BioData Catalyst December Quarterly Meeting
Homepage and Meeting Notes

Tuesday, December 1 - Wednesday, December 2, 2020

We require all attendees to follow the Statement of Conduct and the Community Rules of Engagement throughout the duration of the meeting.

Zoom Information

Please note, we will use different Zoom links for Day One and Day Two.

Day One Zoom: bit.ly/BDCvqmDEC1

- Additional Zoom information
- Optional Fellows Poster Session Day One uses the same link

Day Two Zoom: bit.ly/BDCvqmDEC2

- Additional Zoom information
- Day Two Breakouts: You have been pre-assigned to one of three breakout rooms (PI, Developer, Fellows). See your assignment here. At the appropriate time, we will open the breakout rooms, and you will be moved there. Slack @Amanda Miller if you would like to be moved to a different breakout or if you are unsure of your original assignment.

Relevant Documents and Links

- **Virtual Meeting Slides**
- **Virtual Meeting Recordings**
- **Virtual Meeting Website**
- **BDCatalyst Slack Workspace** **(Face2Face Channel)**
- **Virtual Meeting Homepage and Meeting Notes**
- **Virtual Meeting Zoom Backgrounds**
- **Virtual Meeting Feedback Form**

Agenda (with links to Slides and Notes)

Tuesday, December 1		Day One Zoom
10:30-10:40 ET	Welcome and Housekeeping	Slides | Notes
10:40-10:45	Statement on Consortium Transparency	Slides | Notes
10:45-10:55	NHLBI Opening Thoughts	Slides | Notes
10:55-11:45	Cohort 2 Fellow Project Introductions (3 min each)	Slides | Notes
11:45-11:55	Questions for the Fellows from Consortium	Slides | Notes
11:55-12:05	***Break***	
12:05-12:25	Fellow Interest Groups: Updates	Slides | Notes
12:25-12:30	Fellow Interest Groups: Questions and Discussion	Slides | Notes
12:30-1:30	***Lunch***	
1:30-2:20	Partnerships and Interoperability: Updates and Discussion •Overview of partnership models (Becky) •BDCatalyst CICI (Steve) •NCPI (Stan) •Project 5 (Jon) •MIS-C (Jon) •PCGC (Jon) •CONNECTS (Becky) •Cure SC (Jessica)	Slides | Notes
2:20-2:30	***Break***	
2:30-3:25	Rethinking Supporting Users	Slides | Notes
3:25-3:30	Day 1 Recap and Next Steps	Slides | Notes
5:30-6:15	*Optional: Fellows Poster Session*	Slides | Notes
Wednesday, Dec 2		Day Two Zoom
10:30-10:35 ET	Welcome and Housekeeping	Slides | Notes
10:35-11:30	Working Group and Tiger Team Updates	Slides | Notes
11:30-11:40	***Break***	
11:40-12:40	Breakout Sessions (PI, Developer, Fellows) •PI: Risks moving into Production? (Stan, Becky) •Dev: Finishing Release 1, Preparing for Release 2 (Marcie, Brian) •Fellows: Improving new user onboarding and training (Chris E)	Slides | Notes
12:40-1:40	***Lunch*** *(and EEP breakout: 12:40 p.m. - 1:10 p.m.)*	
1:40-2:05	Breakout Report Backs	Slides | Notes
2:05-2:55	Consortium Project Management	Slides | Notes
2:55-3:05	***Break***	

3:05-3:15	External Expert Panel Retrospective	Slides | Notes
3:15-3:25	NHLBI Final Thoughts	Slides | Notes
3:25-3:30	Recap and Next Steps	Slides | Notes

BDCatalyst Virtual Meeting Roles

Role	Purpose	Assignee & Slack
Maestro	Master of Zoom Ceremonies. Contact Amanda for questions about Zoom issues, breakout rooms, or other general questions.	@Amanda Miller (amiller@renci.org)
Screen Sharing	Will share screen and advance slides.	@Julie Hayes
Mute Master, Raised-Hand Monitor, & Security	Will unmute speakers and mute non-speakers. Will note and lower raised hands. Will kick out bad actors. Contact Patrick if you notice suspicious activity.	@Patrick Patton
Slide Content	Will update slide content throughout the meeting.	Sarah Davis (@sdavis)
Moderator	Moderator listed for each agenda item. Moderator will prompt slide transitions during presentations and foster productive conversation during discussions.	Ingrid Borecki (iborecki28@gmail.com) Becky Boyles (@rboyles) Stan Ahalt (@stan)
Plenary Notetakers	All are encouraged to add comments to the Homepage and Meeting Notes.	@Marcie Rathbun @Tom Madden
Q&A Monitor	Monitor questions in #face2face Slack channel as well as Zoom Chat. Share questions from Slack and Zoom to the Homepage and Meeting Notes document.	@Chris Lenhardt (Zoom) @Chris Erdmann (Slack)
Time Watcher	Will try to keep us on time while still allowing room for important conversations.	@Stefania Knight (am) @Paul Kerr (pm)

Meeting Notes (Day One)

Welcome and Housekeeping

Statement on Consortium Transparency

- See PM Plan V2.0 -- Communications Management

NHLBI Opening Thoughts

- Alastair - Overview of BDCatalyst ***successes*** -- Fellows Cohort I & II; COVID data (specifically ORCHID) incoming; security assessment and authorization progressing; Jira tracking/management continuing to grow -- visibility into progress is critical; strides towards unified help desk (important for users)
    - Overview of BDCatalyst ***Challenges*** -- onboarding users takes a lot of hand holding -- disparate user experience -- clarity needed on user pathways; we need unified help desk, documentation, etc.; data access/usability -- most users are byod; ***data ingest bandwidth -- demand will continue to grow; interoperability is limited
    - Where are we now? Complete Phase I -- on the bring of Phase II (what do we need to do in order to be successful for ***USERS***) -- many questions need to be answered moving forward; ******single unified search is critical*** (genotype, phenotype, semantically based); unified user support -- true integration of data concierge concept and teams -- no ***visible*** “hand-off” across the ecosystem; how to enable NHLBI researchers to become known as having the most reproducible science? -- a challenge/goal for BDCatalyst
        - To answer all these questions and move forward, changes may need to be made; reassess what and how we’re doing things given changes at NIH and elsewhere -- focused on collaboration;
        - Challenges lay ahead but confident we can address and overcome

- Questions / Comments
    - Shu (Sue) Hui Chen: To speak to Alastair's point about data use, as of October 2020 NHLBI has had ~1000 Project applications, which constitutes ~9000 DARs just in dbGaP for the 2020 year
        - Q: Stan: How does that compare to other years?
        - A: Shu (Sue) Hui Chen: @Stan, it's gone up. 2019 had about 8000 DARs so we are about 1000 more this year // FYI it does not count the TOPMed Exchange Area applications only the released data

Cohort 2 Fellow Project Introductions

Alexander Bick, PhD

Melissa Cline, PhD
Jamie Murkey, MPH
Einat Granot-Hershkovitz, PhD
Xuefang Zhao, PhD
Diego Mazzotti, PhD
Randi Johnson, PhD, MPH
Yaling Tang, MD
Jia Wen, PhD
Pranav Rajpurkar, MS
Yonghua Zhuang, PhD
Brandon Lê, BA
Bo Li, PhD
Xu Zhang, PhD
Jinling Liu, PhD
Ravi Mathur, PhD

Requested Changes/Additions/Challenges:

- Availability of additional data (available on TOPMed, e.g. Freeze 9)
- Jamie Murkey - **Lessons Learned examples/use-cases helpful**
    - Randi Johnson: Agreed Jamie! Lessons learned would be useful for more efficient startup - Pranav Rajpurkar also agrees
- Cloud billing guidance (e.g. Terra)
- Ability to write, store within projects
- **dbGaP dataset structure -- naming system, type of data, etc. -- better metadata structure and/or documentation needed (spreadsheet-like format suggested)**
- **Earlier training for platforms -- hard to locate answers to questions regarding tutorials**
- **Videos for tutorials/use cases** (R video e.g. from Einat Granot-Hershkovitz)
- Read/write access to shared files (e.g. Seven Bridges)
- Start/stop of AWS instance
- Unified platform to share harmonized data
- **Onboarding "pipeline" for new users** -- new/future users, where to start, etc.
- **Clarity on STRIDE billing -- tutorials**
- **Clarity on file systems in workspaces, virtual machine and cloud**
- Pathways for Fellows to get involved w/ TOPMed working groups -- access to exchange area data
- Easier way to organize files in project spaces (e.g. transferring from data cruncher instance to a project)

Questions for the Fellows from the Consortium

- Alisa - Hackathon ideas and/or activities for getting Fellows up-to-speed; specific short-term needs to ramp-up?

