Day 1: Monday, May 3

11:00am-12:30pm – Welcome and Working Group Updates

Welcome - Sam Volchenboum (UChicago) and Tanja Davidsen (NCI)

Notetaker: Alan Zheng (NCI)

Notes:

[Tanja Davidsen on logistics]

- Please mute yourself if not speaking
- All sessions will be recorded except for the breakout sessions
- Links to the meeting recordings will be available after the meeting
- We will be taking notes during each session
- Speakers: please turn on your camera when speaking if possible
- Please vote for Fall Workshop Dates (Poll link) . We will announce the Fall meeting dates at the end of the meeting tomorrow (Tues May 4th)
- Each speaker will have 20 min (15 min talk and 5 min Q&A); We will give warnings at 5 min and 1 min left in the talk;
- Please enter your questions in the Chat or raise your hand
- CBIIT tech support is available during the meeting

[Sam Volchenboum on Introduction]

- We will have very technical talks today and tomorrow about data harmonization and standardization
- A view from the perspectives of a pediatric oncology clinician on the final common pathway
- As former FDA chair Robert Callif would always talk about, we live in two parallel universes: in the Clinical Universe, we have Electronic Health Records that are fed with information from other resources/systems, which has been instantiated and institutionalized. We also have this Research Universe. The way that data makes its way to the Research Universe is still very old and archaic. Data has to be manually entered, and get collected by third party tools, and somehow get into Research Universe
- Some examples to show these parallel universes and why they are such a problem. Example of a cancer protocol for children: the way how data got entered and the lack of standards make it hard to extract information.
- Lack of standards is the reason for these problems and why we cannot fix them. There are lack of standards when we collect data, and lack of standards when we transfer data. The theme for the next couple of days is how do we come up with standards for both data and transfer [another example of lacking standards: mapping hemoglobin data from the hospital's database]

- We want to build a common data dictionary that will allow us to collect data in the new harmonized way. We cannot do that until we get consensus. It is a lot of work to get groups together and come up with a common way of speaking. But the work is worth it: at the end you will have a common data model, a common dictionary, and then you can share your results.
- It took us over a year to develop a data dictionary. A lot of efforts and a lot of people got involved. [Example] working with three groups: children's oncology group, and two international groups to come up with standards for head & neck cancer and for sarcoma. We don't use the commons until we have the data dictionary established
- We really need to head toward a single universe where we can tackle all these problems. One of the main answers to helping tackle these problems is to build a standardized dictionary so we can all harmonize the data. It will allow us to populate clinical trial forms, to change the way we think about the two universes of data, and try to make them into one

NIH Coordination Group - Speaker: Valentina Di Francesco (NHGRI)

Notetaker: Natalie Kucher (NHGRI)

Notes:

- Valentina thanked members of the working group, all who have been very involved and contributing in many ways.
- The responsibilities of the Coordination Working Group are to serve as the governance body of the NCPI projects, steward working group progress and support investigators, and liaison with various parts of NIH.
- NIH recruited Asiyah Lin as an ODSS DATA Scholar who is solely committed to support NCPI work. The DATA Scholar program was launched by ODSS to bring to NIH people with strong backgrounds in fields such as machine learning, computational and statistical methods, and ontology development.
- The Coordination Working Group brought on the first new member to NCPI: NCBI of the National Library of Medicine. The group has learned a lot through this process and that information is contributing to the development of "rules of engagement."
- The group has successfully obtained support for NCPI activities. The platforms submitted proposals for activities to cover in the next 6-12 months. This work includes:
    - Improving search across platforms and enabling search of the aggregated data,
    - Performing coordinated outreach among the NCPI platforms via the portal, training, and a data dashboard,
    - Developing tools for users to estimate cloud costs for analysis workflows and perform cost optimization,
    - Enabling workflow execution across NCPI platforms,

