Computable Cohort Representation

Date: October 13, 2021
Time: 12:00 - 13:30 UTC
Meeting Chairs: Melanie Courtot & Susheel Varma

Objective: Computable cohort representation is a new subgroup in GA4GH exploring standards around cohort discovery (from a registry) and computable cohort discovery (cases). We will present a summary of the discussions to date, and invite the community to provide feedback on next steps for standard development.

Attendees: Lindsay Smith, Susheel Varma, Melanie Courtot, Marion Shadbolt, Ian Fore, Nicky Mulder, Alex Wagner, Tim Cezard, Soichi Ogishima, Moni Munoz-Torres, Francis Jeanson, David Hansen, Francesca Frexia, Kees van Bochove, Peter Goodhand, Judit Kumuthini, Shimon Rura, Brian Walsh, Melissa Haendel, Jordi Rambla Thomas Keane, Davide Piscia, Steve L, Oliver Hofmann, Miro Cupak, Ashley Hobb, Michael Baudis, Aina Jene, Angela Page

	Agenda Item	**Speaker**	**Time**
1.0	**Welcome**	Susheel Varma, Melanie Courtot	5 min
2.0	**GA4GH Cohort Standardization Efforts: A Path Forward** (Presentation) - Slides - Landscape Analysis Spreadsheet	Susheel Varma, Melanie Courtot	15 min

3.0	**Discussion & Feedback on Next Steps**
	- Contributing to the landscape analysis
	- Call to arms for contributors/volunteers: Please let us know if you'd like to be involved!
	- Discussion questions:
	  - Are there multiple standards here?
	  - What is a feasible meeting structure?
		ALL	30 min

4.0	**Cohort Discovery - Standards Inventory** Diving into first steps for cohort discovery: - Creating an inventory of standards for cohort discovery - Presentations from landscape analysis contributors	ALL	30 min

	- Is there anything we are missing?		

Minutes:

Some reasons folks have joined:

- beacon is aiming to making discovery of cohorts around the world possible
- see if we can apply the standard to improve interoperability of cohort data
- combination of research and real world data (including EHR) intelligent creation of cohorts, and making data available for advancing research
- how we select and move around collections of patients/subjects and assemble collections of data on them
- how can we do better at scale for federated analysis in the use of human cohort data
- align activities in uk to global discovery and federation across jurisdictions

New standard started over summer '21, landscape analysis, gap identification
Main question: is there a market for the gap?
Survey circulated via social media to gather information around cohorts, phenotypes, etc, (still open)
Some work already available, some tangential
How do different initiatives represent data around cohorts, represent phenotypic info - how is it discovered, accessed, through websites, portals, APIs.
Some paint points:

- Interoperability between diff data models across different cohorts and across the wider landscape?
- Performing analysis
- Semantic interoperability, not just technical
- Standardized demographic variables

Requirements:

- Reuse standards already available
- Computational representation of what a cohort is / represents within a database, extract that for further analysis
- Represent different models, how do you compare/contrast different definitions, as part of publication for reproducibility

(welcoming additional pain points and requirements)

Two Main Boxes:

1) Cohort Discovery -- Computable Cohort Discovery: find a phenotype within a db, find set of patients, eg. semantic version of seql query
    a) What encoding is used in the db, demographics, phenotypic info
    b) Once I've discovered cohort, how do I go in and find exact set of patients
    c) Utilize other GA4GH standards, eg., passports to gain access to the cohort you want
2) Cohort Identification -- Cohort Discovery: find db's with patients of a certain phenotype
    a) Broad set of cohrots
    b) Pull data from a cohort, feed into a pipeline

Metadata fields in the middle that represents a cohort registry (high level) and vocab /ontologies to represent deeper phenotypic information, unlikely to be a single standard, but collection of standards, pull together and leverage APIs to activate them

Defining cohort at a high level and at a subject level

Tabular way to think of it:

- High level = column level, what columns are in the dataset? Don't need to know the individual subjects, just the kind of information contained in the cohort
    - Eg., CINECA project / IHCC Atlas: Interface collected in centralized or, more interestingly, federated db to discover the information, richer data in federated, dynamically query across the globe
- Patient level = rows, looking for a set of patients with specific inclusion/exclusion criteria; encode those criteria and make that reproducible, encode computationally using a definition, take same criteria into another db / by another researcher, repeat the analysis and reproduce the results
    -

HDR UK - running with 9 cohorts so far

- Build query, fireoff, return results
- OMOP, I2B2 - any standard, the data semantic interop between models, how do we encapsulate/adopt that for representing cohort information within GA4GH