    - Jinling - TOPMed WG involvement would be helpful. Many of us are new to the datasets and have had trouble working with some of the data. WG insights would be helpful for understanding the data better.
    - A: Einat Granot-Hershkovitz: Hakathon idea: maybe each fellow, from the first cohort, and maybe from the second cohort, can demonstrate the main way they utilize the platform. This way we can see the breadth of utilizations that the platform offers, and if anyone needs help in a certain utilization, they can later reach out to the experienced fellow.
        - Jinling Liu agrees
        - Sarah Davis to Everyone: @Einat, I like that idea. Let's work to use Cohort 1 and Cohort 2 examples as we onboard Cohort 3!

- Stan - Two questions/points: 1) We'll mine the recommendations for changes for common themes and address them as a group moving forward. 2) Many people mentioned videos, are they the preferred mechanisms? (vs. written articles)
    - Ravi - I like the visuals, which can include screenshots in an article. Prefer this over straight text however.
    - Diego - I like written material, I never have the patience to watch videos.
    - Stephanie & Randi also both prefer text with screenshots.
    - Yaling - prefer videos
    - Melissa - Prefer written material with images.
    - Brandon Lê - I think there's a third option that hasn't been mentioned yet: face-to-face onboarding meetings. Shoutout to Dave and Allison from Seven Bridges: they have been essential in getting me familiar with the platform, mainly through the technical trainings and meetings given to the cohort II fellows. I realize this requires scaling-up of the work they would be doing, but their work in making sure the fellows are acquainted with the platform was invaluable to me, at least.
- Q: Jason Williams: General question for any of the fellows who care to answer: If BDC was not around what might you have used instead? Would the research have gotten done? Would you not have attempted? Or would you have done this in some other way?
    - A: Einat Granot-Hershkovitz: I would have done the same research using a Linux cluster.
    - A: Sheila Gaynor: Cohort 1 fellow answer @Jason: would have done the research on local cluster with notably slower output but less start-up costs
    - A: Yonghua Zhuang: Provide AWS linux instance directly instead of being contained in Docker.
    - A: Pietro Nardelli: Answer for @Jason: I would have done the same research on my local cluster with GPUs. It would have taken longer but it would have probably be less expensive
- Jason Follow-up, is the opinion that BDCatalyst is preferable over other research methods? How is it better than other experiences?

    - Diego - In my experience, the system is useful for breaking technical barriers and is more approachable to people without the expertise required for other research methods. (Less technical folks can ask more complicated questions).
    - Melissa - I'm looking forward to using secured collaborative workspaces. This is not something I have access to outside of this system. Adding someone to our secure VPN is a long road, but this will be much easier.
    - Sheila Gaynor: @Jason: significant benefit for me is the scalability— I could use the methods to analyze locally, but doing it on BDCatalyst I am going to be analyze significantly more phenotypes at a much more rapid pace with reproducible and consistent pipelines (that others can then also use directly)
    - Alexander Bick: @Jason: data download from dbGaP is a major bottleneck with WGS data - and something that the cloud "solves"
    - Sheila Gaynor: Towards @Jason/@Stan's earlier questions— one aspect that has been transformational in our work has been data harmonization/collaboration (Cohort 1 fellow Kenny will present his efforts in the diabetes working group this evening), where researchers across institutions are working on this collaboratively in shared workspaces which we otherwise simply could not
- Q; Brandi: Question regarding reproducibility: how many of you plan on making your tools and analyses available in a workspace or similar as part of a publication?
    - Alexander - Yes, but people looking to reproduce your results will still need access to the same TOPMed data. This may be something NHLBI needs to address.
    - Ravi - I'd like to make my workflows available after the research, as well as interoperable with systems outside of BDCatalyst.
    - Randi Johnson: @Brandi, I'm not sure I know enough yet to determine what's best. Cohort 2 fellows may have better ideas? One I've heard is just exporting to Dockstore.
    - Stephanie Gogarten: @Brandi: I have a github repository with my phenotype simulation code. I haven't thought about making the associated SBG tools available, but I would be open to that. maybe exporting to dockstore would be the best way to do it?
    - Brandon Lê: @Brandi: I'm planning on making variant filtering/calling pipelines and machine learning models available. The variant pipeline is linked to TOPMed data, so I'm not sure how to make that available, but the ML models are theoretically easier to export.
    - Brandi: I think dockstore definitely solves this as far as getting access to tools. My question also related to the specific analysis journey (what exact data files, references, etc) and results. The response regarding data access is tricky but something that could be possible to solve
    - Brandi: Really clear guidelines about how to share for example Brandon's use case is key and is related to some of the workplace 3.1 efforts
    - Stephanie Gogarten: @Brandi for TOPMed data anyone wanting to reproduce would have to apply to dbGaP for all the same studies used in the analysis

(which in my case is many). One could illustrate the methods using e.g. 1000G data, but I don't know if that helps with reproducibility
        - Alexander Bick: Agree with @stephanie! We need a GRU Topmed dataset with a single application to all data
        - +1 Jinling Liu
        - Jon Kaltman: We need to think about how to develop novel dbGaP accessions for synthetically created cohorts, so someone doesn't have to apply for a large number of studies if they are trying to reuse someone's work.
        -
- Stan: On the Data Commons effort there was one study that was completed by a researcher who said he would not have done it if he did not have the tools. Can you say anything about how BDCatalyst feels for research?
    - Alex - Some projects require data so large you can't download it all, and you must perform research where the data is. BDCatalyst enables you to bypass this by performing the research without downloading to a local machine/cluster.
    - Yaling - Working with clinical phenotype data it is better to download (since they are smaller files), but I agree if you're working with larger files you should avoid downloading.
- Yaling - Suggestion for a workshop across Fellows Cohorts to discuss previous pain points, and how they were addressed/solved.
- Jamie - Adding content, lessons learned, etc. from previous Fellows cohort(s) to the onboarding process for new Fellows Cohort(s)
- Randi Johnson I agree that organizing a session where some existing fellows demonstrate their use of each platform during new cohort onboarding would be helpful to get a sense of capability, use, "pain points," etc.
- Kenny Westerman: I think a major issue here is what Yaling described: coming in with the promise that BDC gives you access to datasets like TOPMed, when in reality all of the same requirements for data application via dbGaP, TOPMed, etc. are still needed
    - Yaling: Agree, Kenny. It would be great if we have a page on the website explaining the pathways to access to the data: dbGAP, TOPMed, TOPMed workgroup, etc., including for what type of data, which access you have to work on...
- Ravi Mathur: Could the sharing of fellows experience on the platform be incorporated into the monthly fellows meeting? Basically have a couple fellows present formally about their experience
- Brandi: If there is a session where fellows demo pain points I'd love for our product people to be able to join and ask questions, see how you are working etc.
    - +1 Kira Bradford
- Brandon Lê: We could incorporate @Brandi's session on demo-ing successes/pain points into @Ravi's suggestion on doing them during the fellows meetings.
- Chris Erdmann: Do any of the Cohort 1 Fellows want to present on your experience in an upcoming Fellows Monthly Meeting? Volunteers?

