like the example of comparing existing computable phenotypes. Ultimate goal is to come up with generalize to generate computable phenotypes using rules. Take a set and look across. Been some great work being done with cardiac phenotypes. Derive commonality in representations even though there may be nuance in implementations.
IF: The task is to take a selected set of asthma cohorts and the task is how would you put a cohort together from across those cohorts. Take real cohorts that exist.
SV [from chat]: The outcomes for the hackathon: 1) Crosswalk asthma phenotypes to represent an ideal asthma phenotype 2) Specify the phenotype in an executable format (CQL etc) and 3) test the standard across multiple implementations in a federated manner. Each federated implementation with test it against a real dataset to evaluate what is needed to expand the standard.
SV: Action 2 - launch at GA4GH Connect and have a hackathon spread over two weeks. Through that launch event refine that further.
IF: CINECA datasets Melanie has suggested, there's some links in the doc that show example of what we can do. I identified the asthma cohorts. If asthma is the idea we want to take, perhaps there are some in EGA.
FJ: Cohorts that we're thinking about? HDR UK?
SV: Start with two, one synthetic and one actual. Can use one of MH's standard queries as well.
IF: If this is going to be realistic, have to pick cohorts, which when brought together, make some sense. Doesn't have to be, but helps if it is.
MC: So firm plan based on FASP, at Connect. Need to specify the outcome. The goals right now are a little bit fuzzy. Have a clearer outcome.
IF: With FASP, have to make sure the goals are reasonably achievable. The planning is to be realistic about that.
MC: Clean up the hackathon planning doc, present at Connect, formalize contributors and such.
IF: Have something more concrete than working on the planning. Plan the other steps after Connect.

Maybe meet every couple of weeks.
IF: Still put something on the calendar for Connect. Would it make sense to have more before Connect.
FJ: Get a few more meetings on the calendar prior to that.

Filling out the Connect form?
KVP: Generally report out of discussions thus far, discuss planning for hackathon.
IF: If kickoff, goals would be to have a practical exploration to drive what a computable cohort rep would be. Goal is to do some practical and identify comptable cohrt
MC: Goal would be outreach as well.
[Lindsay to fill out the form for Connect]

March 24th and April 4th or 5th
Lindsay to make Doodle Poll today and circulate, complete by Friday

IF: Looked at whether there were asthma datasets we could use in the hackathon. SV has mentioned CQL, I don't know if on this FHIR server I identified (dbGAP) is using CQL or not. The question is would ask, is are there datasets like this, sitting in FHIR servers or somewhere else that could be merged. Each of these 9 datasets are already available via FHIR. Also easy to make available through Data Connect. Which studies could these be combined with. This lists the fields in there. I don't know if we have creatinine level in there. Data dictionary that lists the variables available from that study. How do I query FEV1 across those cohorts and get a mega-cohort. This showing the schema as well.

The dataset we just looked at can be put into Data Connect, I don't know about FHIR. The question is, if you pull data together from these cohorts, what's the useful query you can do across them. More than just thinking about representation, looking at real data and making sure our representation works. An effective hackathon needs to identify data of this kind.
MH: Seems to me our goal is to query for the same set of CDEs, where we want to find some geno-phenotype association. So if we could come up with some scientifically valid reasons. Suggested choosing one that had an easier association. If we had a set of disease-phenotype variant associations that were knew were scientifically valid, work backwards from variant association.
IF: There's the big goal, and the goal for the hackathon. Does this stimulate anyone, dataset to combine with the things we were looking at?
MB: Shaping the multi-parameter definitions as subsets of phenopackets in a query language and see how we can send off to different protocols
IF: With CINECA data, can take that data and ask what it looks like in a phenopacket. What does that look like in a Beacon
MB: For Beacon, want to test phenopackets as a query language.
MH: A good part of why we conceived the working group was to think about the individual case to the cohort level. Number of different things that we need. Need to query for a set of CDEs. There's also rule based definitions, machine learning definitions, and metadata tagging strategy. Comes back to let's just define the scientific goals and think about the many different ways of answering that goal.
IF: The FEV1 in one place won't be the same as FEV1 in another. Question is are they really common, is it achievable to be common, and if not, how do we cope with that.

Action item to find some additional cohort examples

MH [from chat]: Example diabetes rule-based phenotype: Computable phenotypes are most often "rule based"; i.e.,"Define 'patient with diabetes' as any patient with >=3 ICD-10 codesof E08-E13 in a two-year period AND one of the following:(1) a prescription for any diabetes drug, (2) a maximum HbA1c >= 7%"
PvD [from chat]: I have been looking into the work presented in this paper: https://onlinelibrary.wiley.com/doi/full/10.1002/lrh2.10233, I imagine one of the outcomes is an experiment where we use the asthma definitions, try to represent those in CQL, and then see if this can be executed against OMOP and FHIR

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

2021-12-14 - Report Back Meeting

Chair: Melanie Courtot, Susheel Varma
Attendees "Name (Project Affiliation)": Judit Kumuthini, Brandon Wright (NIH), Ian Fore (NIH), Philip Van Damme, Kees van Bochove, Francis Jeanson, Heikki Lehväslaiho (CSC, Finland), Karen Haase, Tony Brookes, Michele Mattioni (Seven Bridges), Ana Leon, Katie Vance, Shimon Rura, Zhenyu Zhang (GDC), Brian Walsh (OHSU), Robert Carroll (VUMC / AoU), Grant Wood, Subhashini, Nicola Mulder (H3Africa), Adriana Malheiro (NIH), Mogomotsi Matshaba (H3Africa), Mamana Mbiyavanga (H3Africa), Nuria Queralt Rosinach (EJP-RD), Audrey Hendricks, Michael Baudis

	Actions Arising	Assigned To	Deadline
1.	Plan for a Hackathon in late January - join FASP planning - Doodle Poll - Topic Planning - Hackathon Process Information	Ian / Susheel / Kees / Lindsay N	ASAP
2.			
3.			
4.			
5.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping - Beacon v2.0 is now open for public comment. See the main page of ga4gh.org for links to the specifications. - REWS is finalizing their 2022/2023 roadmap. Many projects may be of interest to Clin/Pheno. Add your +1s here to be added to mailing lists in the new year	Susheel Varma / Melanie Coutot / Lindsay Smith	3 min
2.	Report Back: PHENO Team - Slides	Susheel Varma, Kees Van Bochove	20 min

3.	Report Back: MIACC Team - Slides	Judit Kumuthini, Jordi Rambla, Lauren Fromont	20 min
4.	Next Steps	ALL	15 min
5.	A.O.B.	ALL	2 min

Minutes:

PHENO
SV: Trying to create a standard to represent the phenotypic information of a cohort. Broken in MIACC (presented later on), looking at the higher level, and Pheno, looking more granularly. Aiming to leverage existing standards.
Motivation - did a landscape survey in the summer. Collecting who's doing what when where and how. Lots of data models out there. Main pain point that always comes up is lack of interoperability. One of the requirements we have is a standard that goes across omop/fhir/cql, etc.
What DPs have requested - have put down those who are interested or who have contributed in some way. At some point, both of these standards will come together.
IF [from chat]: Add Cancer Research Data Commons to the list]

SV: Looking for patients with a particular phenotypic value, looking at a row view. There are interactions with the standard with HL& FHIR, OHDSI OMOP, community standards like PheKb, HDR UK, N3C
Asthma example from phekb.
OMOP also represents cohorts.
KvB: Approached these problems by precisely defining cohorts with entry and exit criteria. Assume you have a database or set of records where you have persons with phenotypic information. Screenshare shows how the OHDSI tooling supports that definition - user interface to define criteria, rules, specify the time constraints.
SV: FHIR representation. Groups can be defined, a grouping of conditions or observations. There is a possibility to use CQL, originally defined for quality measures but one of the ways to define quality measures is by defining a cohort. It brings us close to clinical healthcare systems, rather than working with or creating something outside.
Strawman reasoning - choosing CQL as a built in abstract data model. If we take this there's a potential to extend to clinical trials. Operate with the Beacon network.
Would like to propose a hackathon in the new year to try and run a harmonization effort.
Link back to MIACC, link to Beacon and Data Connect. Expose endpoints to release phenopackets and DRS as v2.
Timelines - complete a summary report of hackathon in February.

Q/ Looks like you're more interested in the query? Or the modeling as well? Not sure how the query language, why do we need that if we have data connect, search, beacon.

A/ What we're aiming to do with CQL is overloading as a data model. Data connect doesn't provide a data model.
SV: Those interactions we'll explore as part of the hackathon, how to integrate with Beacon and Connect.
RC: Interested about the spectrum. Will you have objects for DRS?
SV: Specific connection to DRS was more of a pull from the FASP side of things.
RC: We're investigating how to link over to DRS references in FHIR. Natural bridge point, sounds great.
IF: We keep gliding over this thing of Data Connect and data models. Connect doesn't do models because there's a diversity of what we need to express in phenotype. LinkML, CRDC and CD2H specifically settled on LinkML having looked at FHIR as being fit for the purpose. They were looking for the modelling ability. This is a social thing asking can we do what we need to do. Would like to setup that we go into this more later. FHIR did not have the capability to do what they needed, so we should ask them about their experience.

MIACC
JK: Still working on the product description and main aims. Looking at the minimum information required to describe what a cohort is, with the level of granularity needed to answer end user's questions.
Have evaluated existing solutions like Maelstrom, IHCC Cohort atlas, DATS, OMOP.
What DPs have shown interest - ELIXIR, H3Africa, possibly CanDIG, EGA, SPHN. Other contributors could be CINEAC, IHCC, Maelstrom group from McGill.
Milestones - 1) Defining cohorts, finalize standards. 2) Consolidate resources 3) Identify DPs to test this out. Hopefully to have this done in the next few months.

SV: One of the things we both are both cognizant of is what is the overlap between the two groups. Want to use that to generate high level or summary statistics. How will the hand offs work - do we push to you or do you pull from us?
JK: Great question, don't have an answer per se. Once we finalize which model and get feedback we'll better understand what is most suitable.

Next Steps
SV: Two directions of travel. Hackathon in the new year to get a handle on the practicalities. Want a real-world example.
RC: Makes sense to try and leverage existing standards.
IF: Hope it would include some of the KF resources
SV: One of the arguments for using CQL is that the amount of work is exactly the same. Leverage the wider community rather than doing everything on our own. FHIR may not be the right level of representation, which is something we can feed back to FHIR.
KvB: Would be good to get together an organizing team.
Aim for FASP hackathon in January.
IF: Share the Doodle Poll with group. Certain amount of organizational effort behind this.
[Lindsay can share the hackathon materials]
SV: Happy to volunteer some synthetic data. Set up an OMOP server. Need people willing to learn how to write up a CQL statement.

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

FASP - Planning document:
https://docs.google.com/document/d/1DOQW5y3XieUY6lrD6JNkI7nlE-lWDq0-QJz4e7YReNw/edit#
Planning for the end of January. Two week window.

ZOOM CHAT

00:56:25 Ian Fore: Please add the Cancer Research Data Commons to the driver project list
01:02:16 susheel.varma@hdruk.ac.uk: Will do
01:04:15 susheel.varma@hdruk.ac.uk: Also interesting to think of outcomes as a domain for clinical trials. I suppose these could come back as future conditions
01:14:58 Ian Fore: Tony - the aim was indeed to represent the query. Cohort Representation. As you well know that can't be isolated from model.
01:15:08 Brian Walsh: * Is there a technical document that describes CQL ?
01:15:31 Robert Carroll: https://cql.hl7.org/
01:15:49 Brian Walsh: * Given a set of FHIR data (https://anvilproject.org/) - is there a implementation guide on how to implement ?
01:15:55 Brian Walsh: Thanks!
01:15:57 Michael Baudis: Data connect maps stuff to SQL etc. but uses external standards.
01:16:37 juditkumuthini1@gmail.com: Susheel you have 10 mins more Q& A
01:17:36 Kees van Bochove: Good questions Tony! It's good to know that CQL is actually more than a typical query language. It provides a way to plug in various data models and harmonize those. And there are already FHIR models maintained for it as well as OMOP models proposed (see https://github.com/PheMA/cql-on-omop/blob/master/docs/DESIGN.md for some more depth)
01:19:28 Michael Baudis: It's more biolink w/ LinkML for the code IMO...?
01:20:47 Kees van Bochove: Either way, if you want to have portable / interoperable phenotype definitions that go across often used standards like OMOP, FHIR etc. then you will have to express first of all what that generic model looks like (e.g. in LinkML) and how it maps to data standard specific cohort expressions (e.g. OHDSI phenotypes, FHIR queries etc.).
01:20:56 susheel.varma@hdruk.ac.uk: All great questions. Happy to discuss in the chat or offline
01:24:01 Tony Brookes: I agree Kees. What, in your experience, would be the main things that general model needs to provide beyond the obvious cohort context/metadata stuff and the widely available structured semantic richness of medical ontologies?
01:24:12 susheel.varma@hdruk.ac.uk: We are not set in our ways long-term whatever works to represent a cohort in a standardised/interoperable way. Our steer to reuse CQL was to using existing standards and leverage existing tooling to reduce burden. It may not be perfect, but as always cognisant to the fact to not let the perfect be the enemy of the good. We are happy to work with either LinkMl, HL7 and OMOP to chart and roadmap forward
01:25:31 Robert Carroll: If there are gaps, it's possible to engage with these communities as well to make sure they are covered in the future. I'm leery of "starting over" with new models every time we hit something that doesn't go perfectly with the existing spaces.