How do we represent cohort, set of phenotypes, algorithm used to represent, summary information display, how do we operationalize that?
Minimum set of information? Align with existing work, eg., BBMRI, CINECA, Gecko ,registry alignment, query alignment, payload alignment, API alignment
Bring different entities together around the payload at both registry and phenotype level

Need experts on how data is structured, how to develop standards, operationalizing on a technical level

Proposal:

- 1 team, 2 projects (cohort registry discovery and computational cohort discovery)
- small teams (5-7 people), mix of people who have discoverable real-world data, informatics experts, technical builders
- Biweekly sprints (90 min) and/or monthly (3hr) hackathons (with bi-monthly meet with larger group)

Need technical builders as well as informatics and data controller expertise, contact Lindsay.smith@ga4gh.org

Question from Thomas: The discovery question seems relevant to everyone here - how much of this is covered already by the Discovery standards? Are we building a new standard or building a demo? Staging - find out what's working within DPs, then how do we interop between two different entities, is there a missing gap

Can we interop between one or more? If not, what's missing? Find a standard that can roll out across multiple DPs

MC: best practices, training,

MH: Interoperability is important, helpful to also do understand what purpose they would serve, terminologies/fields in alignment doesn't necessarily solve a problem - whatever standard(s) we deliver should be able to be used together to solve a problem.

IF: to illustrate a problem, this is what we've looked at in FASP - how to put together a combined set of pancreatic cancer patients across EGA and dbGAP and in Japan, and Australia etc. etc.
Broadening to more continents, basic use case we've been working on for a while (second use case, cohort building)
Scientific questions around asthma, cancer, etc

SV: two problems within this esp, around federation: within a DP, connecting information just for your own project, rolling out existing solutions; and second, between DPs, how do you interop with other projects to discover other parts of the biomedical landscape puzzle

Set of biosample registries want to discover something about, consented for secondary research, how do we drive that use case through?

RCarroll: many of these use cases in NIH, different cloud platforms/programs funded across the institutes, limitations of row and column - not the way it actually looks!
Trying to figure out how to engage old data (csv file), common standard to just talk about the data to interact with it, but also building semantic layer on top of it, data collected in different contexts
Aligned with GA4GH Discovery scope, but they've been thinking about it differently, not clin/pheno specific, Data Connect may not hit on the specific use case

SV: here's the general ga4gh discovery layer, here's how clin/pheno can plug in, we'd translate blood pressure statement into semantically computatable form, in whatever format it's in

IF: Data connect consciously doesn't do what it doesn't do, needs to partner with others, specific use cases, only will find "creatinine measures" in certain datasets, Data Connect leaves it to the use case to define how to do it with more specificity
Significant part of the problem, is when you get to more detailed level, need capability to define things in specific study

RC: to get to next level, need to have some level of agreement, not necessarily same vocab, but enough semantical alignment that it works

We also have set of real world researcher use cases as part of NCPI, may be in public github, help frame actual real world problems

IF: in DC, trying to extend that beyond NCPI, look at EGA, eg., which has many of the same challenges. NCPI FHIR examples, look at those datasets, how do you find the value you're looking for (if through FHIR), how does that work through Beacon, DC, etc., do graphical query languages do this better? This group could begin to explore some of those questions we've been talking about for some time

MC: next steps (semantics being discussed in the chat), do the people on the call agree that the questions are the right question, how to tackle as a group
IF: Yes, in the right ballpark, need to get the right groups working on actual data

MH: responding to Kees' comment in chat (agreeing to MH's comments. Defining a cohort can be done unambiguously if you use a specific patient level data model. Whether you can fo in a model-agnostic way and whether that's meaningful and useful), started this group bc do have computational cohort data models, but constantly need to transform for every model and db instantiation, phenotype def's are computable in one context but not exchangeable, exchange standard so they don't end up in PDF and word docs
Is that possible? Not sure, but have examples in diff db's

KvB: even when you are using the same data model, eg. OMOP v5.3 and OMOP 6.0 already have this issue, specific even to one software, run same cohort in a FHIR server, see if you can get same results there; is it meaningful? Even if we can agree on a semantically sound standard, only useful if communities like odyssey start using that standard, otherwise just out there gathering dust
Some tools, eg., seql, makes possible to formulate clear query model, could look into that

SV: Can we do some work around, let's see if it is possible, or come back with 5 things we need to do next even to be able to make progress, paper - this is what we need to do, these are the steps
KvB signs up :) Looking for additional volunteers, phenotypic semantic alignment around cohorts

Looking for volunteers to move high level metadata around cohorts, often requested feature

PRobinson: One challenge in non-hospital institution, don't have access to typical datasets used for projects like this; what is the strategy for collaborating on basis of some shared dataset?