- Randi Johnson: I have definitely learned from Cohort I fellows through participation in the CWL working group

Fellows Interest Group Updates

- Brandi: These interest groups look great and could be of broad interest beyond the fellows. As we expand the user community would it make sense to open these up more broadly?
    - Ben Heavner: Would the interest groups benefit from shared workspaces?
    - +1 Stanley Ahalt
    - Stephanie Gogarten: @Brandi I think it would depend on how big they get. If there are too many people, it might be harder to have the same level of open discussion we have right now. Not necessarily a barrier, but something to consider.
    - Caitlin P McHugh: @ben we've talked about that in the GWAS/TOPMed group, but haven't acted upon anything yet definitely a good idea to consider implementing
    - Ben Heavner: I can imagine it highlighting some challenging questions regarding controlling data access…
    - Kenny Westerman: Good point, Stephanie — larger groups would be better for developing docs and initiatives, but less amenable to discussion of specific problems for specific people
    - Ben Heavner: Oh yeah - a shared workspace focused on collaborative tool development could be separate from data access...
    - Stephanie Gogarten: and if the groups get bigger and more formal, I think leadership would need to move from fellows to platform reps… I'm not really prepared to manage a much bigger group
    - Ben Heavner: 2 kinds of workspaces: 1 focused on tool development (using GRU/open data); 1 focused on close engagement with data among people with common access…
        - Stanley Ahalt: Yes, @ben, the network effect that is occurring is fascinating, and will probably arc careers if we can sustain the momentum
    - Brandi: Agree Stephanie, putting some support in place to make effective user groups would be critical if they expanded. One of the amazing things that I think can come out of these user groups is expanding folk's network that they wouldn't be connected to otherwise

GWAS - Caitlin McHugh: Met a couple of times. Goals are improving functionality of GWAS within BDCatalyst and illuminating requirements for access to TOPMed data. Next steps are continuing monthly meetings, addressing Cohort 2 onboarding issues, and working towards a BDCatalyst approved workflow for efficient and reusable GWAS tasks.

CWL - Stephanie Gogarten: CWL is mainly for Fellows working on Seven Bridges to share tips, ask questions, and learn to write better workflows. Also discussing sharing and publishing workflows. Cohort 1 Fellows are helping Cohort 2 Fellows. Will continue meeting monthly. A recent question involved pulling code directly from GitHub. Another issue is determining expectations around Fellows publishing workflows and sharing internally with each other. Seeking clarity on how much this is a goal of the Fellowship.

Ingrid: Some Fellows included that in their application. Some workflows are generally usable and repeatable tasks would be helpful, whereas others are less important. For instance, it would be useful for a structural variant caller.

Stan: This may need to be addressed at a systemic level. There have been discussions of creating highly developed software and highly developed and harmonized data collections both publishable. May need to seriously consider finding a venue where we can support this, perhaps GitHub. Also need to give credit for highly used bits. This requires additional thought and discussion.

WDL Development and Terra Cost Estimate - Kenny Westerman: These interest groups are a good place to coalesce common questions and problems from Fellows. Meeting regularly to discuss new features and obstacles. Established Slack channel so Fellows can communicate directly with developers. Interested in piloting new features, including a Terra notebook for run cost retrieval and estimation. Next steps are contributing to the expansion of documentation around creating, publishing and refining workflows, probably focused on Dockstore, and determining best practices for FAIR workflows. Also continuing to develop a utilities collection within the BDCatalyst organization on Dockstore to store workflows with the potential to be widely used. They're seeking BDCatalyst or Dockstore stamp of approval.

- Ben Heavner: And for NHLBI to consider how those contributions might factor into things like funding evaluation…
    - +1 Stanley Ahalt

Cross Harmonization Studies - Yaling Tang: Goal is to establish an avenue for communication of the harmonization process. Meeting monthly to discuss pain points. Created Slack channel so Fellows can ask questions of experts, developers, etc. Aiming to establish environments for sharing harmonized phenotype data pipelines for Fellows. Created spreadsheet to track work completed so far. Also aiming to facilitate developing and publishing harmonization workflow. Next steps are creating a workflow deposit in Docker that can be imported to Seven Bridges or Terra. Need to consider how to access data. Hoping Fellows can contribute to documentation to aid future Fellows.

Fellows Interest Group Questions and Discussion

Would groups benefit from having a shared workspace? Could be a good place for initial sharing of workflow or small apps that can be initially released before being made broadly available. Could also help work out some of the kinks. Harmonization may also benefit from a shared workspace. This depends on data access or else it could be an obstacle.

Stan: As more users access BDCatalyst, there will be recurring questions so we should capture and pass on lessons learned in a reasonable format to people who are onboarding. Should also offer structure and support for groups that are meeting to share problems and solutions.

Ingrid: People have been helped by hearing solutions to common problems. Are there efforts to record and archive tips, pain points or solutions to benefit future users?
Stephanie Gogarten: There is a Google Drive folder called “BYO Tools” with documents that have ended up on GitBook. There is a flow from jotting down informal notes to utilizing them to help write official platform documentation.
Randi Johnson: Benefited from utilizing those. Working group is a great place to learn from Cohort 1 experiences. For these working groups, some training may be needed for new Fellows to be able to participate meaningfully.

Becky: Important to consider how to scale things like working groups as more users come onboard. Something scalable like Stack Overflow may be needed in addition to working groups.

Ingrid: Do any additional topics need a working group? No suggestions were provided.

Day 1 Morning Zoom Chat Link

Updates on Partnerships and Interoperability

Overview of partnership models

Proposed models identified below and within the linked slides:

- Interop across platforms -- common interests, e.g. Gabriella Miller Kids First, Anvil, etc.
- Platform integration -- bringing in existing platforms into ecosystem; e.g. U. Michigan Imputation Server
- Data generators -- anyone looking to contribute data to BDCatalyst; e.g. CONNECTS, PCGC, Cure SC, etc.
- App developers -- to expand capabilities available to the community; e.g. DICOM Viewer

Questions

- Brandi to Everyone: Conceptually I'm not following the major differences between ie the imputation server and dicom viewer
    - Stanley Ahalt to Everyone:I think it is precisely a prioritization issue.
    - Alisa Manning: I also grouped them together in my notes…
    - Rebecca Boyles: We had defined apps as something that would be 1 or multiple containers
    - Steven Cox: @Brandi - Imputation server's a big distributed system (Hadoop, etc) and is multi-tennant. DICOM viewer is a single container app a user can launch an instance of, use, and dispose of.
- Brandi: Do we anticipate that the imputation server type systems would always have their own authorization boundary? and conversely apps need to run within an existing boundary?
    - Rebecca Boyles: That is my assumption Brandi
    - Steven Cox: Sounds right to me Brandi. / Also, there's an actual RFC defining what an app would be but there's not an analogous RFC defining a general method for installing something of the complexity of the Imputation Server.
    - Rebecca Boyles: To be honest, I think many of us thought that it would be rare to bring in a platform but we are getting hints that is not necessarily so
    - Albert Vernon Smith: @brandi: Having integrated authorization available for the Imputation Server is on our to-do list; also, sharing imputed results cross platform

BDCatalyst CICI

- AnVIL interop
    - **Has**: Performed basic analysis on each platform.
    - **Will**: Utilize interop frameworks to improve sample analysis.
- GMKF interop
    - **Has**: Piloted Nimbus data ingest script and identified key areas of improvement.
    - **Will**: Index data in their Gen3, NHLBI systems to use from there.
- UMich Imputation Server
    - **Has**: Migrated MIS to the cloud, aligned website branding to BDCatalyst's branding, and optimized cloud infrastructure.
    - **Will**: Align with BDCatalyst's Authentication/Authorization scheme.
- Deloitte Cloud Cost Model:
    - **Has**: Run experiments to determine ideal workflow setups for tools with anticipated high usage.
    - **Will**: Present a final public report with their findings.