    - Defining guiding principles for technical interoperability and overcome operational barriers, and
    - Adopting RAS and GA4GH as common authorization and authentication.
- One need for interoperability work is allowing platform developers to access data in each of the platforms for testing of interoperability tools. The Office of Science Policy (OSP) has taken the charge on developing a proposal, which the Coordination Working Group will review once it is circulated.
- In April 2020, the NIH Coordination Working Group approved the Five Principles for Interoperating Data Platforms. The Community and Governance Working Group is working on how to go from defining the principles to their implementation.
- The group is working to define NCPI "Rules of Engagement". NCPI has been successful in defining interoperability and testing technical solutions, which has gotten attention from NIH on who can be involved in NCPI. The Rules of Engagement discussions led to five criteria for potential members to consider, which are still in development. The group has also begun defining a decision tree for engagement, where potential participants could be designated an interested party (participates in working groups) or a committed member (voting member with committed use cases and resources).
- There has been progress on the NCPI goals from the October 2020 Workshop. The group has been successful in continuing to host workshops and pursuing additional funding from ODSS for NCPI, and will continue to make progress towards the other goals.
- The goals of this workshop are the same as the driving goals outlined by Jonathan Kaltman for the October 2019 workshop, which is to identify 2-4 concrete use cases or collaborative projects to work on in the next 6-12 months. This will be accomplished by identifying the people to lead implementation of each use case and determining how NIH can support the work.

Community/Governance - Speakers: Bob Grossman (UChicago) & Stan Ahalt (RENCI)
Notetaker: Natalie Kucher (NHGRI)
Notes:

- Before NCPI was established, each platform had carefully considered and determined their approach to data management, security, and other operating principles. Working with platforms that use different models and are based in different places require a number of agreements to achieve interoperability. The goal of this working group is to simplify the diagram of agreements between NCPI platforms for interoperability.
- Even in the cloud, there must be several agreements to allow data sharing by the data owners and for those who access and use data to do so.
- Cloud platforms must meet security and compliance requirements, for which there are different solutions across NCPI.
- The NCPI platforms are moving to RAS for user authentication and authorization, which is part of what is needed to authorize environments. Platforms which hold data and want to distribute it would need users to provide a RAS passport and trust the client that

sends it. Right now, the data and analysis tools are bundled in a system. In the future, the focus can be on whether a user is authorized to access data and whether a platform is authorized to analyze it, focusing less on the monolithic platforms.

- The mixture of security, governance, operating models, and how data can cross system boundaries to support interoperability are complex. In the whitepaper table, they highlight the similarities and differences among the platforms.
- There are two proposed considerations that, if adopted, would allow platforms to remove the different types of arrows demonstrating cross-platform agreements and replace them with agreement on the active considerations. The considerations are summarized in four concepts:
    - 1 - a user is authorized to access a dataset (i.e., has a user agreed to the terms and conditions to use the data),
    - 2 - cloud platform A has right to distribute a particular dataset (i.e., can the data be moved to a different platform),
    - 3 - cloud platform B is an authorized environment for a particular dataset (i.e., is the cloud platform where analysis will be done an approved platform to receive and analyze data),
    - 4 - each dataset has a data trustee (aka data steward) that makes decisions about 1, 2, 3 (i.e., NIH or awardee delegate).
- Interoperability can be achieved if you authorize the users, environment, and trust the authorizations.
- The remaining work is to agree on the agreements for authorization, standard ISAs, interactions of government and third-party systems, and crossing of security boundaries.
- NCPI members were encouraged to review and comment on the whitepaper. The glossary provides definitions that will facilitate precise and productive discussion.
- Q&A
    - How do you approach the issue of who owns the data? Do users give up their rights on how data are used?
    - Bob commented that a researcher can't share data unless there is an agreement that you can put it on a certain platform and share it a certain way. This is integrated in the dbGaP process, and it's also part of the data trustee process which is done with dbGaP or others. Data ownership needs to be agreed on first.
    - Jaime commented in the chat that NIH is the steward of a copy of the data which we distribute on behalf of investigators/institutions in accordance with wishes of participants. Bob noted this is the fundamental principle, and this working group is trying to remove barriers that hinder interoperability.