01:25:54 Tony Brookes: I see the attraction of CQL if it lets one represent information and query it in a standardised way. But there are also arguments for keeping the model part and the query part separate, so groups can adopt either or both and plug into whatever else they may be using
01:26:47 Brian Walsh: For data models at this level (cohort) and for this use case (identify, query and download) the FHIR standard is sufficient and probably represents the shortest time to market. Consumers always have the option to integrate into their own analysis pipelines.
01:27:22 susheel.varma@hdruk.ac.uk: @Toony - I agree We just happen to use a query language for modelling purposes as well. SQL has DDL too 🙂
01:28:33 Michael Baudis: Years go by ...
01:28:38 susheel.varma@hdruk.ac.uk: I would argue the CQL is more a modelling language rather than a query language based on my experience. Maybe @Phillip would be able to help as our resident expert 🙂
01:29:24 Robert Carroll: I don't know much about CQL, but it sounds like a reasonable approach in that one could map over ideas using linkml, leverage the hl7 tools for fhir, etc.
01:29:24 Michael Baudis: We need a cat logo for MIACC...
01:31:22 Kees van Bochove: @Tony if you ask me I would start with the entities used in OMOP cohort definitions, which is one of more elaborate takes on creating reproducible phenotype definitions, and you will find analogous concepts in FHIR and other standards. In the slide I already made a start: Cohort (with sub-entities Cohort Entry Event, Inclusion Criteria, Cohort Exit Criteria), Concept Set. Which in turn use Domain (maybe we should leave that one out), Concept Code (as Ontology - Identifier tuple), Relative Start Date, Relative End Date etc. See https://ohdsi.github.io/TheBookOfOhdsi/Cohorts.html
01:31:46 Tony Brookes: Susheel - Has the decision to use CQL already pretty much been made? I am too new to this to have a view. But it would be good to know. [just glad you're not suggesting RDF though]
01:32:04 Ian Fore: “Generic model” could be what one might call it, but it doesn’t really define it’s a model at a different level that was needed.
01:32:47 Michael Baudis: @Tony RDF ¯_(ツ)_/¯
01:32:56 Ian Fore: @Tony - that doesn’t have general assent from the group. It’s one of the things to look at in the hackathons.
01:35:53 Michael Baudis: Well, Phenopackets was something very different in the first format; adopted a lot of the GA4GH (metadata) etc. thinking during re-write (and GA4GH metadata was partially driven by people active in phenopackets development ...).
01:35:58 Tony Brookes: OK. Thanks for being clear. It will be tested first.
01:36:04 Robert Carroll: For All of Us, we are getting close to a fully synthetic dataset that we can share. I'm not sure on timing yet, but it is an OMOP-centric multi-site and multi-source dataset (eg, EHR data, Survey data, etc)
01:36:35 Ian Fore: @Tony - the design of the tests is key.
01:36:49 Tony Brookes: Given Kees' suggestions, this groups ambition seems very achievable
01:37:08 Brian Walsh: thanks very good discussion
01:41:45 Ian Fore: https://docs.google.com/document/d/1DOQW5y3XieUY6lrD6JNkI7nlE-lWDq0-QJz4e7YReNw/edit#

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

01:44:55 Ian Fore: This is the higher level spreadsheet for the hackathon https://docs.google.com/spreadsheets/d/1LSDG1jbIg2dKWhYi29v5fcSFCWtltEjNtT5omxTshNc/edit?pli=1#gid=729423797

2021-12-02 - PHENO Team

Chair: Kees van Bochove, Susheel Varma
Attendees "Name (Project Affiliation)": Ian Fore, Vanessa Vogel-Farley, Nuria Queralt (EJP RD), Francis Jeanson, Karen Haase, Philip van Damme, Judit Kumuthini.

	Actions Arising	Assigned To	Deadline
1.	Complete Slide Deck / Roadmap for the report back meeting	ALL	Dec. 14
2.			
3.			
4.			
5.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping - REMINDER: Report Back meeting Dec. 14 @ 6am PT / 9am ET / 2pm GMT / 11pm JST / 1am (Wed) AEDT	Susheel Varma / Kees van Bochove	5 min
2.	Using FHIR to support Cohort Definitions - Clinical Reasoning module (e.g. expressions) and CQL	Philip van Damme	15 min
3.	FHIR - BR&R Eligibility Criteria Update Slides	Ian Fore	15 min
4.	Updates on OMOP & CQL?		20 min
5.	Finalizing the Report/ Roadmap / Presentation	All	45 min

Minutes:

The MIACC group is looking at cohort registries in general, how do you find certain attributes, answering questions like do you have a hispanic ethnicity dataset. We're looking at the deeper query, can you give me the count of patients who are hispanic and have this blood count level.

HL7 FHIR and Disease Cohorts
PvD: My research focuses on semantic interoperability, mostly terminologies, FAIR data and FHIR. HL7 FHIR stands for fast healthcare interoperability resources. Fast because fast to implement, free and open source. Goal is interoperability, and the resources are the individual packets that FHIR is based on.
Main components to FHIR: the data model and a restful API. There are also open source tools, FHIR servers and community. FHIR defines ~150 resources. A resource has subelements with different types and cardinalities. 80/20 rule = the spec covers only what is needed by 80% of the systems worldwide, 20% for specific use cases
Profiles - taking resources and constraining them, eg. change cardinalities, remove elements
Extensions - adding new elements
Implementation Guides - rules for how resources should be used.
Mapping Common data models in FHIR - there is a tool called CAMP FHIR, it gives you a way of defining mappings between a CDM like OMOP and FHIR. CAMP FHIR provides a way of defining the mappings and transforming to FHIR.
OMOP on FHIR group is also doing mapping to FHIR.
Defining Cohort/phenotypes - if something is a diagnosis, use a condition. If something is a symptom, use Observation. Once you do that. The group resource is a group of multiple entities. The characteristic can be a reference to a phenotype.
There are different ways of doing profiling, theres a repository based resource at simplify.net. I really like FSH (FHIR short hand), machine readable files.
CQL - Clinical quality language. It's based on libraries. Can define observations which will transform in the background into a FHIR API Call. The IG editor can take a CQL file and get you a JSON representation of the library resource.

SV: Excited about some of the things we can do.
KvB: Terminologies - it seems to be increasingly more popular to create terminology systems and use FHIR. Problem with FSH is that it's meant to carry FHIR specifically, so cannot fulfill the role of being a CDM across. CQL potentially could.
The OMOP on FHIR working group is very large. CQL came up there as well.

IF: I always thought the Q meant query but it means quality. Want to make sure it's not too heavily tied to one use case. The US doesn't have a national health service other than the CMS. They buy services from hospitals and their use case is, is that hospital delivering quality care.
SV: You're right, but to define a quality measure you need to define the characteristics of the patient.

BR&R - biomedical research and regulation in FHIR

IF: Most of the focus is in clinical trials, “regulated research”. BR&R is concerned with a lot more than just FHIR. Adverse event, research study and research subject are the key FHIR resources. Active in trial eligibility criteria.
Going back to what we mean by eligibility criteria. They looked at 27 different studies and identified commonly used concepts. Looked at both inclusion and exclusion criteria, boolean objects, etc. Temporal constraints. Looking at did a patient have an experimental drug within 28 days of the study. I think this group is actively exploring the same things we are. Focused on the group resource. That group resource is maturity level zero, this is open and fluid to suggestions. Great opportunity to get involved with this. Beginning to look at how to express a query inclusion/exclusion criteria in FHIR. Looking at FHIR as part of FASP. Examples of using NCPI FHIR Server. It is still challenging in FHIR to articulate a query based on ethnicity. Maybe this cohort representation query is also something to explore further. We are looking for use cases for the FASP hackathon.

SV: Can draw on some of the use cases from the Dec 14th document.

FJ: With ian it’s diving deeper into the data model. How do you extract knowledge from this unstructured data tha we get from this inclusion/exclusion criteria. Speaks to varying levels of complexity. At what level do we want to dive in given that there are other groups working on this? Ian, when you said there are challenges with querying (e.g. ethnicity), did you mean that from defining the query, or the fact that the data is unstructured.

IF: The first is a consequence of the second. You’ve got multiple levels of abstraction you have to go through. To get to the actual value of ethnicity, it’s hidden several levels deep. I think of the kind of process we’ve gone through over the years, it’s very much more complex in FHIR.

FJ: You think things like CQL would actually help

IF: Not sure, but we should explore it.

SV: We’ve got options on the table for next steps. Things we can feed into our December 14th document. We can identify the three options in the doc and pick which one we want to go through with

In the draft proposal we said we’d go through with CQL. Maybe we should write down our main motivations to do that. These presentations are a great starting point for that.

IF: We seem to have gone deep down into a particular tunnel, the FHIR tunnel. FHIR works well for healthcare data but it doesn’t work so well for research data. We have a healthcare use case but we als have research data to consider. We’ve started looking at the standards but other things have dropped off the map such as data connect.

FHIR is not great for everything. Need to critically consider if it's the right thing to do.

SV: One way to approach this is to write the problem and show the data connect team. From what I understood, data connect is a very general data query system. Whereas we need the semantics of the resource more accurately captured.

IF: More than 50% of what Data Connect is about is defining those resources semantically.

SV: I will try to loop in the data connect team.

IF: there are other GA4GH standards that are equally applicable (beacon, phenopackets)

FJ: LinkML is something that's coming up as well. From what I'm gathering, we're after a work flow rather than defining a spec that is some sort of technology output. If it's a work flow, we might be after defining requirements or constraints. At each stage of making your data available for cohort discovery you need to accomplish a certain set of requirements.

SV: I agree. For the 14th of December, our requirement is to define the problem, the requirements and share that with the wider GA4HG work streams. Then they can come in and tell us where they can help. We can be aware of the WSs but we don't need to bring them in right now.

IF: I'm not even sure that we would fully get the requirements defined by Dec 14th but we can certainly define next steps. The question is can we query for a cohort in FHIR, in data connect etc. We need to go beyond the question to get to the answer. Hackathon activities could be helpful here to take it out of the theoretical.

KB: Agree. Last time we said in the proposal that we'd use a crosswalk, take some concrete examples and express these in CQL. Take asthma cohort definitions…

Maybe one group of the hackathon activity would be to take this and convert it into CQL. Fist half could be conversion, and then execution.

PVD: Do you already store the data as FHIR resources?

SV: No, the raw data comes from a *data bricks layer, then we convert it to FHIR or..IT's a materialized view on the data. But for the hackathon I'll probably give you a synthetic version. We need to go into a road to travel. It will be weakly held and we need input to make sure that's the route we want to take.

KVB: Copied the description from last time (we will do a data comparison, cross walk, etc.) I also tried to make a reasoning for why CQL. Want to make sure we agree on it. You can express cohort definition in FHIR pretty well. Once you've done that you can store it in a library. But this is still ithin FHIR resources. The advantage of CQL is that you not only say I'm using FHIR but you also have to choose a data model. Philip, I noticed you used FHIR helpers in your example.

PVD: Yeah it transforms the CQL, so if you refer to something in CQL it knows what that is in FHIR.

KVB: If we could build a number of those data models, we'd have to build some for OMOP. that's partially done in the CQL and OMOP projects. We could use CQP and you'd only have to replace at the top of the file the definition of the data model and then you can use the same cohort definition. That's what I would propose for the hackathon.

SV: We are hoping to do this in the hackathon. Also harmonization. One work on creating a CQL with a FHIR CMD and then a CQL with an OMOP CDM and run it against the same data. Was thinking of Cynthia* and generating that as FHIR and OMOP.

FJ: Would the CQL be able to query the data connect data base?

KVB: Can you tell me how data connect works.

IF: Take Susheel's example. A sequel data store that's based on OMOP. Expose that table with semantic definitions of the table through a schema. If you had a column on smoking, the definition of that column would say it used that standard...Data connect gives them the necessary information to do what they need to do scientifically. There's no semantic mapping, it's describing it as it is. This column called smkoking can take these values. Or it could say this column called gender has 1s and 2s and that means males and females. Now as a user I know what that means. I can combine those two in the way I need for my scientific purposes.

SV: So this is telling you the range of values or enumerations.

IF: Ideally it wouldn't have to tell you any of that. It would be simpler id the data provider can say "we use this standard and here's the reference to that standard. This is a column that contains blood pressures.Go look at loinc and that will tell you what you need to know."

SV: The problem there is that there's no one to one mapping between column and semantic concept.

FJ: One of the advantages is it can abstract some of those databases. It removes one barrier.

IF: There's a FHIR example in data connect out of the kid's first data source.

SV: On the 14th we will let the GA4GH community tweak and change our proposal to make sure everything is alignedAnother question we need to answer is what are the driver projects willing to run with this.

FJ: Might have an implementation project here in Ontario between 3…

KVB: Going from community consensus and wor docs to CQL seems like a big step

IF: On our roadmap should we have some interaction with the BR&R group?

SV: Yes we should have discussions with anyone who wants to have a discussion around this. If the computable clinical trials starts picking up momentum that oculd be a driver project ina nd of itself.

IF: In a way it's going to the root because it's an HL7 working group and it would feed into their development and standards process. It would result in something that actually becomes an approved part of FHIR. Means that somebody else could do something for us. There's a process there which is useful to have.

SV: You know what we talked about on computable eligibility criteria, based on what we talked about I did try to do that. I reused a phenotype definition. Once you have that, that becomes your computable cohort definition.

IF: BR&R would love what you've just shown.

SV: the next time you meet with them, share what we are doing and we'd love to have them present.

KVB: The roadmap doc is coming together. The main thing I'm trying to wrap my head around is how we are going to condense all this into a 30 minutes slideshow on the 14th.

SV: I'm not too worried about that. I'll volunteer to take a first stab at it from monday next week.

KVB: We want to start with our landscape analysis and use the crosswalk for that. Then present our proposal, the reasoning, next steps, and also ask these questions about which driver projects want to support this.

IF: Be very clear when you say driver project support. What are you asking for from the driver projects?

KVB: You're right.

IF: You could say bring a dataset. Where could you find a dataset from a driver project that is real data.

KVB: We could ask them: Do you have this problem? Where do you have this problem? Do you have examples of it? Does it hurt that you don't have an exchangeable cohort definition that could go across data…

SV: We don't have a specific problem. What we have is that it takes a long time for people to go through the data access process. Move the access problem to a discovery problem.

IF: It's a meta problem. In a sense what we're trying to do is meta because we want to go across driver projects. One way of looking at driver projects is that they're the source of what we want to integrate. There are different ways we can key into a driver project, either as a data source or one that is struggling with this data integration problem

JK: Want to make sure what phenotype related to clinical phetnotpye is going to be presented. Instead of identifying at a higher level what is being presented, we are looking at the maximum level.

SV: There are two conversations. One around the semantics around the phenotypes. What phenotypes get implemented is actually driver project led work.

NQR: Wondering if EJP-RD could be a use case.

SV: Of course. Please reach out to them and see if they're interested.

IF: Yes, it is a very important use case for rare disease.

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

2021-12-02 - MIACC Team

Chair: Jordi Rambla, Laureen Fromont, Judit Kumuthini
Attendees "Name (Project Affiliation)": Ian Fore (NCI), Melanie Courtot (EBI), Isabel Fortier (McGill/Maelstrom), Susheel Varma (HDR UK), Katie Vance (NCI), Nicky Mulder (H3Africa), Michael Baudis, Rita Wissa (Maelstrom), Soichi Ogishima (GEM-Japan)

	Actions Arising	Assigned To	Deadline
1.	Set up Slide Deck for Dec. 14 report back, and follow up with everyone on content over the next few weeks	Lindsay	Dec. 14
2.	Contribute to the roadmap for the report back	ALL	Dec. 14
3.	Present the CCDI catalog at an upcoming meeting	Katie Vance	
4.			
5.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping - REMINDER: Report Back meeting Dec. 14 @ 6am PT / 9am ET / 2pm GMT / 11pm JST / 1am (Wed) AEDT	Jordi, Laureen, Judit, Lindsay	5 min
2.	Presentation: Cohort Atlas	Melanie Courtot	10 min
3.	Presentation: Maelstrom	Isabel Fortier	10 min
4.	Review of standards (referenced in DATS paper)	Ian Fore, A Beltran, P Rocca-Serra	10 min
5.	Discussion and Next Steps - Roadmap template - Slide deck version	All	45 min
6.	A.O.B.	All	5 min

Minutes:

JR: Determine if there is a solution to describe what a cohort is, the contents, to the level of granularity to answer the questions we'd like to ask. Today, looking at existing standards.