NCPI

- Community Governance
    - **Has:** developed Five Principles for Interoperating Data Platforms and white paper that provides definitions and a series of questions to determine platform adherence
    - **Will:** continue to find resolution to blockers, work towards approval of policy, and develop "trust" between platforms
- Systems Interop

    - **Has:** Built consensus around standardized handoff mechanisms & data access. Aligned (as possible) IC funding to support interop.
    - **Will:** [2021 H1] Investigate PFB &/or FHIR question; help researchers achieve use cases; harden solution / attract more portals. [2021 H2] Enable some form of "compute across systems" to solve folks blocked by pipeline description languages; support additional researcher use cases.
- FHIR
    - **Has:** Piloted mapping Kid's First PCGC data & AnVIL CMG data to FHIR; developed API for server that can be queried and linked to other services (Monarch Explorer); created and tracked NCPI FHIR model (Profiles and Resources)
    - **Will:** Continue mapping of diverse data types into FHIR; establish interop between separate FHIR servers
- Training/Outreach
    - **Has:** Portal being developed
    - **Will:** Develop concrete plans for 2021 in 12/1/2020 meeting
- NCPI Governance
    - **Will:** Determine rules for developer access across multiple ecosystems

Project 5

- COVID-related
- Includes (not not limited to): BDC3, All of Us (IOD), & N3C (NCATS)
- **Has**: Data gathering/generation. Working on Hash IDs for data linkages.
- **Will**:
    - December 10 - BioData Catalyst will provide a demo to the Project 5 group
    - January 13 - BioData Catalyst will be part of a group of demos to the NIH COVID-19 Data & IT Steering Committee

PCGC

- **Has:** NHLBI has provisioned a cloud bucket for PCGC
- **Will:** PCGC to upload data and start to migrate tools to cloud.
- PCGC is a paradigm for an NHLBI funded group that wants to migrate both data and tools into BioData Catalyst. We probably want to think strategically about how to make this a reproducible process.

MIS-C

- **Has:** Still in planning phase. Studies are up and running.
- **Will:** Meeting on EHR extraction scheduled for next couple weeks. Meeting with platform developers and study investigators being planned.

CONNECTS

- **Has:** Regular team meetings to understand studies and projected data timelines, captured in the draft study index. CONNECTS ACC has provided feedback on the Data Generator Guidance as well as the Data Ingest Form. Have worked with the C3PO study to complete dbGaP study registration in anticipation of May-June data submission.

- **Will:** Release of Data Generator Guidance. Creation of Tiger Team to include ACTIV 3 and 4 clinical trials.

Cure SC

- **Has:** One dataset currently available in separate instance of PIC-SURE. Moving forward all Cure SCi related data will be ingested directly into BioData Catalyst. In addition to data, Cure SCi wants to create a community for Cure SCi researchers and might be interested in how BioData Catalyst can assist in that effort.
- **Will:** In progress, Cure SCi datasets prioritized by DRMWG for ingestion into BDCatalyst.

Discussion

- Other types of likely collaborations?
    - [Ben] Omics data Freeze 9 & future Freeze 10
    - Robert Grossman: We may also want to work with COVID and imaging data from https://www.midrc.org/. It will be opening shortly.
- What processes or guidance are we missing to engage efficiently and responsibly?
    - [Josh Bis] Process for permissions for users to access the data - still need to solve this in the system
    - [Becky] Process for onboarding whole other platforms? (not just their data)
- How do we think about prioritizing collaboration opportunities
    - [Brandi] We need to approach these new opportunities with the user in mind and how they will be using the system / data - ROI for end user community
        - [Becky] flesh out the use cases in the beginning of the collab process
- What portions of these partnerships can be automated or made a self-service?
- Sweta Ladwa:I believe one of the goals for BDC was to also be the home for all NHLBI supported/funded research data, so how does that fit into this discussion?

Introducing the Updated User Experience Coordination Group (UECG)

Collaborative brainstorming and prioritization

- BioData Catalyst UECG Areas of Focus Collaboration Board: http://bit.ly/UECG_B1
- Outreach Activities: http://bit.ly/UECG_B2
- BioData Catalyst UECG Success Metrics Board: http://bit.ly/UECG_B3

Brandi: BDCatalyst is in phase transition. Scope of UEWG had grown dramatically and activities in other groups not always tightly coupled. An iteration is proposed. 4 key points: provide coordination place for team decision making, complement TCM's focus, develop strategy & advice and coordination of decision making, and spin off smaller working groups and tiger teams for specific deliverables.

Focus areas: 1) Outreach - events, workshops & content generation, publications.
2) Documentation, tutorials, user pathways, help desk & FAQ - many of these are underway but can be connected.

3) User management and user responsive continual improvement - comprehensive strategy from onboarding through deactivation. Need clear and reusable processes. Includes feedback and recommendations.
This list is not comprehensive. Important to understand priorities.

Chris Erdmann: @Alisa @Brandi I'm not sure if we address the externally facing, community coordination in the charter, so something to note I believe. Lauren's comment had me thinking about this and also some of the conversation the Fellows prompted with the user groups.

Alisa: We're dedicated to listening to what people are experiencing and want to empower researchers.

Josh: This is a great way to refocus efforts. User feedback remains an important component.

Ashok: Will this group be involved in onboarding new users? Which aspects cover that?

Brandi: User onboarding would be contemplated here, but the targeted outreach component is where specific researchers will be identified. Groups should work in tandem.

Jack: Approach provides more focus and helps identify gaps. Will also streamline internal communications.

Lauren: This all makes a lot of sense. Some good processes are already in place. This group can be expert consultants.

Brandi: Charter is drafted. Roadmap of deliverables. Subgroups will work on deliverables. Main group provides feedback and approval. Brandi & Alisa volunteer to Co-Chair. Seek to integrate and evolve existing tiger teams and working groups. Will identify additional focus area leads to develop strategy. Transparency is goal. Not required to attend UECG meetings to participate in subgroups. Co-Chairs of UEWG have done a great job so far and resulted in many high quality products. Hope they will continue to participate.

Collaborative brainstorming and prioritization - 3 "Retro Boards": Areas of Focus, Outreach activities, and Success metrics. Review and comment on existing cards and add new ideas. Each person can vote for 6 things per board.

Chris Erdmann: Nice activity @Brandi @Alisa

Alisa: High-level pathways inform users how to approach particular analysis. Brainstormed 5-6 narratives describing how a user solves problems.

Josh: Are analysis workflow pathways more specific?

Alisa: Yes, focus is different.

Brandi: Workshops are top vote. Discuss why workshops are high priority.

Jennifer Brody: Guided journeys are very beneficial, along with the opportunity to ask questions in a group environment.

Ashok: Workshops are a good forum to meet a diverse group of users and learn their needs.

Brandi: User pathways is one of top votes.

Beth: Helps describe how services interact and provides intro to ecosystem's complexity.

- Tiffany Miller: +1 Beth

Brandi: Comprehensive onboarding resources is top vote.

Josh: Onboarding and user pathways seem to go together and should be clearly linked. This will help illuminate how pieces should fit together. Workshops are probably also related.

- Jack DiGiovanna: +1, tutorials seem to overlap in the Venn diagram too

Tiffany Miller: Workshops shouldn't be the main way people learn. Focused on community. Creating scalable resources should be a priority.

Josh: Workshops can also be a learning experience for the consortium to learn what is missing and needed.

Brandi: This will all be used as valuable input. Next board is Outreach Activities.

Alisa: Model this based on other initiatives and successes.

Stan: Is there something more scientifically oriented than Medium for posting?

Lauren: Publications can have different meanings.

Jason: Continue to be careful about having too many things on the to do list

- Chris Erdmann: Yes, yes Jason!

Lauren Hochman: +10000 to Jason on board 2

Stan: Small booth at ASHG could be beneficial. Posters would be good idea.