Systems Interoperability - Speaker: Jack DiGiovanna (Seven Bridges)

Notetaker: Brian O'Connor (Broad)

Notes:

- **Slides:** https://docs.google.com/presentation/d/1Tp4IUcNVMa-iML3JrTgMl3ekydEso_JmbZycUUDTxeM/edit#slide=id.gd59cd672ce_14_11
- Showed the 2020 flow of portal -> workspace -> DRS access
- Showed the state of connectivity as of Fall 2020
- As of now, able to use data from 4 repos (BDCat, AnVIL, Kids First, and CRDC) to workspaces in SBG and Terra-based systems as well as Gen3 workspaces for some
- Prototyping with CRDC… GDC handoff using manifest → PFB import
- Prototyping with Kids First + AnVIL… manifest from AnVIL to client → Cavatica
- AnVIL+BDCAt… portal push to workspace is working in production
- Demo: Jack showed a workspace in SBG that is using TOPMed, Kids First, AnVIL, and CRDC data all in a single workspace, accessing data via DRS
    - Showed search on each portal
    - Process for "handoff" was automated for BDCat and Kids First
    - AnVIL was through a manifest import
    - CRDC was via data browser in SBG
    - Important outcome, workspace that has data references from each data repo
- Use cases:
    - See the use case document
- PFB/FHIR
    - Bridge prototype, can we use FHIR, get a result, and "hand off" via PFB?
    - Prototype was done
- RAS
    - Authentication… log in via RAS across all systems
- DRS 1.2
    - Passports + DRS
    - Authenticate client system
- Lessons learned
    - We relaxed our ideal solution to getting data from portals → workspaces… wanted to get researchers productive
    - Avro/PFB vs. manifests… no free lunch
    - Single AuthN/Z would greatly, greatly simply things
- Human Lessons Learned
    - Are we funded? Need to understand funding better, focus on funded activities
    - Looking to get other groups to present

- RAS Milestone 3… we need this
- Need more active use cases!!
- Summary
    - All 4 portals have a path to workspaces
    - Resolved many tech concerns
    - Two use cases now complete
    - Next
        - Using equivalent tools on multi platforms
        - Connect with NCBI DRS server

FHIR - Speakers: Allison Heath (CHOP) & Eric Torstenson (Vanderbilt)
Notetaker: Maia Nguyen (CHOP)
Notes:

- Slide deck: Link
- Background
    - This WG is just over a year old
    - First 6 months of WG were focused on Project FORGE
    - Prioritizes end-to-end FHIR framework for data systems
    - Focuses on addressing gaps, moving data between platforms, and facilitating representation content conversations
- Where We Left Off Last Time: Framework for Clinical Data Interoperability
    - Verified that FHIR is a good framework for interop
    - Identified opinionated and flexible sections within FHIR structure
    - Interacted with HL7 proper
    - Focused on prototype tools and implementation
- Focus of Last Six Months
    - NCPI Implementation Guide Development
        - Background
            - Set of rules about how FHIR resources should be used to solve a particular problem
            - Requires selection of terminologies of choice
            - Making it accessible for various use cases
            - Anticipating rapid evolution
        - Use Case Gathering
        - Profiling
        - Utilizing FHIR Shorthand (FSH)
    - Path Towards Production