Cohort Atlas
MC: Atlas is work between CINECA and IHCC. Both are a consortia several cohorts, study cohorts. Need for a data harmonization pipeline. Are there data for specific criteria? Wanted that in a cohort browser, fed in from our registry. Info containing the semantics is not always equivalent. The Gecko metadata model is a high level overview of categories, capturing things like lab measures, cohort attributes, questionnaire data. Found from overlapping 10 cohort. Generic and high level, needs to evolve. To promote the standard using semantic standards using world wide web standards. The cohort registry - store the dictionary in the registry, gives us version control and the tools we already had in place. Automated mapping pipeline for cohort owners
The atlas was available online, have a filtering module where you can set country of origin or type of data. Easy interface for discovery, not looking at any individual level. We've built on CINECA, IHCC and GA4GH. Working to import Maelstrom in gecko. Next step was computable cohorts, gecko needs to be extended. OMOP would be an interesting model to use, can reuse the mapping and implement automatically. Have had issues with data being not available or sharable, in CINECA, developed synthetic cohorts for testing, available online. Could use here.

Maelstrom
IsF: Distinction between documentation of individual studies and the data they are collecting, vs the harmonized data. For each, we use existing standards, and are trying to be as complementary as possible with other groups. A few years ago we did a review of existing catalogue, found different layers of precision and the data collected. Maelstrom is just focusing on study description, overview. In the catalogue, have 290 studies that are partnered from 23 networks. The aim of the catalogue is to identify a study or studies that can answer a question I'm asking. To also explore potential to harmonize the information across several studies. Example of mother and child study, have different information about the design, list of PIs, possible to have access to data or samples, etc. Then we have a description of each sub population. The structure is that you have the study, x number of subpopulations, for each sub x number of collection events. We attached the metadata to each collection event, in each dataset you have a list of variables and a number of annotations for each variable. Will generate a list of variables, take each variable and classify. We have 18 categories with subcategories. Allows us to facilitate search.
Because we have the subpopulations, can visualize for each subpopulation. Important for co-analysis. Can also add aggregated information, like the distribution, but don't add on the website. Can extract the categories used for studies, be able to assess the ability to harmonize. Some of the categories were collected by all, as you can see.
Harmonization - have some tools and standards for this. The harmonized data need to be documented differently from individual data.

DATS
IF: Idea behind DATS, broader remit than just cohorts. Some datasets that needed catalogued would not be considered a cohort. Idea of DATS or data tag suite is to do what was done for journal articles. Schematic overview - level of abstraction that can make it hard to see the real world. A dataset is at the core of the model and attributes and related entries. Idea of dimension is covering the structure of the study, eg. if you had looked at 3 different drugs, the drug would be a dimension of the study. What those dimensions are depend on the experimental design. There is an ability to record the terms of access and the licenses under which data was available. Some effort was put into recording variables. The extended elements can look at molecular entities, biological entities. To give an example of how this was applied, the tool built on top was datamed - goal is to discover data sets across data repositories or data aggregator. Scope was broad.

Q/ How do you keep the numbers in sync as cohorts collect more data? Or are your collected data always retrospective?
IsF: Depends on the network, the money to support the update will depend on the cohort. For some networks, there is no more activity so the updates are not done as frequently. Adding new cohorts each month.
MC: So ask the cohort if they have any updates and they come back to you?
IsF: Generally it's coming from the network. Or directly the cohorts tell us. We don't chase.

Q/ How was the variable consensus were reached? Is there standard lists that were considered? Is it possible to get all the variables that were considered i.e. maximum? Finally, what were the rationale for the selection of variables you showed?
IsF: For the cohorts, just taking the info that they collect, really metadata or the way the information is collected. Collecting all of the data dictionaries and that is what we put in the catalogue. For the harmonization, we have two types of support - prospective harmonization, helping the cohort to develop the core variables, and supporting retrospective harmonization. Will not be possible to say "here is how to collect smoking" - depending on the question you ask, you might not want the same variable. Align variable to research question.
IF: Harmonization is a huge problem, one that we need to do some work on. My question was originally whether it's one thing to harmonize all the data within datasets, the other kind of harmonization is metadata, the data about the cohorts. What are the variables they measured. How do we record in a harmonized way, how a new variable is recorded.
IsF: Depends on the definition of the core variables. We have guidelines for harmonization and templates on how to document the core variable. R package that will be out next work to help the process of harmonization. https://portal.canpath.ca/
IF: What is the data model you use to harmonize?
IsF: Dividing line - we're not harmonizing the 16 different measure of smoking, but what we want to say is these are the 16 measure. The question is what is the data model for that database.

JR: We have been talking about different aspects of the continuum. Framework, categories/variables, and values. Then the harmonization process. IMO, the parts that we need to better understand are the framework and variables. My question is: If I want to go and access your catalogues, do I have an API where I can get the information.

IsF: Mixture, when we developed the catalogue there was not one standard that covered all that we wanted to capture. Mixture of existing standards, and when there was none, we filled in the holes with things we develop ourselves
JF: So no standard from the framework either, but you have some for categories.
IsF: Compatible with DDI where relevant, but we added more if needed. Tried as much as possible to use existing standards. Not 1, but 26 standards we used, but nothing is perfect fit.
MC: There is no API on the atlas. We are trying to do automated mapping, and for that, we store all of the mapping (https://mapping.ihccglobal.app). If there was a way to include generic models like OMOP, OHDSI, etc. those would immediately be reusable.
JR: Moving from gecko to omop, to fhir, etc. - are you doing that yet?
MC: We've talked to Robert Carroll at AoU, but not done anything yet.

Q/ Is it right to think of DATS like as bioschemas as more of a generic/structural representation of any kind of datasets?
IF: Yes it is more generic. I think that doesn't disqualify it. It seems to me we have a mountain of harmonization of data itself, but could also refer to it as a hole. Maybe we should be limited ourselves to the question of cohort discovery. Would be useful to see that data model about cohorts. If we can get to what does that data model look like and do them side-by-side with an use case in mind. Until we look at those models, this group has a simpler problem.
IsF: Documenting the data collected and metadata about when it was collected and at what time, all of that info is quite different from how to coanalyze smoking across studies. If you want to document and give easy access to what is collected by the cohort, if you want to try to fit with OMOP or CDISC that's something else.
JR: Now if you need to know about a change, you need to ask them or have them inform you. What we expect is something that happens with Beacon, describe the cohort in a standard way and light it up, everyone can see that information. Instead of mountain, it's a pyramid and telling everyone to bring the right brick.
IF: The point is that there still needs to be rules for the bricks
IsF: The quality is extremely important.
MC [from chat]: I really like the vision here of using Beacon as 'automated pipeline' and the point that we need count of data as well
SV: Trying to reuse standards as much as possible. Don't have a reliable way of cross mapping all of the different concepts across OMOP, FHIR, etc. Question I have is can we reuse existing standard to cross map. Is there a standard or mechanisms to represent information that you can translate to another standard. If that's too difficult, another option is working with OMOP or HL7 and invent a new standard as part of their organization. Regardless of the choice, need to write an adaptor. Look at all of the different ways phenotypes are represented.
JR: Cohort packets. Something that any solution could map to
SV: At the cohort bucket level, want to ask if there are cohorts with ethnicity hispanic, get a list of cohorts.

IF: The simplification is some way of saying yes, this cohort has ethnicity in it, but avoiding the standard definitions of what ethnicity is. Just want to know is there something I can use in there.
SV: Overlap there. There are two ways of defining asthma - one is by some clinician that has put in the ICD code, the other way is to find by lists of medications and predict you have

asthma. From the cohort registry, you just want to ask the question do you have asthma patients.
MB: Want to design a formatted interface that identifies cohorts.
MC: Trying to enable discovery of cohorts. Build the parameter Beacon could ask and what the cohort can expose.
JR: In Beacon, you can do the query, but also say give me cohort #1 information.
MB: Have an endpoint that exposes ontology terms, can also include the counts of occurrence.

MB: Middle ground, there is sometimes recommendations which ontologies should be used, not a prescription.
JR: We don't have a standard to describe a cohort. Have an idea of some activities and the standards out there. Outcome of meeting today is that a standard doesn't exist to represent a cohort that includes everything we want. Compare the models.

IF: The conclusion is that we don't have an answer. But we have several models of how to catalogue.
IsF: We don't have a standard, but we know how to document.
MC [from chat]: Maybe we can start with a small number of fields, try it out, and build up as we get more use cases. We don't have defined use cases yet?
IF: Need the use cases
JK: For Dec. 14, we will also get more input and make available in public domain.

JR: Present that there is no single standard
IsF: The thing with the cohort is that there is longitudinal data as well as the subpopulation

MC: Do you have a cohort descriptor, the type of study, whether its longitudinal?
IsF: Yes, all there and searchable. Have design, type of participants, etc.
MC: Model is available?
IsF: In the paper, yes. https://www.maelstrom-research.org/page/publications

JR: Need to define some use cases.
IF: My one comment is that we need to have some care about the models we choose. The model we need for this problem is not one that deals with the values.
SV: At that level, we can push to the other group. We already have cohort registries, so can we start harmonizing.

2021-11-18 - PHENO Team

Chair: Kees van Bochove, Susheel Varma
Attendees "Name (Project Affiliation)": Karen Haase, Ian Fore, Lauren Fromont, Alice Mann (GA4GH), Aina Jene, Peter Robinson, Vanessa Vogel-Farley

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

	Actions Arising	Assigned To	Deadline
1.	Create readout presentation Template - See here	Lindsay	Nov 19 - DONE
2.	Create slides for OMOP	Kees	Dec 1 - DONE
3.	Create introduction slides for FHIR	Ian Contacted Smita Hastak and Hugh Glover (HL7 BR&R co-chairs) they are starting working on trial inclusion criteria in FHIR. First discussion is on the agenda for the BR&R Nov 30 call.	Dec 1
4.	Add slides on CQL with Philip	Susheel / Philip	Dec 1
5.	Add slides on working plan / roadmap	Kees	Dec 1
6.	Review current draft roadmap and add ideas	All	Dec 1

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping	Susheel Varma / Kees van Bochove	5 min
2.	Using FHIR to support Cohort Definitions - Clinical Reasoning module (e.g. expressions) and CQL	Philip van Damme? Kees van Bochove	
3.	Which cross-translations would make sense for GA4GH to further look at, e.g. example of CQL on OMOP	Kees van Bochove	
4.	Next Steps / Roadmap Writing / Planning for Next Meeting		
5.	A.O.B.	All	5 min

Minutes:

SV: Phillip offered to do a short presentation on FHIR for our next meeting. This group is about how we represent phenotypic information. Previous meeting was about OMOP.
KVB: Last time we looked at OMOP, thought it would be helpful to look at FHIR. Looking from FHIR, there is a clinical reason. Scope of expressing some logic and querying data. There's a

list of use cases than help in understanding querying in FHIR. FHIR is just reusing another HL7 standard, CQL. Downloaded an example, definition of chlamydia.

IF: FHIR has this search interface. One of the key factors about FHIR it's an out and out REST implementation. Vanilla REST. How does CQL relate to that search interface?
KVB: Yes, the restful API and search API - I think search is a default action but doesn't release much data beyond the basic.
IF: SQL gets you into the full graph
KVB: Don't just want to formulate as a search query, want the search definition as well.
SV: Is the criteria you mentioned composable? If you were to develop a query, can I redefine another query and say that query that Kees defined is one of my criteria as well.Imagine you defined a search query for a set of diabetes patients, and I want to do another query definition based on diabetes and asthma.
KVB: You have the basic framework, then there are a number of entities supporting implementation. Just looking at this, if we dive into clinical reasoning, there are a number of common use cases of how to create reasoning modules.
IN order to look at quality, also need to create a consistent representation of a cohort of patients. Can use FHIR path, or CQL, or FHIR queries.
IF: Bigger questions here: they way we representation cohort, we want it to be independent of a particular implementation. There's also model, and these may be closely connected or not. TO reaffirm our goal, we want to represent cohort independently of whether is FHIR or OMOP or anything else on the backend. But can't get around in having a cohort definition representation, one has to in some way commit to a model, so do we adopt an existing model. What path are we thinking?
SV: Thinking about staging our resources.
IF: If looking at meta level, should be looking into Data Connect API. Leaving aside the bigger philosophical Qs, thinking about FHIR, we could draw from the NCPI FHIR initiative. They have a FHIR server up and running to access KF studies. In FASP, we've done some work with FHIR examples. But how to formulate a query in FHIR. dbGAP on FHIR has been doing this work as well. dbGAP has put a FHIR wrapper around their research studies.
SV: Great thing that we have a lot of FHIR initiatives we can relate to.
IF: If we're moving towards a presentation from FHIR, would be great to have their experiences
PR: Obvious resource out there is eMerge. Basic problem with defining cohorts in clinical data is using billing codes. eMerge is a set of rules that can be used in hospitals to identify cohort. Use billing codes to identify cohorts. Electronic phenotypic algorithms. Main difference between that and what I think we're talking about, is the emerge is PDF files on how to write queries in your own hospital.
SV: I think the differentiator is that we're in the middle ground between PDF and OMOP or FHIR. Is there a computable version of that PDF. Maybe the task is to feed back to emerge.
PR: Important to be critical when planning. I would say that the fact that there is emerge is that we're doomed to failure. Some much heterogeneity in what we're trying to do. If we want to get beyond that, need to define what is the minimum amount of homogeneity in data sources. Is there a transformation step.
SV: Our intention is not to invent a standard, see the state of the art of the possible. The report would say if it's possible, this is the route we much take. By December, we're not saying we must implement some computable standard.