Ingrid: Would be great to have a 3-hour workshop.

Workshop hosted at targeted university got the most votes.
Stan: This and hackathon could be joined. Could be good for future quarterly meetings too.

Jason Williams: University events can be positive. Also the XSEDE campus champion model could work at places with a high biomedical research footprint

- +1 Rebecca Boyles to Everyone (3:16 PM)
- +1 Jack DiGiovanna to Everyone (3:17 PM)

Mitchell Flagg: particularly if targeted to institutes/departments at universities of interest w highly interdisciplinary groups and projects - can be a great opportunity to engage people across silos

Brandi: This can build a community of people supporting each other.

Alisa: Number of citations in publications got most votes. Number of funded projects with budgeted compute funds is second place.

Jason Williams: Re: citations, when papers are being written the researcher probably hasn't used BDCatalyst in 2-3 months. Re: reproducibility, most users could probably not execute on a data management plan before using BDCatalyst. We should capture user transformations to learn the benefit BDCatalyst provided, how they're using tools in a new way, what new behaviors they can now do, and how they're now able to succeed. This isn't a traditional metric but would be very impactful. What's the minimum viable reproducibility?

Stan: Full blown reproducibility will be very tough but a possible question is whether the research is reproducible within a month.

Ben: Perhaps if a person composes a workflow for their analysis, we automatically make a citable DOI for that workflow. Then if we have a list of the DOIs of workflows that people are using for analysis, we can track things like the number of DOIs generated and the citations to those DOIs, and we can see which are the most popular workflows and tools. This is a way of reputation building for particular analysis flows and would also give credit to the developers.

- Josh Bis: +1 citations are going to be a lagging indicator on the scale of months to years
- Jason Williams: +1000 Ben
- Beth Sheets: Dockstore has a DOI feature
- Chris: I shared the neurips checklist with Becky earlier as a example https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

Alastair Thomson: Great exercise Brandi!
Josh Bis: nice format for promoting discussion
Rebecca Boyles: Thank you Brandi and Alisa
Brandi: Thanks for the great participation everyone!

Retro Boards
Area of Focus "Retro Board": http://bit.ly/UECG_B1
Outreach activities "Retro Board": http://bit.ly/UECG_B2

Success metrics “Retro Board”: http://bit.ly/UECG_B3

Day 1 Recap and Next Steps

Alistair is interested in moving in a slightly different direction and says change is coming, but it shouldn’t be off putting. There are a lot of challenges to deal with.

Progress of Fellows is impressive. They’re critically important to ecosystem and we should continue benefiting from their energy and insights. We need more tutorials and they need to be easily findable. Need to be able to predict cloud costs. When data isn’t available it causes frustration. Need to help organize their tools and workflows - will likely need a workshop.

Now that we have basic functionality, we need to meet expectations of partnerships. Getting data on and accessible is a key issue, as is getting tools. Need to understand the entire spectrum of technical difficulty to describe to leadership why some things they’re asking for are challenging and costly.

It was helpful to think through UECG priorities. We need to make incremental progress on reproducibility, including writing an article.

Day 1 Afternoon Zoom Chat Link

Optional Fellows Poster Session (recording)

Michele Daya, Diego Mazzotti, Jean Monlong, Fayuan Wen, Kenneth Westerman

Meeting Notes (Day Two)

Welcome and Housekeeping

Working Group and Tiger Team Updates

- Tools & Apps
    - App Dev Guide 1.1 in progress
    - Refresh tools and workflows pipeline
    - Best practices, trust levels and code signing in-progress
    - Need dev environments -- moving from hypothetical to concrete solutions; understanding of chains of trust levels with enforcement
- Data Access
    - Data Access across platforms -- best practices becoming available

    - DAR and IRB updates forthcoming
    - Ongoing collaboration w/ DRMWG and other WGs
    - Considering sunsetting DAWG in favor of a Tiger Team/on-demand approach
        - Alison - Addressing user support for ongoing data access issues; develop templates for troubleshooting data access issues
    - Data Use Agreements per Dataset vs. standalone policies
- DRMWG
    - Ingestion priorities identified for the remainder of 2020 and Q1 2021 (in-progress)
    - ORCHID ingestion current priority → CureSC and BioLINCC training data to follow
    - Increasing ingestion and efficiency is a priority → NIMBUS script successful test-case with PCGC
    - Evaluate lessons learned, track progress in Jira (and support Google docs), continue to actively engage NHLBI Cloud Services Team and DHWG
    - Prioritize data quality and security -- establish data versioning protocol
    -
- Data Harmonization
    - Collaborating with Fellows Data Harm Interest Group to guide/inform future efforts
    - Goal to develop metadata ecosystem that's agnostic to the platform data model
    - RFC Draft 8 on file-level minimal metadata out for comment → priority for consortium adoption/use
        - Expand thereafter -- platform developer and user considerations
        - Agreement/path forward needed around platform interaction
    - Tool development to make data more FAIR ongoing
    - Q: Alisa Manning: Ben — can you describe how a metadata API would be used?
    - Q: Yaling Tang: Would the new meta-data affect the reproducibility of the existed Terra workspace?
    - Alisa Manning: (I'm trying to think about how we currently keep track of data files accessed through dbGap, and it's a manual process… so such an API could be very useful)
    - Ben Heavner: Interesting thought, @alisa! I think the idea of metadata API(s) in BDC is early phase enough that it could go in lots of different ways. Gen3 currently has a metadata service that allows software to use a GET method to query for metadata associated with a GUID and returns a JSON blob with some limited fields. I know that SBG also has tools to work with metadata on its platform. I don't know whether the gen3 metadata API supports any PUT methods, or ways to reference other APIs - so that's one direction I'm interested in discussing. We currently obtain metadata for single-subject files from a dbGaP API - if there are other external metadata sources, it would be good to identify them and develop workflows to interact with those metadata sources and propagate it across BDC. So much of the metadata work is internal engineering/architecture. The primary user-facing work cases at the moment relate to adding metadata (such as variable tags) or querying existing metadata

(searching and filtering over metadata elements)

I'd like to learn more about what you're thinking about with respect to "currently keep track of data files accessed through dbGap, and it's a manual process"

- josh bis: e.g. phenotype files from the exchange area are obtained manually, potentially off-platform and can appear as BYOD.

- Alisa Manning: @ben, I was trying to understand more about what you meant by meta-data API, and understanding that it's internal engineering/architecture is helpful.
- Stephanie Gogarten: Kenny wrote a WDL workflow to fetch dbGaP data: https://dockstore.org/workflows/github.com/kwesterman/fetch-dbgap-data-workflow:master?tab=info
  - Alisa Manning: Yes, but please read and understand how this WDL should be used (i.e. locally or within a workspace that's not shared with collaborators)
  - Stephanie Gogarten: Even collaborators who have permission to use that dbGaP data? Are the repository keys unique to each downloader, or shared among the same dbGaP application?
    - Kenny Westerman: @Stephanie, the key is unique to each application as I understand
    - Kenny Westerman: So I might download using that workflow in a workspace shared only with our dbGaP application (using our application-specific key), then import from there into a data harm workspace shared among many collaborators with global TOPMed approval

- Integration Testing
  - Suite of tests running for Team Ca; scoping work for Seven Bridges and PIC-SURE → continue work to build connections and establish further integration testing
  - Q: Alisa Manning: Kevin — regarding the Integration Test Team, could the 'test endpoints' be aligned with our ecosystem metrics?
    - A: Kevin Osborn: @Alisa yes, we can report on tests run, successes and failures
- Change Control Board
  - WP3.0 Change Request completion (target 12/21) for documentation/reference
  - CR Process in Jira continuing to refine/improve
- Cloud Cost
  - Cloud Credit Request Form (and guidance) available
  - Review Committee established → may need to expand
  - Surveys planned around cloud credit feedback
  - Timelines for recovering unused credits; data costs; etc.
  - Working on Jira/Freshdesk integration
- Load Testing