        - Background
            - Driven by platform teams
        - Server evaluation
        - RAS and Controlled Access
    - Tooling and Initial Utilization
        - Background
            - Look to eventually use real data for analysis
        - PIC-SURE bulk FHIR import
        - PFB to FHIR
        - NCPI Dashboard
- What are FHIR Implementation Guides?
    - Set of rules about how FHIR resources should be used to solve a particular problem
    - Allison reviewed newly published version of IG
        - Links directly to FHIR resources being used
        - Technical aspects are described
        - Examples of representation are available in different languages
        - Informs Raw > Harmonized discussions
- Existing Study Data - CARING Example (POPS)
    - Used to assess how quickly spreadsheets can be translated into a FHIR server
        - Then layered with APIs
    - Performing light harmonization
        - Reasonable code additions
        - Terminology usage (ex. HPO)
- FSH used for Implementation Guide Development
    - FHIR Shorthand (FSH) was selected for use
    - Requires less scaffolding
    - Shorter code overall - less scrolling
    - Easier to read, write, validate and curate FHIR resources
    - Rapid collaboration and accessible tracking changes
- Implementation Guide Development on Github
    - Restricting profiling to only places where it is required
        - Ex. Disease and phenotype, family pedigrees, DRS
        - Allows URIs to work appropriately
- FHIR Server/Platform Evaluation
    - Working to establish common fashion using a test suite that only uses FHIR
    - Weighted score and high-level/detailed reports
    - Obtaining use cases AND example data (bulk export)

    - Framework is available on GitHub and high-level Google doc
    - Continue work with “testbed” servers
- Summary and Next Steps
    - Refining NCPI IG
        - Use case and background documentation
        - Guidelines on using existing FHIR resources
        - Terminology selection
        - GA4GH pedigree cross-informing
    - Platform Specific FHIR Servers
        - Kids First DRC (end of May, similar timeline for CARING)
        - dbGaP
        - AnVIL
        - Continue to support NCPI FHIR “testbed” servers with KFDRC and synthetic data
    - Tooling and API Usage
        - Interchange, Search, Mapping, and Provenance
        - Prioritize based on emerging needs
        - Integrations using Jupyter Notebooks and Shiny Apps in cloud workspaces
- Q+A
    - Will the FHIR test suite address data interop and/or system interop?
        - I think those would largely be expected to be done outside of the test suite. However, if you have ideas for how to write tests for such a task, we are definitely open to incorporate whatever folks are interested in
    - Do you expect that NCPI systems will all use the same FHIR server instance? The same FHIR server/type with different instances? Or is this testing just to help NCPI systems choose a FHIR server offering that's "compatible" with other NCPI FHIR servers?
        - The latter for the platforms themselves, but for the "testbed" we'll probably use this to be more selective in terms of what we support for development/collaborative purposes
        - One possibility for near term would be to have an NCPI terminology server with individual fhir servers to host data islands with some tooling to facilitate pulling data from each of the islands
    - General discussion
        - Support for FHIR to OMOP flow (focus on downstream analytics)
        - Support for terminology server
        - Support for central repository - or registry for terminologies

- Optional Next Steps for Attendees
    - Provide feedback on Implementation Guide draft
    - Join the bi-weekly IG meetings (every other Tuesday from 1:00-1:30)!

1:00-1:20pm – Working Group Updates continued

Outreach and Training - Speaker: Anton Nekrutenko (PSU)

Notetaker: Stephen Mosher

Notes:

- Reviewing NCPI Portal
- Training docs for each platform
- Reviewing NCPI Dataset Catalog
- Not meant to be a cohort builder, but solve the problem of "what datasets exist?"
- Can choose datasets based on platform, dbGaP consent codes, data types, related diseases
    - Can download a TSV file of resultant dataset tables
    - Can share a link by "Copy URL"
- Closer look at "diseases"
    - Currently MESH terms
    - Working w/ Asiyah Lin to map those to MESH terms to Ontology terms
- Thank you to Dr Asiyah Lin for helping to connect all the groups
- Question - will you all implement a "handoff" feature to workspace environments?
- Answer - this is incremental, will focus on building out more once the dataset catalog is production ready
- Question - will be great for those that do not yet have dbGaP access, this will be helpful to see what studies exist.
- Answer - yes, we have documentation on how to request access. Would be great to automate services to ensure where access requests stand at any point in time.
- Comment - dbGaP is available to interface where needed and will be happy to help with those that want to build tools.