IF: I think this focus of this is on North American health care records, but there's enormous diversity there. I feel like we wont be able to change the world and make that go away, but people need mechanisms to cope and find a way through that diversity. One solution is a cleaned up version, but I don't think that's something we can do. So, moves us in the direction of being realistic about what it will take. Set of resources that will help translate and help people understand the data. What's the tooling that can make those tasks easier.
KVB: I think the situation is changing over time. Using CDMs and standards is on the rise, see it well with OMOP. Not everyone is using OMOP. If you look at certain FHIR resources is not too diverse. I feel that there may be possibilities for us to output some work. We need to assume some kind of data model - was interested in looking at CQL. QUICK is the data model that relates it to FHIR. If we have a similar representation of the data model for OMOP, or OMOP version, then using CQL, could still define queries. Maybe CQL is the common ground we're looking for. IN roadmap, could assess using the same CQL grammar but swap in different data models.
SV: There will be a learning curve for those coming from OMOP to FHIR. I think the standards are on the rise, I agree. Possibility of proposing a roadmap here to look at decisions.
IF: It's not just a question that diversity is there because of different approaches, but biology and medicine is enormously diverse. There are specific variables.
Part of research is to say, this measure is a much better indicator of stenosis than this other one. There will be research done to validate that. We shouldn't disclude because it isn't the standard.
SV: Question again is can we make the definition computable
IF: We're just focused on making it computable, not whether it's a standard measure.

KVB: Thinking about what to propose in our roadmap. Omission that we didn't look at pheKB. We should look at Phekb definition in asthma, vs the one in OMOP. Does it make sense to create some intermediary specification. Maybe using CQL for that, and bringing in various data models. Build a concrete example would be useful for all.
SV: Yes, need to split out if we can make phenotype definitions comparable, and then, do we need a standard to make that comparable. First step is atleast being able to compare phenotype definitions.
IF: You have us the use case Kees is looking for, there are 8 definitions of asthma. We're going to allow those to co-exist, how do we represent those 8 in a consistent way.
SV: Problem we have is that we can't even compare among on our own. We have a list of codes.
We have ICD-10, have to use heuristics. Make due with what current state of the art is in use today, and the historic data. Don't have a measurement based approach.
PR: One of things important for GA4GH is statistics. I think we're entering uncharted territory - need to figure out the special angle. How to understand data well enough that's coming from multiple hospitals where we might be using clinical predictors.
SV: Roadmap doc will set that out. I like the action of looking at asthma end to end.
IF: What's the goal of comparability. Is there an use case for comparing the 8 definitions of asthma or are we supporting the use case of the person who has their definition, and we want to enable them to search databases to find people who match their definition.
SV: Comparability rooted in the findability of the subtype of asthma you're looking for. Also in reusability. How similar are the 8, and what are the code differences between them.

IF: Brings it back to us expressing refinements.
VV: Are you saying that rather than developing a new standard, use the existing ones in ways to code data more effectively?
SV: Better characterizing the existing data for a particular phenotype.
IF: There's a rabbit hole we can get sucked into, but we're not saying how to change the backends, we're just saying, I want to express asthma as x, and if you're below the curve, that's what we mean by asthma. Just how to express.
VV: We don't have to change the way people are coding it. THen we don't become redundant to emerge.
KVB: Started a comparison. Now that we have more machine readable and FAIR data, is it possible to evolve to something more machine actionable [Sharing screen].

KVB: There are some recurring elements in these models. OHDSI and OMOP have arguably the most powerful way of defining cohorts. Commonality in all cases is that there is some list of codes, at that level, something we could abstract. Could reuse the lists.
IF: That's saying if I'm putting that cohort together, if you have these variables for people who met the criteria, those are things we'd like to see in the dataset.

KVB: If anyone knows this PheMa project, they've been doing some interesting work. They have CQL on OMOP. Could start with just a list of diagnosis codes. PheMA takes the CQL and transforms in OHDSI and JSON. They also use QUICK. The CQL is almost like a grammar but the data model QUICK is synchronized with FHIR. Can write data in CQL. Convert to OMOP cohort definition. Example of how to implement CQL, and how to use the same definition in CQL to execute in FHIR.

CQL and OMOP need to parse QUICK, and make the jump from QUICK (which is basically FHIR) to OMOP.

Tracking down the PheMa folks would be interesting to track down (projectphema.org)

SV: Need a standard or mechanism to quickly iterate while defining a cohort. Having a mechanism to iteratively iterate that definition is powerful. Don't want to provide access to the data, but counts

KVB: Want to produce a roadmap document. Could take asthma example and target three definitions. Assess if this could be transcribed into a common CQL definition that could be executed in both OMOP CDM databases and FHIR endpoints.
SV: Add in another example like ethnicity, it's easy to define but hard to pin down.
I think the deliverable would be a guide.

SV: Do want to say the proposed route forward? Learning would be to define or adopt a standard already out there.

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

2021-11-12 - MIACC Team

Chair: Jordi Rambla, Lauren Fromont, Judit Kumuthini
Attendees "Name (Project Affiliation)": Lauren Fromont, Soichi Ogishima, Ian Fore, Katie Vance, Nicky Mulder, Subhashini Jagu, Michael Baudis

	Actions Arising	Assigned To	Deadline
1.	Presentation planning for the next workshop - reach out to contacts: - FHIR Research Study, CKAN - OMOP - Cohort Atlas - Maelstrom	- Ian - Jordi and Lauren - Nicky Mulder - Jordi, Lauren	Next meeting is Dec. 2
2.	Contacted A Beltran and P Rocca-Serra about review of standards (referenced in DATS paper).	Ian	Contact - done Follow up - Dec 2
3.			
4.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping		5 min
2.	Summary: Revision of the list of standards, and selection of the ones we wish to consider	Jordi Rambla	15min
3.	Identification of specialists and experienced users of these standards	All (Jordi Rambla lead)	15min
4.	Workshop Dec 2nd preparation - Roadmap template	All	15min
5.	AOB		

Minutes:

JR: Subgroup for describing what a cohort is, knowing everything about the cohort, how the cohort is described itself. Idea is that for this meeting, we already have a spreadsheet with some potential standards, and for the next meeting, we would like to discuss the standards and have some experts describe the standards and decide whether something already exists or not.
IF: What we're talking about here is minimum info, and we are only querying things at a cohort level, really querying for cohorts that are the cohort we use in existing studies. The other use case is one that rather than finding these studies cohorts that already exist, but looking at building their own cohorts, so looking into each cohort and pulling out age ranges, diagnoses, etc.Here, we're just talking about describing a cohort in a queryable way with existing cohorts. Distinction between established cohorts and creating them.
JK: Agree with Ian, would be beneficial to have a definition of the cohort. Is it general population or domain agnostic for example, retrospective or prospective. I did indicate that if we add clarity that searching within a cohort vs searching across and how to best harmonize these elements and variable across existing cohorts. Would help if we define the cohort, maybe all of them.
JR: If we are able to describe all existing cohorts, then we can one step father and look at the individuals and characteristics.
Let's say you describe a clinical trial cohort, have all the detail about what a cohort is. WIth that description, maybe you're interested in the age and that they have diabetes. If we describe the larger cohort, we are at the same time described the specified thing.
IF: I would hope we're not going to query at this level, looking at the higher-level data types.
JR: We see an overlap but not 100%
IF: I can see that as part of a cohort description there is data on hypertension, but going deeper how is hypertension defined. Could be a definition of the cohort to say the subjects have hypertension.
JK: Once we do a review of existing standards and how they have mapped these already, there will be overlaps. Will lead for us to understand what and how we should press.
NM: The biggest win is to have detailed information at individual level, if you had that you could populate and then go the next level and describe the particular values. Even is people won't share the data, even have a yes no of whether that data is included is helpful.
IF: Excellent reason for having these two groups. The other group is looking at it, order of magnitude different problem. Need harmonization. This group has a problem that is much easier to produce a result on or tools that scientists can use.
JR: My approach was lets understand what is existing and what level of granularity this shows.
IF: With some of the things the list, what we would look at is a bunch of people wanting to catalogue cohorts. Can observe what happens in the wild - eg. you thought this attribute was important, why? Good way to look into requirements.

JR: Looking into behaviour in the field. Suggest to go over the list/spreadsheet of mixed things we have, and see which could be potential standards. Try to get someone to present at the next meeting.
Clinical statistical profiles

IF: This list was put together for a broader purpose, and can eliminate rows rather quickly because they're not specifically doing what we've been talking about. Mentioned FHIR, that suggests it's more about individual values. FHIR has this concept of research study. Don't know if this particular system is using it, but the research study object is close to what we're talking about.
FHIR Research Study *Resource* is something to include in the list.
OMOP as well.
IHCC - have the cohort atlas, based their model on GECKO. Created an ingestion mechanism and browser. Quite high level data, metadata.
NM: GECKO is linked to multiple ontologies.
IF: Add the ones I sent via email, heavily influenced by biomedical use cases.
[Lindsay to add Ian's list to the spreadsheet]

IF: If you click on the one that's NCPI data catalogue, this shows a list of studies, what data types are they focusing on, study design, etc. If we look at different cohort catalogs and see this kind of thing, it's helpful in seeing what different groups found useful.

Anything we're missing that is human cohort centric?
NM: Maelstrom. It's a catalog to look across multiple cohorts.
JR: Not sure if they use a standard. Can contact them and ask if they have been looking into any standards.
LF: We can reach out to them

IF: There's a platform and standard underneath that many catalogs are built on. - CKAN
S: We can present the childhood cancer data initiative model.
IF: Going to call it an in the wild initiative
JT: will look through and add more detail

JR: Have all presentations within 1 hour
Plan for short presentation and provide more details when we know how many speakers.

2021-11-02 - PHENO Team

Chair: Kees van Bochove, Susheel Varma
Attendees "Name (Project Affiliation)": Ian Fore, Jim Vlasblom, Karen Plevock Haase, Nuria Queralt Rosinach (ELIXIR-NL/EJP RD, LUMC, NL), Francis Jeanson, Jordi Rambla

	Actions Arising	Assigned To	Deadline
1.			

2.			
3.			
4.			

	Agenda Item	Speaker	Time
1.	Welcome, Housekeeping	Susheel Varma	5 min
	Analysis of vocabularies involved in phenotype definitions on today's agenda OMOP Vocab, FHIR Profiles, NHS-D TermServer	Susheel Varma	10 min
2.	Walk-through of OMOP Cohort Definitions using example	Kees van Bochove	30 min
3.	Walk-through of HDR UK Phenotype Definitions using example	Susheel Varma	15 min
4.	Walk-through of N3C Cohort Definitions using example	Melissa Haendel?	15 min
5.	End result and deliverable of this PHENO group: a roadmap proposal using this Roadmap Template	Susheel Varma?	5 min
6.	Agenda for next call: Potential for using FHIR to specify cohort definitions, relation to CQL	Kees van Bochove	5 min
7.	A.O.B.	All	5 min

Minutes:

SV: We split the cohort representation groups into two. One group is focusing on phenotyping algorithms and the other one is much more focused on cohort registry. Of course, you're welcome to join either. But today we're going to be talking much more about phenotyping. And how we represent phenotypic information for cohort boards and how we can make that discoverable. And also how we can operationalize a standard or even an infrastructure to make that more discoverable. If you are familiar with clinical terminology and clinical informatics, you'll be aware of OMOP. *demo of how OMOP works* There are a huge number of vocabularies out there. You've got OMOP concept codes where terminology may not exist. There are also compatibility issues between OMOP codes. But even even if you step one step behind asthma, what are the set of phenotypic conditions that encode the fact that you've been diagnosed with

asthma? You know, what is your lab test results? How is that encoded? Can you probabilistically calculate if the person has asthma or not. So there is a lot of discrepancy and uncertainty in that entire chain. I'm just trying to set out the scene of the complexity of healthcare from I'm sure you're all aware of this in your relevant healthcare system. And this is the same issue that we have. So there are, there are work there is already work around this. So this is just my own example with from the UK.

FJ: Do you think this group is intended to create a sort of discovery accord discovery service based on EHR data alone?

SV: No, no, absolutely not, just highlighting my own background.

KVB: So we have three meetings planned and the scope and the purpose of these meetings is to, in the end, produce a roadmap. We'll look at that later in the agenda. But the purpose of that roadmap would be to specify how a follow up effort could potentially create a cohort definition that's sort of stretches across some of these different communities like HDR UK, OMOP, N3C, etc.

OMOP and Ohdsi

- Ohdsi Care Diagnostics App
- Takes an ohdsi cohort definition (domain specific language) → a cohort is not just a list of codes, there's also very clearly defined temporal constraints
- this is really at a level where this phenotype or code definition is executable, meaning I can run this and what I get back is a table
- you can already see from the human readable definition, that there is one reference here to COVID-19. And that's actually a reference to an underlying construct, which is called a concept set. So in Ohdsi, a cohort consists of a number of rules. But there is a basic book called concept set. And a concept set is always defined as a list of concepts.
- if we go to Athena, and we check this concept ID as you can see that there is within the realm of standardized vocabularies, we have a number of other concepts that are considered the same. And there are a number of other concepts that are non standard concept that are mapped to the standard concepts

IF: How are these definitions being applied to those databases? Who knew brought those data out of those databases into this system? Or does this get translated out to some API which connects to those databases and queries those databases? And if so, who does the translation from the cohort representation to how that query would be applied to those databases?

KVB: That's going towards the mechanics of how studies work in Ohdsi.Typically you create a study package which is loosely connected in an open source fashion. They need to have a database and an R environment. Cohort diagnostics is an R package.

IF: with that list of databases there How would you add another database to it, but I think in some sense to you, that you what you've described sort of illustrates what would be involved.

SV: we need to work with the implementers as well. we already have the cohort beacons, for example, phenopacket standard. OMOP is definitely a provider, and also FHIR. Definitely something to consider in roadmap.

JR: There seems to be overlap here with the MIACC group. In this group are we trying to define the query, or are we trying to define the cohort?

SV: Other group is trying to define the cohort registries (i.e., what are the set of concepts that a cohort represents?)

IF: The essential difference to me between the two groups, is this is about querying within cohorts. I think what you're referring to the query is what I think is the implementation, which is well, which the definition here is expressing the thing that you're calling the question. The devil is in the other part, and I can see why you're going there. Right, which is the query, you're gonna take that definition that if you take a particular database, how am I going to implement that as a query against some system? That's enormously complicated? And, you know, I think we're not, we're not setting out to solve that here.

SV: we have many implementations of this, but we just don't have a interoperable way of doing this across different systems. And that's why I think do projects will be able to help standardize that

Scope:

- Creating cohorts from existing EHR data
- Creating cohorts from data in research studies (?)

Example: how to define Asthma?

- Condition code: SNOMED-CT, ICD-9/10 etc. codes

- FEV value: LOINC code + measurement value threshold

“Cohort Definition”

- Concept Sets
    - Domain (Condition, Measurement, Procedure)
    - List of codes / Resolvability
        - OMOP Concept Set Expression:
            - Descendants / Mapped codes
            - Exclusion of codes
- Cohort Definition
    - Entry criteria
    - Exit criteria

Key to reference Implementers in our Roadmap.