    - Points of integration with other testing efforts/teams to be identified/defined
    - Test cases identified;
- Manuscript
    - Ongoing efforts for paper development; annotation explorer for rare variant association testing; exploratory data analysis via PIC-SURE; image analysis
    - Draft → solicit feedback → general consortium-wide feedback → publication
- User Engagement
    - Transition to coordination group; BYOTools developed in coordination with Fellows; user pathway development; cross-platform metrics progressing; “open doors” checklist drafted; help desk and FAQ guidance → sub-groups established to move forward
    - UEWG → UECG; subgroups to push efforts forward
- Help Desk Updates
    - SLAs being drafted (focused around response times); table to host this information
    - “CC” process for copying individual platform Help Desks to talk to each-other;
    - Developing training and documentation around the above

Breakout Session Notes

PI Breakout Notes - Moving from Development to Operations

Moderator: Becky / Stan	
Notetaker: Patrick	
Attendees: Ingrid, Becky, Jason W., Kira, Tim H., Donna A, Jon K, Asia, Benedict P, Stan, Alisa M, Brandi, Anthony, Albert, David M, Ash, Ben H, Robert C, Chip, Paul A, Alessandro, Jessica L, Robert G, Warren K, Allison H, Liz Wagner, Ashok	
Topic Ideas	Action Items
Imaging Data Needs	
Smoothing Data Access	
Platform / GitHub Integration	

Notes:

Upcoming Challenges (moving to phase 2)

- Kira - **imaging data** and the need to identify user needs. Not sure we have a way for users to get to imaging data easily.
    - David - The file size and network constraints can cause issues, as well as headers differing.

    - Kira - Concerned over the ingest of COVID-19 data. We should develop a plan of action for ingest before uploading to avoid needing to retroactively fix issues caused by a rushed/messy ingest.
    - David - do we have a definition of <u>how</u> imaging data should come in?
        - Intention was to have data de-identified working with Deloitte and their algorithm.
    - Bob - We're working with NIBIB to bring in COVID data
    - Brandi - Two things to think about related to imaging is 1) It would be nice to have visibility to what data is coming down the pipe, 2) what user communities will be interested in the data and cater to them as soon as possible.
    - Warren - Communication in a large group like this is hard, but it's important for everyone to have an idea of what each party is doing.
    - Stan - I don't see how we're going to get the various groups/interested parties together to agree on a common set of tools/techniques.
    - Ashok - How do you actually do machine learning on large datasets? Right now on BDCatalyst there is a limitation on running large datasets. We have been wondering if there's a way to use WDL or CWL to get this working, but we're waiting on imaging data access.
    - Kira summary: The mechanism for ingesting data is not well defined, and unsure we'll be able to provide user access as quickly as we wish. We do not need to reinvent the wheel, but there are a level of agreements that need to be made across the consortium and by NHLBI.
        - Jon: NCI is a main concern and have a meeting scheduled with NHLBI in mid-Dec. To discuss data ingest.
        - David: Suggested speaking to Ron Summers to discuss ML on large datasets.
- Ben (in chat) - Developing / smoothing user data access process. Agreement by all that this is a concern for them as well.
    - Jon: I think this has to be wrapped in the idea of sharing derived data on the platforms. Not just a matter of creating a synthetic accession, but also where you place the data. (Want it all in one place, question remains where this gets stored and how access is controlled). This also relates to Jason's point that a top-level goal should be creating the ecosystem as a paradigm of science, and show that with BDCatalyst you can do things that were not possible before.
        - UN of sharing data & workflows back into the system should be near the top of our priority list.
    - Ben - In the near-term I don't think there is a group of technical people charged with thinking about data access and DAR's. There are some lightweight opportunities for smoothing out this process via apps or forms. Not sure if there is a UN or Feature in WP3.1 focused on this.
        - Alisa - the dbGaP access mechanism and current DAR structure is a big concern. It's unclear how users will expand beyond this as their research expands.

        - Jon - A lot of IRB process is not necessary, and a lot of this effort is about educating staff to be as open as possible (but not go beyond the consent). We would also encourage combining data sets and create a new accession, which would have data access fall under the most strict limitations. (Allows users to apply once, to the most strict and move from there).
    - Paul - BDCatalyst should also be a space where people could combine and share datasets via tools such as Jupyter notebooks. Ex. of the ORCHID study and how one of the most valuable elements they published was the code they used to perform their workflow. Suggestion that studies could include at the end their code or a link to their methodologies. Streamlined reproducibility.
        - Warren - Agreeing and adding to this use case, all reviewers should be able to do this review. Encourage transparency for methodology in BDCatalyst.
        - Stan - Restate: We have technical limitations, but if we could set up a prompt at the point of review for people to provide their code. Also for any people to get data access we need to go back to dbGaP to get permission...there needs to be a clear process for describing how a user gets access to data.
            - Restated point that we need to develop the ability to have mechanisms in place to grant data access without too much back and forth.
            - Alisa - I'm concerned we'll have a repeat of TOPMed's current situation (researchers who have ideas, but don't have access and don't understand the full picture of what steps are required to obtain full access.)
- Becky - What is the capability of the various workspaces to interface with GitHub?
    - Alisa - I've been using this already as an Investigator.
    - Anthony - We usually do that via DockStore. We have in our backlog the ability to mint via DOI.
        - Brandi - This would be great, and would be a gateway of quality of code writing.
        - Ben - Why not mint via DOI as a de-identifier?
            - Benedict - There's a not-small cost element associated with it. My belief is all flagged releases should have DOI, but not all commits.
    - Becky - sounds like we have some tools for reproducibility, but need to work on messaging (example of not everyone knowing about DockStore/GitHub integration).
    - Brandi - I also think for the manuscript we're planning on running some analysis, it would be good to have these in public workspaces.
- Stan - Where in our priority list is the ability to drain the swamp of data access issues? (Directed to Jon/Alastair)
    - Jon - I think it's pretty high in priority. Critical issue in terms of saving money/effort by reducing the need for additional tools/resources for reproducing studies. If we

have a use case to drive this it would be great to demonstrate (Alisa ex.) and shake out policy issues that should be addressed.
    - Alastair - I'm hearing conversations of interoperability that I've never heard before. There is a motivation to drive this and move it forward and we should use it.
    - Becky recap: These two things could fall under the new UEWG structure: the creation of a FAQ on Fellows contributing their workflow to GitHub, and the creation of a TT to support Alisa's efforts to drive this data use case.
        - Brandi: This work is included in WP3.1, but suspect a lot of this work is going to be involved in policy and less in development work.
- Any closing concerns?
    - Paul A. - Concern on the BDCatalyst slides being used not representing what is actually in production data-wise. Want to be sure the deck is accurate.
        - Stan - A large chunk of the blocker is getting agreement from all the teams, and making sure NHLBI/Comms is happy with what's on the slide. We need to ID a mechanism to speed this up.
        - Becky - We're also happy and open to jumping on a call to discuss any proposed changes.

Zoom chat notes:
In response to Kira's discussion of limits on number of object files in a workspace:
Brandi - Should be 250k+/workspace in next month or so and then scaling to 1m+. However at that scale within one workspace it becomes really important to properly use metadata etc to find what you're looking for

Ben - I have 2 user-oriented issues to raise: 1) do we have effort dedicated to helping smooth the data access request process (passports?); 2) how do we work with data generators to plan for the next 6 months (thinking of BDC working with TOPMed in planning for WGS freeze 9, 10, etc; and -omics, other data). As the TOPMed liaison, I don't know who to approach to collaborate on planning for the TOPMed program data generation and management for the next 3 quarters.
Jon K - right now that person would be Sweta. Eventually, that will be the Data Management Core.