3:00-3:20pm – NCBI's Journey in Support of a Federated Cloud Data Sharing Ecosystem

Speaker: Mike Feolo (NCBI)

Notetaker: Durga Addepalli (NCI)

Notes:

- **Overview -** Distributed controlled access data
- Architecture

    - GPAs, Submitters and DACs perform NCBI registration and save in cloud storage and approved users
    - Study Registration
        - Done by NIH GPAs and PIs
        - OMB/PRA approved form, certification, data storage
- Interactions with NCPI - Consent groups- unit of approval is through consent groups
- GPAs have easy access to it
- dbGaP - Submission and Processing
    - QA, QC, configure data for release
    - Study metadata, sample and subject ids, phenotype, molecular data
- NCPI interactions
    - Provide study level metadata accessioning
    - future interactions - build FHIR server, API access to metadata
- Authorized access system - 3 critical functions
    - PIs request data - gateway to provide access to data
    - Current interactions - telemetry reports
    - Future - RAS and coordinate with NCBI (release systems)
- SRA
    - SIs, DCCs, SS
    - Biosample, Bioproject interaction internal to NCBI
    - Link accessions- Cloud locations
    - Current interactions- metadata with Cloud locations and telemetry reports
    - Future interactions-- API access and NCPI platforms might submit metadata directly to dbGaP
- dbGaP seq data in cloud
    - AWS and GCP through STRIDES
    - Old data in cold storage
    - Available through Cloud Data Delivery
    - Current interactions- Cloud locations are included in SRA metadata submissions and can be got from SRA run selector or using SRA toolkit
    - Future interactions-- get tutorials for users on how to get controlled access data from NCBI, Integration of SRA services with RAS
- NCBI RAS Development
    - NCBI does interact with GA4GH WGfor RAS passports
- Future
    - RAS authorizations
    - Include FHIR API with RAS
    - Will include dbGaP data into the cloud - provide that with RAS authorization

- Once RAS is implemented with all NCI systems- users can access data from multiple sites
- Transferring from one site to other is on the radar

Q&A

1. What is the model for Egress for dbGAP resources and governance required
    a. At NCBI Cloud- (Rodney) - cloud as backend subtitute for on-prem.Build up a federation model
2. Roadmap as data might be submitted to NCPI platforms - how can they be made to work well with dbGaP - how would users mint identifiers
    - most studies at NCBI are registered with a consent and also subject and sample identifiers therefore we have run level identifiers at SRA- natural evolution would be to cloud platforms would be submitting the metadata than the PIs.

Questions in Chat:

Question: to confirm, is it correct that identifiers (study, subject, variable, any others) are generated by dbGaP curators during submission and processing of study data, not as part of Study Registration? (this has implications for things like search and access of subject-level data for data hosted by other NCPI systems that may currently be depending on dbGaP identifiers from released studies). Do you have thoughts about how this might evolve if data is submitted to other NCPI platforms?

from Ian Fore (internal) to Everyone: 3:15 PM

Mike - does what you showed mean access to SRA data through the SRA toolset when the cloud buckets are owned by the NCPI platforms vs in NCBI owned storage?

from Kurt W. Rodarmer to Everyone: 3:20 PM

@Ian - SRA Toolkit is able to access any (simple) URL.

from Mike Feolo to Everyone: 3:21 PM

yes if we have accessions and cloud locations there is no reason we should not be able to have our tools work with these data.

from Ian Fore (internal) to Everyone: 3:22 PM

@Kurt - does simple in this case mean signed URLs which grant access to .

from Kurt W. Rodarmer to Everyone: 3:22 PM

Signed URLs work, yes.

from Kurt W. Rodarmer to Everyone: 3:23 PM

If we need to know any magic in the headers, then maybe not.

from Kathy Reinold to Everyone: 3:23 PM

Would love to see a minimal model at the hub level to promote interoperability.

from Ian Fore (internal) to Everyone: 3:25 PM

@Mike - makes sense - think we need to make sure it's working as expected. We hit glitches for example with 'requester pays' buckets. Basically - all sounds good - but could do with more collaborative testing across. this group.

from Mike Feolo to Everyone: 3:30 PM

We do not have plans to host an analysis platform we are focused more towards access to the data. hopefully we do not have to move the really large data.