2021-10-19

Chair: Melanie Courtot, Susheel Varma
Attendees “Name (Project Affiliation)”: Peter Robinson, Soichi Ogishima, Jordi Rambla, Katie Vance, Shimon Rura, Ian Fore, Kees van Bochove, Judit Kumithini, Nuria Queralt Rosinach, Jim Vlasblom, Sylvia Gayle, Michael Baudis

	Actions Arising	Assigned To	Deadline
1.	Set up mailing lists for the MIACC and Pheno grups	Lindsay	
2.	Circulate Doodle Polls to each group to determine cadence of meetings before December	Lindsay / Kees	
3.	Move the next full Cohort Rep meeting to December - at this meeting, each group will present their report/plan for a standard. Invite Clin/Pheno, Discovery, and other interested work streams.	Lindsay	
4.			

	Agenda Item	Speaker	Time

1.	Welcome, Announcements - Minutes from GA4GH Connect - Connect Summary - Recordings are available on the Attendee Hub	Melanie, Susheel , Lindsay	5 min
2.	Connect Workshop Debrief - Feedback? Thoughts?	Melanie, Susheel	15 min
3.	Planning Next Steps Discussion - Volunteer listing (did we miss anyone?) - Setting meeting schedules for each group - how to manage comms? New mailing lists? - First activities - Adjusting scheduling (moving this meeting to bi-monthly…)	All	20 min
4.	A.O.B.	All	5 min

Volunteers from Connect:
MIACC (describing sets of attributes required to describe a cohort consistently (eg. age, biological sex):
MC, Jordi, Nicky, MB, Judit Kumuthini, Soichi Ogishima, Ian Fore, Katie Vance

PHENO (looking at existing phenotypic standards, combine pheno info into cohort rep), e.g. OMOP FIHR, i2b2 etc.:
Peter Robinson, KvB, Jordi, SV, MMT, Judit Kumuthini, Francis Jeanson, Ian Fore, Shimon Rura, Núria Q Rosinach, Jim Vlasblom, Karen Plevock Haase

Minutes:

MC: Liked that we ended up with practical next steps, thought it went well. Need to have more for review of standards and whats already out there - eg. the CINECA report.
JR: Momentum is good, and we need to push forward..

How to organize the groups?
- Slack channels for both subgroups? Slack can be hard to get access.
- Create MIACC and Pheno mailing list
- Leads have been highlighted
- Aiming for monthly meetings?

- For Pheno, ideally 3 sessions in this year and conclude with something, either a report it can be done, or it needs specific follow up. Commit to three longer sessions
- Define the frequency and the endpoint/milestone we're trying to get to with some kind of result - meet with the rest of Clin/Pheno and other work streams to present what we're up. Aim for a roadmap, proposal for a standard to the steering committee by DECEMBER

MIACC is for the high level representation, Pheno is for the lower level

Rather than trying to be prescriptive, define the end goals and how we are going to get to that end goal.
JR: Would prefer 1 3 hour meeting a month.
IF: Suggestion for the mid or end point. For both groups, would be important to get a sense of of other work streams. Should set as a target how we will work together, what the overlap is. Have any other work streams partially got this covered.
SV: As part of that document, the first meeting can be establishing what that cross walk or mini-review will look like. Step 1 - we kind have already done a survey. Let's write that up. Can circulate that for feedback.
MC: Should we have a full group meeting?
Aim for Full group mid-December. 30 minutes for each group.
IF: Closure should invite all of the other WSs (Clin/Pheno and Disc).
KvB: I can send out a Doodle Poll for Pheno.
Lindsay to send out doodle for MIACC and check to browse Google group archives.
Pheno - First endpoint should be finalizing the landscape review for GA4GH and non-GA4GH.
At the end of the meeting, should have the literature review complete.
KvB: Two or three sessions with ground work already been done. Working as an example would be helpful.
MIACC - There are ways to represent cohorts. Could describe again in more details, or go with use case and say how we will do that in gecko, OMOP, etc. Need people that understand the representations.

- Landscape analysis: https://docs.google.com/spreadsheets/d/1KXUr2DeRNCbiaHim9l_mnogkqhKAQAid0LbphTHN1Xg/edit?usp=sharing
- PHENO Library: https://docs.google.com/spreadsheets/d/1KXUr2DeRNCbiaHim9l_mnogkqhKAQAid0LbphTHN1Xg/edit#gid=1404998437
-

JK: Should we think about changing it to the value rather than information. Need to get the buy in to explain why these variables are important.
SV: For data sets in general, what we lacked is a standard to collect metadata. What is that minimum and how to ascribe to be able to go back to cohort registries and ask based on this utility score. As we collected metadata can build up the utility

JK: Especially adding new variables and elements.
IF: If we are going to come up with metadata tags or something we want scientists to use. THe question is what our leverage is to be able to get them to do that. Routes through which that happens - partly through funders, partly through the repositories where they add that data. Aspect that we have to think about.
SV: Add the incentives.
MC: Obvious targets like HDR UK, but need to show the value
KvB: Topic should be a part of the report to the group. What do we think is possible, but also the why and explain how that would be adopted.

SV: Examples and use cases in both of the groups.

2021-10-05

Chair: Melanie Courtot, Susheel Varma
Attendees "Name (Project Affiliation)": Melissa Haendel (Monarch), Moni Munoz-Torres (Monarch), Sylvia Gayle, Francis Jeanson, David Hansen (AGHA/CSIRO), Ian Fore

	Actions Arising	Assigned To	Deadline
1.	Last call for Connect slide deck + agenda comments	ALL	Oct. 11
2.			
3.			
4.			

	Agenda Item	Speaker	Time
1.	Welcome, Announcements - GA4GH Connect Registration	Melanie, Susheel , Lindsay	2 min
2.	Checking-In on Connect Planning - Connect Agenda - Presentation Slides	Melanie, Susheel	15 min
3.	A.O.B.	All	3 min

Minutes:

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Brief 20 min check-in for Connect prep.
We reviewed the agenda and slide deck and made comments directly in the documents. There are a few outstanding actions, including adding some additional overview slides to the slide deck and potentially inviting contributors to the landscape analysis speakers to briefly speak.

2021-09-21

Chair: Melanie Courtot, Susheel Varma
Attendees "Name (Project Affiliation)": Francis Jeanson, Neerjah Skantharajah (GA4GH), Ian Fore (NCI), Katie Vance (NCI), Lauren Fromont (CRG), Karen Plevock (NCI), Sean Garin (NHGRI), Kees van Bochove (Hyve), Soichi Ogishima (GEM-J), Emily Boja (NCI)

	Actions Arising	Assigned To	Deadline
1.	Send proposal slides to the mailist list + new Slack channel info	Lindsay	DONE
2.	Refine the slide deck for GA4GH Connect	Susheel, Melanie + All	Oct. 12
3.	Adjust next meeting (Oct. 5) to be a short check-in	Lindsay	DONE
4.	Discuss hackathon frequency	ALL	

	Agenda Item	Speaker	Time
1.	Welcome, Announcements - GA4GH 9th Plenary Registration - GA4GH Connect Registration	Melanie, Susheel , Lindsay	5 min
2.	Presentation: A proposal for a path forward	Melanie, Susheel	10 min
3.	Discussion - Planning a roadmap	ALL	25 min
4.	Planning the Connect Agenda	ALL	15 min
5.	A.O.B.	All	5 min

Minutes:

SV: Did a landscape review where we invited existing projects to fill in information about cohort work. Collected in an excel spreadsheet and we'll share that later. A lot of cohort representation go into detail in data models, knowledgebases, less so on registries and summaries. Trying to address two aspects of interoperability - what cohort are available and how to do we find them, and how do we build infrastructure to bridge across these different APIs.
Pain Points - lack of interoperability, lack of standardization, managing differing baseline measure. Having a reusable computable cohort definition would be helpful and having that interop at the semantic level.
Trying to go in and define those work packages. One layer on the left side is around cohort registry discovery. How can we build a standard that allows people to build a search engine over those cohort registries. What you're really returning is the cohort link where you can get access to the data. This is column level - the right hand side is more of the value level. Computable cohort is asking for the number of patients in a cohort or a particular set of cohorts. There's some work that needs to be done in the middle - left and right is the implementation of cohort discovery work, but the middle is where that metadata standard comes in.
Cohort registry discovery - looking at a column level view. Example is the IHCC global interface that allows you to query. Can query in a federated way where many registries can expose this information.
Computable cohort - looking at row level, collecting up rows that match a query. How to build reproducible computable representation, no matter how it's coded. In HDR UK, the focus is around OMOP, but underneath have ICD and SNOMED codes. Fire off a query into a federated network, counting up the subjects.
The difficult part is the middle bit, how to layer up these things together - inclusion/exclusion criteria at a value level and a phenotype level.

Proposed way of working:

- Two teams, one on cohort registry discovery and one on computable cohort, and have a joint group to harmonize that middle piece
- Suggesting having 5-7 member team with at least one data controller to give us the structure of the data, builders,.
- Bi-weekly 90min issues, and monthly 3 hour hackathon style
- Quarterly show and tell for the whole group
- Running in sprints, 3 month sprints with teams that run in parallel

FJ: Great plan. I really like the ambition and it's showing unique value. Might need that API or backend able to do the mapping with the ontologies.
MC: If we can plug a UI on top that's great. For IHCC example, the left hand side is that cohort registry level, about standardizing the attributes, where the right hand side is about standardizing the values.
SV: Design decision we need to make once those teams are set up. There's a backend and frontend component to both. UI is important, but left to the projects where those interfaces will be implemented.

IF: I think this is great. One thing we should draw out more is that a standalone implementation from one platform doesn't prove what we want to do. Only works when we can do a federated query across sources. Any demo has to show at least two, maybe three platforms involved.
MC: Leveraging the work Beacon has done. Could be a technical solution we could rely on to demonstrate something that works.
LF: In the new Beacon model we have implemented cohorts, but it's not a cohort Beacon. There could be some overlap and I think in terms of mapping there are things we could learn from each other.
MC: Beacon network. If each cohort exposes a Beacon which has information about the attributes and values of that cohort in a standardized way, harmonized or mapped to Beacon, can use it already.
LF: really dependent on the mapping, but would be able to do federated queries
IF: Beacon deals with the right-hand side. Cohort beacon means discovering the cohorts, which is the left?
SV: Yes, don't get subject level values, but allows you quickly filter down which cohorts are of interest.
MC: With CINECA we're using Beacon to do that
LF: There are different ways to consider cohorts. Can make up a cohort based on the ad hoc search or query that you're making.
IF: With cohort registry, you're discovery cohorts other people created. Someone did a study and have a cohort where there are blood measurements.
KB: Adoption is important for those kind of standards. The standards design part of this is more crucial to make that happen. A network study of all of these moving parts. First there's the definition of the data systems, metadata attached to those. Then when running the study, you make these phenotypes, and that leads to counts in all of those databases, which is a virtual cohort we've created. If we would have HDR UK gateway and IHCC atlas all use the same standard for defining a cohort, that would be great. We should be searching for fairly universal standards.
SV: If there is a standard we don't need to invent, all for it.
IF: Could use the word dataset in every place you used the word cohort. We've been aggregating data from within datasets for some time. Haven't created something new here. Cohort is a particular kind of dataset, but almost everything still applies.
SV: The problem we have is that if you use dataset it becomes too general. If you focus in on the problem we're trying to solve, can we then generalize it.

MC: Is the group in agreement that these are the right problems to think about? Are the two aspects Susheel presented the right thing?
IF: A lot of overlap between them. Those two teams are just focused on different parts of the same data model. The registry is dealing with finding things. The two things are just different SQL queries on different tables.
MC: Right problems, but just one small team focused on the two ends
FJ: First is the stage 1 high-level properties, rows will be those attributes or values that can be distinguishing from different rows. That second stage is like looking at refinements and diving a bit deeper into additional mapping.

IF: Overlap is the column names should be the same, shouldn't have two teams working on defining column names. I think delivering on one is much easier than the other. We have people working on tools for doing that (OmicsDI). Some benefit in separating the two
SV: Operationalizing. To get to the deep value level, need to convince people with level 1.
IF: The thing that will make the difference is getting people involved who have the data.
KB: When we were looking for a generic standard to describe the study, cohort definition, outcomes, it's not really there. Made an extension of bioschemas, but that's the point where I don't want to be the only ones.

Agreement that those are the right problems, but better to not have 2 teams as they are just targeting different facts of the same problem. Stage 1 vs stage 2, where stage 2 is a refinement.

Lindsay - send slides to the mailing list + slack channel

MC: We are talking about one team looking at cohort discovery with the sprint schedule and will start with review of existing efforts.
SV: Yes.
IF: Creating a workshop in January on Search, one aspect is Cohort discovery, some overlap in efforts.

MC: Had discussed having an extended version of the slides at Connect, make a call to arms for warm bodies to help. Refine slides to describe one team. Have a bit of description of the standards we've already considered or are familiar with.
KB: Start with an inventory.
IF: Use connect as a kick off for a hackathon.
SV: Let's refine these slides for GA4GH Connect. After that, arrange a hackathon.
Shorten next meeting to "office hours" to check the slides

Actions

- Refine slides for GA4GH connect
- Discuss hackathon frequency
- All to bring ideas/standards etc to Connect

2021-07-27

Chair: Melanie Courtot, Susheel Varma

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Attendees "Name (Project Affiliation)": Jordi Rambla (CRG), Katie Vance (NCI), Francis Jeanson, Soichi Ogishima (GEM-J), Elizabeth Janes (GA4GH), Neerjah Skantharajah (GA4GH), Ian Fore, Melissa Haendel (Monarch Initiative), Monica Munoz-Torres (Monarch Initiative), Sylvia Gayle (NCI), Karen Haase (NCI), Jonathan Fuerth (DNAStack)

	Actions Arising	Assigned To	Deadline
1.	Identify a simple query, and apply it to one source using Beacon 2.0 or Data Connect API. ● Jonathan will share links to Jupyter notebook used for Connect API: Data Connect queries over the "384 Phenopackets" collection https://colab.research.google.com/drive/1jhq5ONPj59jrQ4bmZGJtjakdur2tCFGd#scrollTo=u2WwrfudIbLp and Data Connect over KidsFirst FHIR Observations https://colab.research.google.com/drive/1Fr6l02-1clmouxVuhUgtsRVe7cVphjQA#scrollTo=qO_rYcj7glbS ● Melissa mentions we can query N3C - the largest clinical cohort in the US. https://ncats.nih.gov/n3c		
2.	Run same query across different sources and integrate results (summary statistics)		
3.	Create a document to track queries	Susheel	
4.			