Imaging-related chats:
Allison Heath - Is there any link to the NIBIB resource or information about it - couldn't find anything with a quick search, because also agree there's a lot out there to reuse with imaging?
Robert Grossman -https://www.midrc.org/ (it goes live in a few weeks)

Data access request/registration related discussions:
Alisa Manning - RISK: Jon's suggestion for opening access to GRU data might not align with the current data access structure for many NHLBI projects
anthony - Would it be feasible to move away from local IRB and move to a centrlized IRB?

Ashok Krishnamurthy - Would publications be willing to accept as reproducible research data and code that exists in a system that has access restrictions?

Rebecca Boyles - It would likely have to be in GitHub. But I wonder if we could mirror it

Warren Kibbe - @Ashok They already do - they take it on faith that the data and the analysis was done, because they can't reproduce

Alastair Thomson - @Ashok - I think they are going to have to. We discussed COVID pubs from N3C data where the data cannot leave the N3C enclave. Need to be able to capture the data and analysis, freeze it and mint a DOI to reference it I think.

Warren Kibbe - @Ashok. The other side is clarifying the IRB requirements makes this much more tractable. And perhaps there needs to be a dbGaP process for granting access to reviewers for supporting data reproducibility

Ashok Krishnamurthy - @Alastair, @Warren, great, great suggestions. I wonder if a publication results from BDC data/code, if there is a way to create a lighter burden ability to reproduce based on a DOI as suggested by Alastair

Ben - So if Alisa needs to register the study with dbGaP first, how can we ease that process?

Alisa Manning - How is "study" defined in this case?

Stanley Ahalt - @Anthony can you point to the process description for the studies that are making this easier?

anthony - the two studies I mentioned are emerge and ccdg. Let me look for the documentation
anthony - this publication describes it for emerge
https://genome.cshlp.org/content/21/7/1001.full.html

Rebecca Boyles - I could envision having a reproducibility portion of the website where we can start to highlight efforts by means of communication.

Jon Kaltman - @Becky - great idea. And I really like the idea of modeling this in the marker paper.

Alastair Thomson - The concept is to create a "Freeport" for data where the policies and security controls are aligned to expand the concept of "enclave" to encompass multiple systems.

Warren Kibbe - It becomes a matter of precedent. If we can do it for COVID it becomes possible to do it more generally

Chat Log

Developer Breakout Notes - Finishing Release 1, Preparing for Release 2

Moderator: Brian / Marcie	
Notetaker:Tom	
Attendees: Danielle P, Tom M, Hannah H, Marcie R, Jacob P, Andrew R, Kevin O, Jack D, Brian O, Steve C, Alison L, Howard L, Josh B, Alex V, PJ L, Isma G, Tiffany M, Dave R, Teresa B, Murali K, Lon B, Michael B, Charles O, Brian H, Chris L, Ash, Noble D, John C	
Topic Ideas	Action Items
Dev Breakout Slides	
EasyRetro	Discussed "risk" Features: ● (Interoperable System v2) Feature A: RAS support ● (Phase II Hosting New Data) Feature D: Streamlining and multi-threading ingestion of new data sets ○ ***Focus on versioning concerns → DRMWG ownership R2*** ○ Consider development of Feature Planning Doc ● (Interoperable System v2) Feature D: bulk FHIR ingestion ○ Discuss on future TCM ● [Ca] Gen3 to add a Study Viewer page to Windmill ○ [**Marcie**] Alessandro/Gen3 note where/what the blocker is (Marcie to follow-up) ● Scaling for DICOM Images ○ TCM as forum for future discussion/solution ○ Imaging TT (potential)
(Interoperable System v2) Feature A: RAS support	
(Interoperable System v2) Feature E: Connect to NCI Imaging Data Commons	
(Coordinated DevOps) Feature F: Penetration Testing, Security Audits, and other security considerations	

All of the Release #2 Features ...	

Notes:

- Brian's Notes
    - Release 1
        - RAS
            - AuthN not at risk for R1, in a good place -- Alex
                - Fix mid-December
                -
            - What are we concerned about?
                - No concerns really for R1
        - Streamline upload
            - Some progress in R1 for staging data and automation
        - FHIR Ingest
            - Add this to the TCM agenda
            - **Focus on ingest of case report form**
        - Study page in Windmill
            - This epic is blocked… but it's an internal block.
    - Release 2
        - RAS
            - PIC-SURE for R2
                - **Multiple IdPs** -> this is an issue with RAS team...
                - Want to have better ability to test with **real telemetry** data… an update to come on this
            - **Gen3… scalability concerns about Visas… a lot more API calls**
            - **How to have multiple passport brokers**
            - **How to secure APIs beyond file access** e.g. Google Health API
            - **Discuss In**: TCM for BDCat specific issues
        - Streamlined upload
            - How to handle versioning with new data
                - We don't get historic auth information from dbGaP!!!
                - **Are we expected to maintain old versions of data?**
                - **Talk about this in DRMWG!!! Should we make a Feature Planning Document???**
                - **Consider splitting provenance and versioning into two distinct topics**
            - Improvements to make with different datasets in parallel… releases per dataset
            - Provenance in addition to versioning
            - **Discuss In**: DRMWG
        - Penetration testing
            - NHLBI will use their team for penetration testing

            - Each group shares results via POAMs too. (plan of action and milestones).
        - DICOM images and scaling
            - Will we have data (these are COPDgene data files) by 3/31?
            - Could be a billion files in the long run
            - **Google Health API… claim it scales. Apps to access it them?**
            - **What about Gen3 for IndexD?**
            - **What about PFB handoff?**
            - **Maybe DRS bundles would help?**
            - **Discuss in**: TCM, focus on the **use cases** in order to understand the solutions. Imaging Tiger Team would be appropriate
- Tom's Notes
    - What Features are at risk, causing concerns, and/or need further collaboration/development work?
        - (Interoperable System v2) Feature A: RAS support
            - Gen3 - R1 on target (auth/login); potential scalability concerns
            - PIC-SURE R1 -- collaboration/initial discussions; target R2 for release on PIC-SURE; PIC-SURE needs live telemetry data for testing -- potential risk if never been tested (even as optional login)
                - Request is in review by Gen3 Head of Security
            - Multiple identity providers --
                - RAS vs. FENCE considerations; authorization impact
        - (Phase II Hosting New Data) Feature D: Streamlining and multi-threading ingestion of new data sets
            - Versioning protocol needed (releasing updates) -- likely R2 priority
            - Versioning/provenance intersection → on DRMWG's list of priorities
                - Auth for different versions; larger amounts of time needed to dedicate to versioning → user workflow considerations (e.g working on one dataset when another become available)
        - (Interoperable System v2) Feature D: bulk FHIR ingestion
            - Discuss on future TCM -- potential R2
        - [Ca] Gen3 to add a Study Viewer page to Windmill
            - Blocked for R1 (internal Gen3 block) → Alessando to note what the blocker is
        - Scaling for DICOM images
            - TCM as a forum for discussion/resolution *OR Imaging TT*
    - Penetration testing across the consortium by NHLBI IT Team
- Teresa's Notes
    - ....

Fellows Breakout: Improving new user onboarding

Moderator: Chris E

Notetaker: Paul
Attendees: Randi Johnson, Bo Li, Beth Sheets, Brandon Le, Caitlin McHugh, Diego Mazzotti, Fayuan Wen, Jean Monlong, Jennifer Brody, Jinling Liu, Kenny Westerman, Lauren Hochman, Mark Craven, Melissa Cline, Pietro Nardelli, Ravi Mathur, Sarah Gerard, Stephanie Gogarten, Stephanie Suber, Chris Erdmann, Paul Kerr, William Disman, Zilin Li, Laura Raffield, Dandi Qiao, Einat Granot-Hershkovitz, Michelle Daya, Harrison Brand, Kristin Wuichet, Ashok Krishnamurthy, Jamie Murkey, Xu Zhang, Xuefang Zhao, Yonghua Zhuang, Matthew Satusky, Tiffany Miller, Dave Roberson, Yaling Tang

Notes:

Share your questions, thoughts, ideas, examples… towards improving the following:

Bo: Are there plans to unify cloud solutions, workflows? Delay in running analysis?
Stephanie: explanation of history of why two platforms (Terra vs Seven Bridges)
Chris: This is complex and there are efforts to improve workflows and interoperability, right now have to communicate directly to get accounts connected, have credits used on both platforms
Bo: Some issues with this assignment using STRIDES etc, could be a place for improvements, Bo willing to give feedback etc on this
Tiffany Miller: Also interested in billing interoperability. Can STRIDEs provide billing accounts ahead of the cohorts joining/starting? If we can take care of that earlier, that would be better.