SV: want teams to have someone with that data accessible, if not possible, create version of datasets used to evaluate the system, proof will be in real world deployments, after governance framework is developed

IF: push forward db's eg., dbGAP and EGA as large repositories with necessary diversity to make sure general applicability, need to talk about studies within the db's

Data in dbgap and ega mostly research studies, not clinical - we can certainly get the first, may need clinical care data as well

PR: thing about dbgap, cohorts have already been selected, so no algorithmic challenge to find a cohort, could use for a test, but isn't challenge to go into mixed cohorts of patients to ID a study and reliable control.

IF: one use case is to find a cohort from multiple sources, pancreatic cancer study in dbgap and ega, use case is to create a cohort that selects patients who match my criteria to create a new cohort with subjects from both, computational exercise, that's *one of* the use cases; Use case to discover pre-assembled use cases

PR: any time you do a retrospective study, confounding is the single most important thing that will screw you up, idea of mixing data from dbgap with data from a hospital. Limited time and resources, maybe such a broad focus is not best approach.

SV: what does that mean for the user, query two datasets together; deeper problem as well

MBaudis: PR made a good point, cohort aggregation by query, can be flexible, but also through work of search options, etc., have to predefine something by certain set of criteria as an agglomeration, isn't as important anymore, bc search options and federation gets better and more applicable to diverse resources
Concept of what cohort formal definition group has shifted away from as I'd seen it before, now what metadata/formats/how are they expressed? Want to do this with Beacon.

IF: This is a use case I hear over and over again, can you compare the data, even if some assay for ostensibly the same molecule in two different studies? Scientifically yes, can it all be done by query? No, you have to aggregate the data and work with it. That's getting to common use case, needs to be taken account of; initial description is naive. Actually complicated, people want to do it, how can that be done?

Volunteers:

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

PHENO (looking at existing phenotypic standards, combine pheno info into cohort rep), e.g. OMOP FIHR, i2b2 etc.:
PR, KvB, Jordi, SV, MMT, Judit Kumuthini, Francis Jeanson, Ian Fore

IF: MIACC being worked on in NLM as DATMM, looked at DATS, need to be careful about inventing standards on the fly;
SV: MIACC group, not the MIACC standard! Group of people coming together to say what exists and what is still needed
MC: avoid creating a new standard if not needed. If you have projects that would be relevant that you've worked with, please share them so we can fill the gaps

Meeting Cadence:
Pheno may be more asynchronous, exploring landscape before deciding
Reporting back to clin/pheno work stream and wider GA4GH community

IF: How would MIACC differ from how you define dataset in bioschemas?

MC: phenopackets, bioschema may not have cohort description in them already, that will be part of the work, to look at those and see what's already in use, GECKO, others

SV: medical study by schema definition; need to do a crosswalk, **[ACTION: bimonthly report backs to GA4GH, let them decide cadence of their own meetings]**

MH: clin/pheno leads will be passing torches soon, thinking about projects that we've been working on, pedigree and phenopackets, getting pretty mature now, would be important to evolve in context of what's done with cohort work, ready in the work stream for this to become primary effort of next couple years, new leads may have additional ideas, but this seems to be most pressing with enough momentum

SV: need voices of support and criticism! Not going to be an easy effort.

LS: Full work stream meeting next week, use it as a touch base to confirm names, availability etc.

MH: different sc hema types of efforts, addressing alignment in fits and starts, how to make interop and deploy in APIs in consistent way, how to reconcile different approaches in terms of interop, suggest looking at LinkML - flexible way of having a common modelling framework to deploy schemas with whatever schemas, etc

Pheno group - look at whether or not to get them to do so is linkML

FJ: being cognizant of various sources of data, not just thinking about traditional data repositories, but also differnet modalities, clinical trials, real world data and anything in between

NM: H3Africa is a "cohort of cohorts" - everything submitted to EGA has minimal level of data, PIs have more data, need to think about a handover, it's available, but how to get extra data via collaboration,

SV: Here's an example of where you can and cannot do a handoff and still do high end discovery

PR: intersection of phenopacket and this work

MB: interested from beacon side, also how do we represent cohorts overlapping with studies, what's a cohort and what's a study? How do we work with original data, who provides link to the original study, derived, remixed, representations, provenance issues...

MMT: here to connect to phenotype side, phenopackets, even landscape analysis will be useful

Possible data sources:
MH: N3C has synthetic data
FJ: ADDI has some synthetic sets, Other open data from Dryad, Open Neuro, and others

Overall Takeaway:

- Two teams focusing on MIACC and phenotype alignment
- Collecting volunteers for both "pizza teams" - will set their own efforts and report back bi-monthly


GA4GH
CONNECT