	Agenda Item	Speaker	Time
1.	Welcome, Announcements - GA4GH 9th Plenary Registration - Please note that a GA4GH Schema Library: Requirements Gathering discussion will take place at the Clin/Pheno full WS meeting on Wednesday July 28 (29th Asia/Pacific)	Melanie, Susheel , Lindsay	5 min
2.	Summarize landscape analysis	Susheel	10 min
3.	Review cohort representation scope document	ALL	40 min

4.	A.O.B.	All	5 min

Minutes:

Landscape Analysis
SV: Thank you to all who have submitted responses. We have a number of APIs and resources about cohort discovery and similar platforms. Propose we start categorizing and find commonalities between data models, APIs and the knowledge bases. Find overlap with current standards like phenopackets and Beacons. Try and define the extraction protocol to define a particular cohort or phenotype. We seem to be mixing two different concepts. One is around a phenotype, or a set of criteria, and a cohort. The additional bits missing from phenotype is the ability to define phenotypes together with a defined inclusion/exclusion criteria. There is a type of definition for phenotypes or cohorts to represent an individual or type of individual, and there is also the definition of a query, which is a bit different. Encourage everyone to take a look at the landscape spreadsheet, identify anything you may already be doing as a part of your work and split into these three main components.
MC: Remind that we need this group to fill in the gaps, so we don't want to re-do the efforts of any of these initiatives.
MH: Have a generalizable strategy for defining a computable phenotype. Make a computable phenotype I can execute against any resource.
MC: Add this to the scope document.
IF: A computable phenotype that we can execute against different sources - is that essentially what susheel is calling a query?
MH: Yes, it's a set of parameters. Have a query that defines a cohort or phenotype with any combination of parameters. Right now, people encode that individual in their own way each time they want to run it. How we can turn it from being a pdf recipe that can cook itself.

- From chat: Phenotype is the most widely different used word. But yes, a computable phenotype is a specific type of query parameters that are run against a population to define a cohort that meet that specified set of parameters:
    - - In basic science it's an observable entity.
    - - In molecular medical genetics, its a syndrome or disease
    - - In EHR clinical informatics, its a set of query parameters that define a cohort

IF: Bringing a whole set of activities within GA4GH and trying to harmonize our language. I wonder whether it helps to encourage some simplicity and focus on this being a query. In the interest of communication, talking about this computable phenotype as a query against phenotypic data communicates a lot more easily to people.
SV: We use phenotypes and cohorts a lot, which means two different things. A cohort query definition is a list of phenotypes and this combo is what I need to extract. Phenotype is an overloaded word.
IF: There is something more to cohort than a query. A cohort is the set of subjects or patients that I've determined are of interest to me. I might assemble that from one or more queries. We want to be able to represent that set.

[Chat]:
SO: https://github.com/BBMRI-ERIC/miabis/blob/master/Data-describing-Sample-Collection.md

Scope Document

SV: We've tried to capture some of those definitions in the scope document. A cohort would just reuse the ohdsi definition. Computable cohort is a schema to represent that cohort that can be run/executed against a resource. There is functional representation where you have a cohort study that represent patients with particular themes, but could have multiple cohorts within that. This is all up for discussion. Is this the totality of what we mean with computable cohorts?
MH: Great diagram and matches with my understanding.
FJ: Phenopackets comes to mind as a data representation for phenotypic information that is compatible with FHIR. Phenopackets is a data model and could build a query around it. Doesn't mean it will be compatible with everything else. If you have other data models that becomes a challenge. The Data Connect is trying to solve that will a tool to pull the data out. Need to juggle. If phenopackets is successful, might not need to go beyond phenopackets. There is value in creating a query tool that can create an output.
SV: How the response comes back, that could be a phenopacket.
MC: In the spirit of reusing what's already been done, we could use Beacon or Data Connect API. Beacon already includes the Phenopacket representation.
JR: In my understanding, both things are compatible. Need to have a cohort representation and an individual. Added cohort into the Beacon because users want to find it as a unit. We have longstanding cohorts, and temporary cohorts someone is generating on the fly. The only one that really matters is the one created and proposed, and the others are derivations. The concept of cohort with a proposal is the one missing.
MC: If I were to take a example. In this document we define a study as a predefined collection. There is a relationship where cohort is a subtype of cohort study.
SV: Umbrella cohort study, within that we might have one cohort definition, or you could have multiple. Once you've defined that cohort, you want others to come in and ask if my cohort definition is applicable in this study.
JR: Many times we see the cohort is actually a group of cohorts. If we look at cohort study as the entry point, we are aggregating a collection of collections. Could make things a level of magnitude harder. Stepping stone for me is the individual cohort.
MH: We've talked about the relationship between a computable phenotype and phenopackets. Want to be able to know if I define a particular cohort, would my definition be that selects all of those patients. Has to allow for variability we would expect in a cohort, like partially penetrant phenotypes.
MC: Had been thinking about different guidelines depending on the country for the disease you're looking at.
SV: Depends on the type of study, if competitive yes want to run one with and one without. If trying to extract a cohort itself, almost want one definition for everything.
MC: Case definition would have different levels of certainty.
SV: We're almost reinventing SQL, but using the semantics of the various terminologies to try and define extraction. We want to be able to say "what do we mean by x disease", and if you don't have that, there could be a list of other terms available. We could do it at a lower level, or a higher level as a single massive query.

In terms of cohort definition, we've added in those operators - logical and time interval. Being able to specify as part of the definition.

MH: One thing that's missing in a lot of provenance and attribution. Challenge with mapping across terminologies. For me to run in one place to another, might have to map the terminological content which is always lossy. The rules for mapping in those contexts also need to be included.
MC: If we want to look in terms of milestones or practical steps, first is to look at representation of a simple phenotype and how it could map across different cohorts. Second step would be looking into logical operators on top of that, and then having a way to share the computational phenotypes for re-use.
SV: Cohort summary is taking that query and providing summary statistics available from that cohort definition. After that query is done.
IF: WIth this discussion, are we talking about defining a query language? Are there not query languages that do a fine job and have tooling around them. Seems like we're dancing around that question.
SV: We're trying to get the representation right, and they may become a query for someone else to run.
IF: There is the model part, and the query language way of expressing it.
MH: I feel like when we think about what's happening, ther query language is SQL or SPARQL, but there is no standard for how to define the structure of the query.
JR: Not trying to create a query language I hope. Query by example, throw an example of what you expect with some exclusions.
SV: If you look at OMOP or i2b2, they have a SQL mechanism. What we're trying to do is look at these different places and agree on a cohort definition. Need something agnostic to the implementation.

What we're trying to do is work through an example use case. I personally feel that having an example data set would be better rather than trying to build the definition first. Invite the community to provide some examples.
MC: A simple example may be something to find in a lot of places.
IF: We can think about use case in terms of what is the query, but I think part of our use case should be applying this to different sources. Take this representation and ask how it is applied to different sources. Our use case should include a set of selected sources that we would want to ask if they have something that matches the cohort representation.
MC: Is it worth considering the Beacon and Data Connect implementations?
IF: It wouldn't be data connect itself, it would be the sources it reaches through to. If we're going to take a query, that only goes so far as an use case. The real use case is applying that query to different sources, like a Beacon.
JR: When we do the poor man definition of cohort in Beacon, we would love to have a model to provide a better query. We need to have the sources, which is another challenge.
SV: Could simplify our task by starting with a synthetic data set. Put a Beacon and Data Connect API in front of it.
MH: I don't think synthetic data will take us where we need to go. Would be too conflated with the variables used to generate it.

SV: Actions: identify a set of query, perhaps I'll create a separate document. We don't need to be prescriptive here. Another action would be to set up some sources that Beacon and Data Connect could set up in front of.

JR: Might be a bit premature because Beacon v2 is not yet approved. Suggest we do something manually first. With 3 or 4 criteria it could be enough.
MC: In interest of testing, maybe Jordi and Jonathan if you know of sources. Try to query against a specific source but I don't know what is available. I assume you've looked into some test queries.
JF: We've used the 384 phenopackets collection to test some things. We've also tried similar queries over the KidsFirst collection which is FHIR. I don't know of a third one. We have some jupyter notebooks where you can see this (will get link).
MH: We can query N3C as well. I think if we can find some generalizable phenotypes that are common across communities, like diabetes. Do a small inventory of a few phenotypes.
SV: Cross-walk of a single phenotype.
MH: How are the different groups defining the schema of how to query.
SV: Structural differences as opposed to the values.
MH: Yes, how do they define inclusion/exclusion and penetrance. I don't know if anyone has done that across more than a couple.
FJ: In machine learning there is a growing interest in converting natural language to database queries. Type query in a free form text and it converts to database query format. Might speak to how we create those queries in the first place. Might be something to look at later on.
MC: Are you happy adding that to the use cases?
FJ: Yes, I'll add it.

2021-07-13

Chair: Melanie Courtot, Susheel Varma
Attendees "Name (Project Affiliation)": Moni Munoz-Torres, John Grimes, Elizabeth Janes, David Hansen, Francis Jeanson, Jordi Rambla, Aina Jene, Sylvia Gayle, Alejandro Metke, Katie Vance, Soichi Ogishima

	Actions Arising	Assigned To	Deadline
1.	Add to landscape analysis	Everyone	
2.	Work on scope document	Susheel, Melanie	Next Meeting
3.			
4.			

	Agenda Item	Speaker	Time
1.	Welcome, Announcements - Reminder to fill out the use case collection/landscape analysis form - GA4GH 9th Plenary Registration	Melanie, Susheel , Lindsay	5 min
2.	CSIRO Cohort Activities	John Grimes	10 min
3.	~~Reviewing the draft scope document~~ *Moved to next meeting*		25 min
4.	Next Steps	All	15 min
	A.O.B.	All	5 min

Minutes:

Pathling
JG: Specialized type of FHIR server, input FHIR data directly into and translate into a form for analytic queries. Provides operations that are not possible using the regular API such as aggregation. Better for processing large datasets. Can work alongside a FHIR terminology server.
Allowing for exploratory data analysis, advance patient cohort selection (exclusion/inclusion criteria), and data preparation (taking FHIR data and transforming for data science tasks). THe aggregate operation is a tool for “pivot table as an API”. Aggregations can be used to power dashboards and visualizations. Example with a cohort of autoimmune rheumatoid arthritis, queried how many had been prescribed a TNF inhibitor and contracted a lung infection. We can have any number of aggregate and grouping expressions. Here our aggregation is simply a count function, then we have our first grouping expression. Get a true or false answer for each group in the data for whether the patient was prescribed a TNF drug. Taking a look at the code in the resource and testing whether they are a member of a particular value set. We use SNOMED in this example but could use any terminology. Can do translations between code systems, so if we had rx norm codes, we could use pathling to translate at query times.
Similar expression for the two filters. The result of this is a table that is combination of the different boolean groupings.
Search is different from aggregate in that it returns the full FHIR resources themselves rather than aggregated numbers. Can be used for patient cohort selection and can be used to drill down aggregate results. Can grab one of the rows, and that row comes back with a drill down expression we can plug into a search operation.
In this project we had a study done within renal genetic clinics and captured data at three stages within the process: referral, pre-investigation and post-investigation. As we apply more resources and use more specialist practitioners, is the diagnosis becoming more specific, is it more refined. The diagnoses were recorded in SNOMED CT. Came up with counts for each group, and use the response to populate this diagram.

Roadmap - working on extract operation, creating arbitrary extracts from FHIR. Real-time analytics, synch between a real-time FHIR data source and pathling. Subscriptions from Pathling itself. Cell suppression. APIs for Python and R, tableau and powerBI integration so we can do sophisticated querying and reshaping within those environments.
This is all free and open source, permissive licensing. Open to collaboration.

JR: Curious about the technology behind?
JG: One of the problems with complex clinical terminologies like SNOMED is that it goes beyond regular reference data. We're using apache spark (?), and write to a column data format which is optimized for numerous people working on it, slicing and dicing. Because of the runtime integration with Ontoserver, we can do arbitrary computation within queries.
JR: Have you stress tested the load?
JG: I've done up to 1.5 mill synthetic patient data set, which could have up to hundreds of millions of rows in some of the tables. Supports datasets of arbitrary size.
JR: THe FHIR server has a standard enough API that could pluck and place it into any other FHIR server?
JG: Standard API, uses Happy at the front end. Because the FHIR spec is extensive we've taken advantage of that, defined extended operations to be a part of our API. Intended to augment a regular FHIR server, not meant to replace.

SV: How wedded are you to FHIR API and model. We've started doing some work to land GP record data through OMOP into spark. Is there flexibility in terms of using another common data model?
JG: Potentially. FHIR path is a model independent language, could actually apply the same design to interrogating something like OMOP. The other way is to continue working on the FHIR to OMOP story, which is probably closer. We've focused on FHIR because there wasn't a good story for FHIR at all. The tooling for OMOP is far superior to FHIR. Applying this model to another CDM is on our radar, it is conceptually possible.
SV: We were converting FHIR to OMOP to take advantage of advanced query function.
JG: More primitive functions in spark sql. Can have something you add to class path which gives you some of the terminology functions without being bound to FHIR or the pathling implementation at all.
SV: I didn't see negations in there, can you do negation?
JG: You can do it in both FHIR path and expression constraint language in snomed.

SV: Conversation we might want to have around the representation of that cohort. Want to be able to say "give me patients with a particular set of comorbidities" that could be complied into Pathling or omop. Integration point. Another with GA4GH around the Data Connect API, they have a generalized query table API, if there was a link you could make to support FHIR path, that would be another added benefit. Once you have a cohort defined, you can fire it off in the data connect API.

SV: On my to-do list is to get a draft of the scope document together for the next meeting. Key is highlighting what we are not going to do. In the interest of time, want to spend a couple of minutes presenting the projects already out there.
Please add projects to the landscape analysis form.

Pick a few examples to show the range of activities:
ISB-CGC is a cohort browser with pre-packaged queries. Similar approach with the ALS browser. Each of these browsers have pre-baked queries, but want to enable researchers defining their own queries.
Another is just listing out all of the code (eg. Covid-19). This code list is used to construct a FHIR or OMOP query.
PheKb is another example, but all the codes are embedded, which you need to be able to extract to convert into a query.
In HDR UK, we just have a list of codes, but want to get to a point where we can define these rules.
Taking the finngen browser with pre-baked queries. This is how pretty much every phenotyping group has approached this problem. Fingenn does run summary statistics (pre-computed). What we are trying to get to is a definition of this.