3:20-4:20pm - 1 hour: Breakout Groups Report Back

Topic 1: Data harmonization and interoperability, including models, terminologies, mapping, provenance - Speaker: Chris Chute (JHU)

Notetaker: Tricia Francis (JHU)

Notes:

Report back slides:
https://docs.google.com/presentation/d/1GK1ALd7i-uqdjmd6750fGS5PCESpQjDJ0fHQnVRzNqA/edit?usp=sharing

CDMs are difficult. But you can have pragmatic derivatives. Ex: analysis - data projection… Three layer model: FHIR interoperability with specified semantics built as a super set of standard implementation guides and triangulation (FHIR to OMOP) third tier - annotation set/analytic set.

Three tiers - like how that is laid out. Realize it is a lot of work and most want to just be in the analytic layer. But to analyze across programs - there will be problems.

Similar questions between genomic and clinical.

From NCI - critical for data harmonization, but not at systems level, but at NIH. Programs are being funded and we accept what those programs are required to put on it in their format. We are just making the data available as is required.

Where is the forum where programmatic leaders can have that dialogue on what makes sense?

Topic 2: Search – Speakers: Kathy Reinold (Broad) & Steven Cox (RENSI)

Notetaker: Jay Ronquillo (NCI)

Notes:

Report back slides (which includes notes/takeaways from “Post-Breakout” Report Back Discussion) :

https://docs.google.com/presentation/d/1MvbOLXwDA3LFRKl91yGV2P-W0RU5KQ-pYjzhXwJWXS8/edit#slide=id.p1

Takeaways from Post-Breakout Report Back Discussion

- Search very timely because of increased interoperability
- Strong desire for practical demonstration of use cases
    - More than simply integrating datasets, can users search across these datasets?
    - Concrete use cases in next 6 months to demonstrate ability search/extract data across platforms
- Impact of data access (open vs. controlled) on ability to search
    - Before applying for access/authorization, can user find out how many samples are in dataset or which studies are applicable?
    - How to engage investigators while getting/waiting for data approval?
- Data harmonization and identifier creation vital for search as well

Topic 3: RAS interoperability – Speakers: Andre Paredes (UChicago) & Brian O'Connor (Broad)

Notetaker: Teresa Barsanti, Ann Van (Nimbus Informatics)

Notes:

[Work in Progress]

- See report back slides: https://docs.google.com/presentation/d/1Bi-JAL57LT9pZMJ7e8u9D4cWKpfmbIxX8witnxx2Gt8/edit#slide=id.gd6db545531_0_90
- Notes
    - Conclusions
        - Trust
            - Some systems (BDCat) will require mutual SSL certificate verification for client and DRS server
            - Repackaged Passports with intact RAS visas are allowed but not sufficient for all systems to identify the client
                - Give a route to adding non-dbGaP visas
        - Performance
            - POST of Passports must be supported by DRS and implementations

        - Downscoping of Visas is being designed now but is not sufficient to address performance issues nor will it allow passports to be passed via a bearer header token.
    - Policy
        - Can Governance and Policy group identify the systems that require mutual SSL certificate verification?
            - BDCat and ???

4:20-4:30pm - Conclusion of day - Sam Volchenboum (UChicago) and Tanja Davidsen (NCI)

Notetaker: Alan Zheng (NCI)

Notes:

[Sam Volchenboum]

- Thanks to all speakers, participants, and note takers, especially Tanja Davidsen for organizing the meeting
- Tomorrow (Tues, May 4th) Melissa Haendel will be the MC for the meeting with Tanja. We will start at the same time

[Tanja Davidsen]

- Thanks to Sam for being an excellent MC
- Speakers - please send your slides to Tanja by end of day tomorrow (Tues, May 4th)
- If you have not registered for this meeting, please do so (Registration link)
- Please take the poll for Fall meeting dates (Poll link)
- Some experienced difficulty joining breakout sessions - please use the WebEx app, NOT the browser version
- Conversations in Chat are being saved
- The meeting agenda will have all the links after the meeting