Gut feeling is to build something around phenopackets, and then work with groups like Beacon, Data Connect API.
Problem is the cambrian explosion of all of these groups doing certain things in certain ways so we can't collaborate effectively.
JR: There is no standard so far? Do they have descriptions of cohorts that are similar to what you have shown us? Phenopackets would just be a list of phenopackets to make a cohort.
SV: OMOP has a de-facto standard, interface for cohort query. Atlas cohort discovery and extraction tool. Can do sql queries with FHIR. These are all defacto implementations, would have to transcode your definition of a cohort into each of these. Is that the problem we want to solve? If that's not the problem, there is an underlying one, how to represent a cohort defined by FHIR vs OMOP and get summary statistics out of different systems.
JR: Any others solving this problem?
SV: Not that I know of.
JR: With phenopackets, getting more into health. With cohort, stepping farther into health and maybe abandoning the genomics part. Everyone assumes that genomics have to be included.
SV: Merging the genomics and health worlds together.
FJ: Seconding the comment of health. Genomics to phenotypes to behavioural is increasingly popular. I think this is a great project. Build an overarching query builder that is interoperable between the different models. One would be to understand the major models we've brought up already. Have a schema for query building that could map down to those various models. Other approach is to keep it high level and work with the community building those model to adopt the overarching model this group would eventually output.
SV: Dealing with messiness of data, and the queryability of that data because of the large number of datasets out there. Starting simple can be hard to do. Put a simple layer on top of systems to do a federated query across them. Build integrations from the beginning.
FJ: Frameworks and models that are relational based, also a growing use of graph based representations. I think it would be interesting to explore that and see how it could be incorporated.
SV: Definitely. A few initiatives are going down that path, but I'm not aware of anything in production.
FJ: Future milestone.

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

2021-06-29

Chair: Melanie Courtot, Susheel Varma
Attendees "Name (Project Affiliation)": Katie Vance (NIH), Karen Plevock Haase (NIH), Peter Robinson (Jax/Monarch), Michael Baudis (SPHN/UZH), Neerjah Skantharajah (GA4GH), Lindsy Smith (GA4GH), Jules Jacobsen (Monarch/QMUL), Jordi Rambla (CRG), Ian Fore (NCI), Sean Garin (NHGRI), Francis Jeanson, Aina Jene, Lauren Fromont, Nuria Queralt Rosinach, Jonathan Fuerth

	Actions Arising	Assigned To	Deadline
1.	Share landscape analysis form widely, please complete if you are using cohort data	ALL	
2.	First draft of the scope document	Susheel and Melanie	Next Meeting (July 13)
3.	Landscape spreadsheet - please add relevant APIs/initiatives/platforms, etc. that we should be aware of	ALL	
4.			

	Agenda Item	Speaker	Time
1.	Welcome, Announcements - Reminder to fill out the use case collection/landscape analysis form - Phenopackets v2.0 has been approved by the GA4GH Steering Committee - EDI Onboarding Project - Video	Melanie, Susheel , Lindsay	5 min
2.	Landscape Analysis/Tracking Relevant Activities - Spreadsheet (work in progress!)	Lindsay	5 min

3.	Presentations Debrief, Planning Next Steps - Roadmap Form - GA4GH Oct. Connect	All	45 min
	A.O.B.	All	5 min

Minutes:

SV: Please feel free to share the use case collection document among your colleagues. We are trying to capture the work already underway across the globe. We don't want to invent a new standard, want to reuse where we can.
LS: Have started to fill in the spreadsheet based on the information presented at the last meeting. Please feel free to fill this in further and add any other resources/API/initiatives that you are aware of. Please feel free to add new enums for resource type as well.
SV: All of the information collected in the form will be added to this spreadsheet. Next stage will be to review this list of resources, and working out where all of these projects are at. Build a strawman of what cohort representation would look like, what is the phenotypic information that needs to be represented. Once that strawman is built, look at which bits of Phenopackets and Beacon we can reuse. If there gaps, we can work on that here. Could view this subgroup as an interoperablity layer between cohort and beacons.

FJ: How to define a cohort? For me, a definition of cohort is essentially a group of participants that would be likely to participate in a study. There might be some platforms where they've recruited a number of participants from studies, might have data portals that allow you to search for studies. Are we looking more for portals or platforms that have curated collections of participants or allow you to create a collection of participants?
SV: Great question and comes to the crux of the problem in terms of cohort definition. In my mind, the static representation of a cohort, so if you look at a clinical trial, think of the cohorts in that clinical trial as defined per inclusion/exclusion criteria. That is a static representation and could reuse that static representation to say how this clinical trial is compared to another. Could have dynamic cohort discovery and extraction where you say I have a bunch of patients sitting in FHIR or OMOP and I'd like to extract a certain number based on these inclusion/exclusion criteria I define. Try to have a standard that straddles both of those scenarios.
IF: In both of the examples you are talking about, it's a grouping of patients based on some set of criteria. Even the person that assembles that cohort from multiple sources, that's their set to do research on. Such a grouping could be put together in multiple ways. The details of what a cohort are are down to the particular investigator. Some collection of patients together for some interest.
MB: Important to have an umbrella term that can express one meaning and then have subtypes that are stricter defined to the wishes of stakeholders. Don't say 'cohorts are this' but the cohort model allows for this version to express what a cohort is, for x use case.
IF: I'm less in favour of defined distinctions. Have an inclusive definition but stay out of the weeds of the details.
MB: Ontology of cohort types

SV: As part of the first phase, cohort is a group level of subjects, things matched up with a set of criteria. We can iterate on that standard to get into the weeds and see how different groups implement. Not trying to reinvent a new standard, want to define a cohort definition and see how that fits within Beacon or Phenopackets, see if we can extend those models.

First step is trying to get this form out into your networks as wide as possible. Being able to share statistical profiles of cohorts. Can use those statistics to discover cohorts and collaborate without sharing the data from the cohort itself. By the next meeting, we'll put together a small working group to put together a strawperson of the representation.

JJ: The cohort representation could use a phenopackets as a profile of what the cohort could contain?
SV: My mental model is that there is a cohort representation that uses the phenopacket standard for inclusion/exclusion criteria.
JJ: Phenopackets has the cohort message which is just a container. There is no way of defining what the inclusion criteria within the schema.
MB: These entities and attributes could be defined by the subject of the cohort. Not defining the cohort in phenopackets, but using the phenopacket to say for these parameters…
SV: Imagine you define asthma using snomedCT - give me all the patients with asthma but not with allergic asthma. Trying to define and constrain based on a set of criteria. That criteria could be defined by phenopackets. Almost boolean logic.
JJ: Likely needs to set outside of a phenopacket, but wrap around it. Data could be presented as a cohort of phenopackets. One problem is how to define what we're looking for.
SV: It's a layer above phenopacket, but the transmission of data could be through phenopackets.
IF: Some focus of interest, the selection of entities. The phenotypic data about those entities varies with the study. Cohort in a different disease might be looking at completely different attributes. With a cohort, you want to define what the entities but also what the data you have about them is. A cohort will be a table of data essentially, a table of columns. What we need is the ability to define that. Data Connect API has been looking at that. Phenopackets is about defining a set of attributes for particular use cases, but for different studies we probably have to leave it to a new structure or phenopacket to describe.
SV: How can I have a phenopacket as an exclusion criteria for my search. Someone has already done the hard work in defining a phenopacket from the data, can I use that definition as a criteria. If I can use that to hit a cohort Beacon, and that can run a federated search on all of the different cohort beacons, do a data connect query to get a table of entities that I'm interested in.
IF: We could be going down a road where we're defining a cohort as a query. It's the set of criteria in machine readable form. In highly desirable worlds, could apply that set to a data source.
SV: We're not trying to boil the ocean. Starting with the first premise of defining what a cohort is, then can it be queryable. Find out which bits can be queryable.
IF: Once you have a cohort, querying over that cohort, that's a much easier proposition.
JR: When we decided to add cohorts in Beacon, our stakeholders needed a definition to define the data types. I don't see that as something too open or abstract. Suggest to focus on specific cohorts.

JJ: Could envisage using phenopackets as a schema for the query. Whatever results that get returned is the cohort.
What are criteria used to include those individuals, and then what are the counts.
JR: Looking into the counts in the inclusion criteria.
SV: In HDR UK, have to provide a recruitment flow diagram. What I'm trying to get through is a standard to represent that, define the set of inclusion criteria. We need to first be able to define that certain definition of what inclusion/exclusion criteria to compare. I want to be able to compare my cohort definition to others and see why there is a signal difference. Action for myself to put this together into a Github or doc. This is what we are doing, and this is what we are not doing. If we agree on that, can move forward from there.
IF: Data Connect API has some relevance to this group here - one is the ability to define the schema. We looked at queries over phenopackets, addresses how the schema is represented, and the question of the query language.

2021-06-15

Chair: Melanie Courtot, Susheel Varma
Attendees "Name (Project Affiliation)": Miro Cupak, Jordi Rambla, Moni Munoz-Torres, Gideon Giacomelli, Ian Fore, Sean Garin, Christopher Chute, Francis Jeanson, Susan Fairley, Susheel Varma, Melissa Haendel, Nuria Queralt Rosinach, David Hansen, Lindsay Smith, Thomas Keane

Meeting Recording

	Actions Arising	Assigned To	Deadline
1.	Share an introductory slide about Cohort Representation goals to the GA4GH SC	Melanie, Susheel (Lindsay has tagged you both in a comment on the slide)	June 24
2.	Fill out the use case collection/landscape analysis form	ALL	
3.			
4.			

	Agenda Item	Speaker	Time

1.	Welcome, Housekeeping - Reminder to fill out the use case collection/landscape analysis form - Please email lindsay.smith@ga4gh.org to be added to the cohort representation mailing list	Melanie, Susheel , Lindsay	5 min
2.	Current Cohort Activities (revisiting past presentations)		
a.	Cohort Statistical Profiles	Chris Chute	5 min
b.	N3C	Melissa Haendel	5 min
c.	CINECA & IHCC - Slides	Melanie Courtot	5 min
d.	HDR UK - Slides	Susheel Varma	5 min
e.	Beacon v2 - slides	Jordi Rambla	5 min
f.	Phenopackets v2.0 - Cohort Element	Peter Robinson, Jules Jacobsen	5 min
g.	Data Connect API - Slides	Miro Cupak, Jonathan Fuerth	5 min
3.	Next Steps	All	20 min
	A.O.B.	All	

Minutes:

MH: Come up with one or more strategies for how GA4GH can standardize cohort representation, and how do these align with the overarching goals of GA4GH. Interested in clinical profiles, which is a statistical representation of a cohort that is computable and sharable in a completely deidentified way outside of the context of EHR data. Way of realizing characterization of cohorts.

Statistical Cohort Profiles

CC: Provide a public data source about clinical data that is completely open, transparent and has nothing identifiable. It is de facto a “statistical summary” of real data for defined phenotype. Could have a clinical profile about a subtype of diabetes or rare disease. Includes counts, std devs, distributions, percentages from the usual quantifiable suspects: labs, meds, diagnoses, demographics. Tries to preserve first order correlation among variables. Rendered originally as a FHIR resource. Looking at the algorithms for how we would do correlation since they’re not all numerical values. This is the matrix across procedures, vitals, labs and devices, look at all the combinations. Different ways of doing association metrics. Example of JSON Asthma - we divided by demographic groups.

Inputs we could use include FHIR and the CDMs, where you generate the clinical profile, harmonize, annotate and validate them, and publish them as first rank artifacts with any level of resolution or detail for any given cohort. Can create metaprofiles from multiple organizations, open public sharing resource without any concern about privacy because they are just statistical derivatives of the data. It is not yet ready for prime time, but needs more maturity.

MC: Statistical clinical profile example?
CC: “distribution of average sodium value in diabetic 15 yo white male with juvenile diabetes”
IF: One that’s required would vary according to study and the purpose.
CC: You don’t know which is required and that’s the whole point. Get a summary of the whole thing (mean, median, variation, etc.) with statistical profiles. Distributions.
SV: This really depends on your sensitivity and specificity of the phenotype itself, eg. allergic asthma vs asthma in general
CC: Correct, you could generate an arbitrary number of clinical profiles on increasingly granular definitions of phenotype.

N3C
MH: Chris and I lead the national cohort covid collaborative (N3C). 55 clinical institutions have clinical data harmonized in the system. The way in which those patient records are selected are from a computable phenotype. Hope to do this for all rare disease and not just in US common data models. We have 4 model coming from US institutions: ACT, odyssey, picornet and trinetx. Looks for patients with a positive covid test, or a ICD code, or two or more covid-like diagnoses. Match one positive patient with 2 “equivalent” ones negative (equivalent eg for age, ethnicity…) Need for computable phenotype that can evolve.
https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype

How to implement and have a more generic and deployable phenotype: create the phenotype for each combination of each data model. Each implementation has their own script to run those phenotypes. Each model does not have the exact same combination of things.

MC: Is it right to look at each implementation having a layer on top to provide a uniform API for querying?
MH: Not an API, it’s a local script run to extract those records to then provision those data. Have combinatorial problem of each phenotype bein defined for each data model for each database structure.
MC: And are the records loaded in a centralized database?
MH: Yes. Could imagine a standard target CDM from a combination of computable phenotype alongside the database structure you have, but then that populated local data warehouse is where you put the API on top, or export phenopackets, or whatever sharing mechanism. Some sort of common transform process will required a CDM specific version of a phenotype and an extracting script
MC: We have similar challenges, do you put each site in charge of transforming, central loading, who is in charge of mapping, etc.
MH: Ability to generate harmonized data at the source is not terrific, but can't assume each will have the ability to harmonize centrally either.

CC: Example is units of measure. The transforms are fairly trivial, but in case of not knowing what the unit of measure is we can harmonize. We can also 'salvage' values - eg we know the premature baby weight in grams distribution so we can fix data eg if missing unit.
IF: Statistical summaries, are they being calculated dynamically in how the user selects the cohort?
CC: When we do harmonization, we create a new column that is the harmonized value. That becomes the analytic target.
MH: Can you select based upon a subset of criteria eg. show me the comorbidities of patients in x range of a given lab.
CC: Make an additional caveat that epidemiologically, we have an arbitrary case and selection control based on COVID.

IHCC & CINECA
MC: CINECA and IHCC are both consortia of cohorts. 1.5 million total participants. IHCC focuses specifically on cohort 100k+. For harmonization, we have built a simple model built on the consensus overlap across CINECA cohorts. Looked at overlap between dictionaries. Formalized as an ontology.
The usage of that model right now has been for cohort discovery. Built a UI on top of the data to find cohorts of interest. For example, finding a cohort that has specific imaging data across all of the cohorts in IHCC. At this stage, we've been trying to address the syntactic harmonization and the semantics when looking a blood pressure for example - some cohorts might use medications, some cohort say yes or no, some have a value of the blood pressure.
Each cohort dictionary captures mappings and we have a system of automatic ingestion.
Would be interested to know what others are doing for value harmonization.

We've taken human readable (we have pdfs!) and converted them into OWL so it could be machine readable.

HDR UK
SV: We've been trying to build together a phenotype library. Collect phenotype definitions for disease. Taking Asthma as an example - see metadata, primary care codes, etc. Can we get to standardized way or representing a diagnosis so it can be shared and we all mean the same thing. Exclusion and inclusion criteria. We then make it computable by generating Phenotyping algorithm workflows in Phenoflow. We have not tried to harmonize what asthma means, because you can have multiple definitions. All of this is in Github. We have gone in and said once you have that definition, run the code that can extract that from a database. In terms of phenotyping, we're not dealing with probabilistic. Once we have defined a phenotype with all of these codes, we have built a cohort discovery tool. Not at the level of IHCC but more at the granular query level.
MC: Interested in the three definitions of Asthma. Which country you're considering, which guidelines. There is probably some common thread around breathing problems.
SV: Phenotype definition, what do we mean by phenotype. Two is the statistical profiling. Cohort definition, combinations of phenotypes together that you want to be able to define as a query. Final, how to run a cohort query service, like a Beacon, with additional capabilities.
MC: that gives us a nice workflow from cohort discovery to requesting access, use GA4GH Passport/DUO and get the data eg to load into trusted research environment

FJ: Are you thinking about Phenopackets as a guidance here?
SV: I suspect Phenopackets will cover as much as we want from that definition. Would suggest we add into Phenopackets rather than create something new. That kind of metadata layer we can add on top.

Beacon v2.0
JR: You have a question and Beacon helps to provide a response. In Beacon, we have two concepts of cohorts - study defined and user defined. The beacon model has the concept of collections, which could be a bunch of records joined together or a cohort that include individuals with inclusion criteria and so on. Beacon allows the use of different schemas - could return an OMOP model, FHIR patient model, etc. Beacon is the point where you can trigger an operation, the query, the workflow, etc. and returning the data in the format preferred by both parties. Differently from what has been presented so far, Beacon aims to allow simple queries but allows for a diversity of responses. There is a boolean response, or the detailed schema. Cohort fields in metadata, inclusion criteria, data usage, statistics, data types and collection events.
We have the odyssey model as a starting point with some gecko elements on it.

SV: The model you have presented is great. The question then is the implementers of the cohort Beacons would have to have some mapping. One level of complexity or abstraction above is when they have phenotypes defined already, how can we represent that within the phenopackets and return.
JR: I removed phenpackets from the list of examples here, but it is an option.

Phenopackets
MMT: Phenopackets is a model for the exchange of phenotypic data - when observed, how observed, how severe. The latest version has a bit more information about the representation of cancer and COVID phenotypes. In the context of phenopackets, a cohort is a group of individuals related in some phenotypic or genotypic aspect. You could have individuals, all of whom have been diagnosed with a certain disease or phenotypic feature. Cohort element includes an id, members, hts_files and meta_data. Expected that every valid phenopacket will have a metadata element.

Data Connect API
Miro C: RestAPI that provides a mechanism for describing data and its data model - don't prescribe our own model, all we've standardized is how you tell us what your data model looks like. The Table API component provides a way to organize data into "Tables". Relying on JSON schema. Also enables searching the data with the given data model. Using SQL. We can do queries across different data types. Realize these are general and can solve many use cases, mostly restricted by the underlying data. Any data that can be collected as JSON objects, so tabular or document data, can be made discoverable and searchable.
Example: Find me all of the adult males diagnosed with autism having a harmful in SHANK1.
Need to develop pair-wise adaptations to the tools you want to use. Data connect can act as the single adaptor for all of these tools and data sources.
Where we can fit in the context of this group is provide a flexible query language for cohort selection, could also be an integration layer with other standards and systems. If we go more on

the data model side, we could help with ensuring the data model is written in a way that could integrate with data connect.

SV: Many people attacking the same problem from different lenses, which signifies a need. For Steering Committee meeting, we need to set up the layers of abstraction, and figure out how to build up that vertical gap. If there isn't a gap there, we need to build the glue.
MC: Interested in figuring out shared use cases and overlap.

2021-05-13: Computable Cohort Representation Kick-Off

Chair: TBD!
Attendees "Name (Project Affiliation)": Melanie Courtot, Ian Fore, Marc Fiume, Susheel Varma, Moni Munoz-Torres (Monarch Initiative), Melissa Haendel (Monarch Initiative), Miro Cupak, Aina Jene, Jonathan Fuerth, Michael Baudis, Lauren Fromont, Stephanie Li, Jules Jacobsen (Monarch Initiative), Jordi Rambla, Nuria Queralt Rosinach (LUMC, ELIXIR-NL, EJP RD)

	Actions Arising	Assigned To	Deadline
1.	Re-visit previous presentations in a cross WS context + Beacon and Search Overviews		Next meeting
2.	Create a spot to record use case and requirements	Lindsay & Moni	
3.	Circulate some suggested times for the next meeting	Lindsay	
4.	Follow up with Susheel re: Subgroup Lead	Melissa, Moni, Lindsay	

	Agenda Item	Speaker	Time
1.	Welcome		5 min
2.	Recap of current GA4GH efforts - Cohort Representation Breakout at 8th Plenary (incl. IHCC, CINECA, N3C)	Lindsay, Melissa	5 min

	- Beacon v2.0 cohort schema - Document capturing the email thread - Computable Cohort Shared Google Folder		
3.	Group Discussion: - What are the goals/purpose of this group? - Why is standardization needed? - What could be potential deliverables? (Cohort definition schema? Method? API?) - Start defining requirements	All	30 min
4.	Formal Use Case/Requirements Collection - Driver project Google Form (may need tweaking depending on what the deliverable is) - Form is just meant for collection, responses can be collated, categorized and refined collaboratively - Need volunteers to perform technical requirements and landscape analysis	All	10 min
5.	Logistics - Regular scheduled meeting times? - Group Chairs? - Would a landscape analysis discussion be helpful? - Product Proposal considerations	Lindsay, All	10 min
6.	A.O.B.	All	

Minutes:
MH: Have been noodling with having a cohort schema standard in Clin/Pheno for about a year. Have had meetings and a series of presentations to talk about it. Presentations focus on three types of initiatives. When first developing phenopackets, the idea was that there would be a way to describe an individual case, but at the end of the day if you had a collection of phenopackets you would have a cohort packet. Ultimately, cohort was out of scope for phenopacket schema. Three main presentations given on 1) Ontological classification 2) Computable phenotypes (exclusion/inclusion criteria) 3) Statistical correlations and models of disease that represent the cohort
MH: Question that comes up is “How do we query a set of phenopackets?”
MH: Have not started on deeper requirement analysis. Clin/Pheno group had also stated a need for, once phenopackets v2 is released, an API for phenotype profile matching and phenotype querying.
MH: What are the requirements and how can we make something that integrates standards across GA4GH?

JJ: Have already had a cohort representation in phenopackets, but it is bare bones. Just a big list.
IF: There may be a semantic issue here in terms of how we use the terminology for the use cases discussed here. There are things in common amongst the number of groups that are looking at this, but it'st being referred to in different ways. Need to get on the same page beyond a term like computable cohort. What do we mean? What particularly is being looked for out of the sense that it's computable?
MH: Those three flavours (above) are pretty well defined in terms of use cases and scope. One is the ability o model and query a cohort based on a schema. Second id computable phenotype - how do you lift a set of patients based on inclusion/exclusion criteria? Third case is retrieving statistical correlations. Those three cases focus on phenotypic characterization, but not sure how/if it relates to other contexts.
IF: It may be well defined, but may not be well understood.
MH: We can do the presentations again across the Work Streams. It is well understood in Clin/Pheno group.
TK: Can see the first use case. Modeling a cohort (first use case) is definitely an area that GA4GH hasn't looked at yet. Trying to figure out how the second and third use cases could overlap with existing standards (e.g. Beacon v2). Writing out use cases will help to understand the goals.
MH: Didn't know what was happening in Beacon. Would be good to see the Beacon presentation to see how it relates to phenotypic characterization. What's most useful to focus on next and what should GA4GH be doing about it? Huge need is computable phenotypes to exchange around the world. And statistical correlations are a strong interest.
SV: Context from use case pov - different data models across all 4 nations of the UK. Everyone has their own cohort definition standard. Hoping that GA4GH can explain how to harmonize and interoperate between different standards of how you represent a cohort. First use case is straightforward from my pov. Hope is to build out standards so we can have an interoperable framework. Computable phenotypes are ad hoc right now. Need a mechanism to collect and curate. That in itself can be translatable and interoperable. Agree with statistical profiling initiative as well. Once we define cohort definition we can move on to how to represent statistical profiles. Happy to work with GA4GH to build out those standards.
MH: Help me document your ideas. Some of them are in scope for GA4GH and some are not. Harmonizing clinical data models out of scope for GA4GH, not a standard more of a method. Can have an offline convo about how to transform your clinical data. One of the problems that we have with PKB for example is they're very platform specific and not very computable. Susheel would be a good lead to figure out how do we want to port it from the UK to the US. GA4GH can help standardize the framework.
MB: For Beacon v2 there was a clear need to have some kind of cohort definition in place. Have started to shift from data set centric to grouping centric. Your use case descriptions were spot on and helpful in detailing what cohort schema should fill in. I don't think that beacon wants to push its own cohort standard. Beacon doesn't define which ontology should be used but we find that a certain schema should be used. Some people have issues with use of the term cohort. Has a specific meaning in certain epistemic areas, but in some cases, like beacon use cases, "cohort" has a much wider definition than what some clinical study design people might think. More of a warning than a disagreement.

MH: Would be good to understand what that meaning is that is wider than the general clinical cohort.
MB: Not so much from my pov. Metadata task team...
JR: Beacon only tries to open up the path. Two metaphors: 1) in beacon we try to be the plastic fork - something useful that solves a quick problem while keeping your silver forks in the right place. Beacon tries to do this with the schema. 2) Beacon trying to be a pioneer. Someone defined pioneer as someone with arrows in the back because they've crossed enemy lines. Beacon is providing something incomplete, incoherent, not necessarily well defined
JR: How many people are using it? How many people can populate it? Would like GA4GH to provide a complete and coherent cohort representation schema. Can be used as an alternative schema. Lessons from Beacon could inform GA4GH Cohort working group.
MC: Was going to suggest something similar. There are a couple of schemas/models. Looking into aligning Beacon, sinica? and OMAP? Maybe looking into landscape and how those things align. Suspect they're quite overlapping. There's probably a set of cohort attributes that probably don't pertain so much to phenopackets Maybe better defining which aspects we aim to cover, what models/schemas are already available and where we want to dig deeper.
MF: Almost every driver that we've talked to in Discovery has had this need "How do we perform selection across a cohort?". Discovery has focused on punting on some of the specifics - definition of cohort, define functions that are shareable. FASP takes a similar posture in that we want to be a vehicle for discovery through access and analytics but staying general. How does this live on that continuum. What are the functions, like mutational burden? Is that what this intends to do? Create the library of 1) a definition of the cohort and 2) a library useful to statistical functions that can sit on top of this ecosystem.
MH: In terms of knowledge exchange and representation about an individual cohort within an institution that statistical representation is very useful and surprisingly different across institutions. Would want to understand how different groups would want to leverage that as a resource in their analytical workflows. Maybe in a diagnostic setting or in a care assessment setting you're asking yourself how does my patient compare to a cohort statistical representation within my own institution vs other institutions? Identifying what cohort within a collection of patients meets a certain set of criteria in a reproducible way across systems is an obvious key goal. How that relates to search or discovery is one of the questions we're trying to get at here. What do people do with that computable phenotype once they run it? For us, we say okay in these 80 institutions which patients meet the criteria. Not pulling all the records, only the ones that meet the criteria. But may have different questions like of those that meet that characteristic in a distributed network how many meet x,y, and z which might be genomic information. That might be a search use case.
SV: Both those use cases are exactly where we're at. Being able to do a search query to be able to find information about patients.You're looking at pop. for a particular disease, and disease is a computable phenotype. The first use case is that once I've discovered my set of patients within large population of data custodians, run this analysis.Those are the two main use cases we have. Need to answer the first part before we can address analysis.
IF: Marc alluded to generality in data connect, want to talk about where that comes from and how it relates. The convenient thing for data connect to refer to, had this question around focusing on tables, and that caused some problems. The useful shift there is to think about the product as a data frame - that's the unit that people analyze with various tools.The cohort is the data frame of data where each row is probably a person/subject. Depending one what your use

case is you have some set of data that goes into that. Concern is how we produce those people and how we move them around. If when we say computable cohort, we are talking about this chunk of stuff that we want to move around, but the generality is we need to leave the construction of the details to the individual project. There's standardization that can happen (e.g. description and construction of data frame).Not to say we don't want standardization. Generality is a positive characteristic.
MH: I need to understand more what the use cases are for generality. Could we have some notes or a presentation on what that might look like?
IF: To do the analysis on that set of data that specific scientific investigation wants to do. In any case they would have to articulate a use case. General principle is to leave it open.
MH: Need to walk through what a specific example might look like because I'm still not sure how connect would query for phenotypic attributes without a schema.
IF: Could put this on a future agenda.
MH: We need requirements, examples and deeper needs. Lindsay put together form that we can send out to driver projects. GA4GH should do some due diligence about what is the value that GA4GH has that isn't otherwise met by other organizations or standards and furthers our hope of having more interoperable ecosystem of standards. How can we better document requirements?
LS: We have this form, it's relatively general and simple. Can tweak this as people see fit. What we could do is use this document that has notes from the initial email thread. Use the form as a collection mechanism and then put them into a more formal table or area where people can comment on them.
MH: When I think of requirement analysis I think about it as a more formal process. Form is great for driver projects but would like to propose that we take a more formal requirements analysis for this purpose which involves understanding key functional requirements, e.g. queries like this should function that way, different resources that are needed like a schema, api, best practices guide, and then also where's the landscape analysis and gap analysis of what already exists. And actively make decisions about what we're going to build. Collectively we're not doing a good job of coordinating deeper requirement analysis. Don't think we should wait for the forms to come back to start that process but they will be useful to inform the process. After formal analysis go back to drivers with what we're planning and ask them what they think.
LS: How often do we want to meet?
MH: Clin/Pheno only meets monthly so if we can use one of the leftover slots that could be an option
LS: may need to look at a rotating schedule to accommodate people in Australia. Free slot is on Wednesdays at 3pm EST.
IF: Might need more frequent meetings to start off and then settle into a more regular rhythm
MH: biweekly, alternating in time to be more inclusive. Nominate Susheel as chair.
SV: Happy to lead this group.
MH: Next meeting

- revisit the three clin/pheno use cases that I described so everyone is up to speed.
- overview of the search use case and beacon schema
- In the meantime, create a spot for people to place their use case and documentation.
- Lindsay to send out an email to identify a time for the next meeting.
- Melissa and Moni to follow up with Susheel on subgroup lead role



Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.