Ravi: Appreciated user stories at the poster session.

Chris: User pathways may be one way around this- get narrative of how people got started on the platform- could also be used as part of the cohort mentorship side- may want to have story idea for non-fellows new members too
Randi: Intro demo for an hour and one hour on creating tools- after these could have been a great time to connect with a fellow, make videos from prior fellows to help build a narrative of stories of what can be done on the platform- more valuable after the basic and tool building training

Question from Beth: Do you want more of a broad overview (like in poster session) or a direct demo? Randi: Demo seemed helpful
Could do these demos and recordings as part of monthly fellows meeting
Kenny- suggestion of pairing of mentor (or from Chris- Fellows alumni ambassador program) based on interests of new and former fellows- if interests close enough more mutually beneficial, from both a networking and an actually being helpful for new fellow perspective
Videos may be quite helpful from a scalability perspective.

Tiffany- Timing of onboarding could be useful to identify gaps etc

Question from Xu: Is it possible for fellows to have a more extended dbGaP data access? So far we have to renew the data access status on Terra every month. I just have my dbGaP access expired this month and it will take quite a while to review. It is just a complicated process.

Fellows Onboarding Plan (concern brought up yesterday was with timing. When was timing an issue? What about data access?)
From Tiffany- consider turn around time in onboarding- what are the gaps?
From Caitlin- Different fellows will have different timelines due to differing needs for lit review, method development, etc, etc. - makes it important that trainings are something you can go back to, watch again when you actually start making an application, etc
From Lauren- content is available but finding what is needed can be harder, what is constantly needed and makes sense to record and document vs what won't be needed very often- 3 hour videos may never be watched
From Bo: Could videos be chopped into very short pieces? Then more watchable/useable.
From Lauren- General orientation is probably doable to record and watch all at once, but other pieces need to consider how often something is needed
Dave- maybe go back internally and see how often certain things have been requested and ask about making videos.
Lauren Hochman-For videos - emphasis on more useful and easily found, not just more quantity.

Chris- Sounds like even a pretty minimal number of videos would be quite helpful, even if not super polished- Caitlin mentioned Stephanie's Bring your own tools document as a good example of something that gets used a lot that is fellow created
Increased use of the forum could be a help, good place to post documentation and look up questions with low start up cost
Stephanie- we need Stack Overflow for BioData Catalyst- is this what the forum should be? Seems intimidating to submit "help ticket" sometimes
It would be great if some of these fellow questions were addressed in a publicly searchable way (they are already being asked, just not recorded in a way other fellows can use and find)
From Yaling- not clear always who to turn to for help, so Stephanie's idea might help
Dandi: +1 I think maybe people will start using the forum if we direct questions to the support desk to the forum
Caitlin- stack overflow would also give some good street cred to the ecosystem
From Chris- Fellows have access to Slack etc, so need more scalable Forum and Help Desk solution that will be available to other new users who aren't fellows
Forum may need to be more easily findable, many people don't know it's there, how to access
From Yaling- Tried BDC forums, etc but found one on one conversations on Slack is easiest and most efficient for fellow- from Ravi- this makes sense, but Slack does lead to redundancy- Dandi agrees that questions need to be more public
From Chris: Terra has more direct linking to Forum vs Seven Bridges- community forum is right above contact us button
From Melissa: It can be confusing if you don't know where to look first but better than not having forums in the first place

Chris: Documentation search improvements are still coming, and may be helpful for some of these relevant questions
From Dandi:
I think we need to have one centered forum, and the other thing is to notify people of questions that have been posted
I don't usually go on the forum to see questions posted, but I get notification of slack messages

Orientation
From Yaling- it was complicated to understand difference in Exchange areas vs public dbGaP applications
Data access is a major source of complications and confusion
Stephanie: many of initial fellows were already part of TOPMed, which can lead to confusion as non-TOPMed affiliated fellows are brought into the project
Chris: Made some updates to the FAQs to try to address this, but this may be a question to elevate earlier, as it is causing confusion
Stephanie- May want to link to public TOPMed page on "how to get data access"
Lauren- want to be responsible data stewards but also want to link people to the data they need where we can, may need to be more proactive in push messaging for
Ravi- partnerships- TOPMed is one of first and largest, but many more evolving, is exchange area concept really relevant there
Chris- Templated approach may be more relevant for these new partnerships
There are improvements already planned to the data access page, but do maybe need to get more info/links out on TOPMed data access in particular as part of the orientation process
From Lauren Hochman:
Did a little hunting to see where we have language about needing data access approvals and found it in a couple of places. Of particular interest might be its inclusion (though not detailed) in the overview document under Ecosystem Access, Hosted Data, and System Services (How does data access work?
The BioData Catalyst ecosystem manages access to the hosted controlled data using data access approvals from the NIH Database of Genotypes and Phenotypes (dbGaP). https://bdcatalyst.gitbook.io/biodata-catalyst-documentation/.

Yonghua- Docker images how to deal with it getting "erased" every day, how to control this without running up large costs- Dave will get in touch
Office Hours - every other month
Chris: If we don't have enough time per office hours, the general thinking is that we would offer it every other month where users can register ahead of time, ask questions ahead of time in a **collaborative doc**, and then we would have platform liaisons there to answer questions.
Tiffany: sometimes things can be handled more quickly with quick back and forth
May be more scalable if meeting one on one for more complex issues as opposed to things that could be dealt with quickly pre office hours
Workshops

Additional:

(Cohort) Mentorship
    Documenting
Help Desk/User Support
    One-on-Ones
    Forum
    FAQs
Interest Groups
Tutorials/Documentation- Preference for videos or screenshots along with text discussed yesterday
    Discoverability
    Data Availability/Access
    Cloud Credits
    User Pathways
    Videos/Webinars
    Workflows
Communication Channels (e.g., Newsletter, Social Media, Blog)
Publishing Research/Reproducibility - From Chris: Dockstore is a place where hoping to do this From Beth: Bring your own tools documentation is useful for orientation for workflow languages in Docker, how do you publish, can link to Github, publish a version with a DOI, can link Orcid ID, etc.
Alumni/Ambassador Programs - Buddy system
Anything else?

Breakout Report Back

PI Breakout Report Back

PI Breakout summary slide

- Ben Heavner to Everyone: note: [context of developing smoother data access] study registration is a component of this discussion, too.
- Kevin Osborn: Note that Dockstore has a way to mint a DOI

Dev Breakout Report Back

Dev Breakout summary slide

Fellows Breakout Report Back

Fellows Breakout summary slide

- Lauren Hochman: For videos - emphasis on more useful and easily found, not just more quantity.
    - +1 from Brandi

Consortium Project Management

- BioData Catalyst Project Management V2 [Draft]

PM Plan is living document. Will incorporate UECG over next few weeks. Post-award re-baselining complete. Multi-team features now have RPs. NHLBI is monitoring progress in Jira. Feature Planning documents are linked in Jira.

Jon Kaltman: It is very important to NHLBI that we improve the rigor of our project management and strict use of JIRA is the first step toward this goal.

Brandi: Feature planning is valuable in aligning everyone. Is there a venue where people could synthesize the feature planning documents into a couple of slides to present them to a broader audience? How do we ensure we're giving sufficient visibility into the whole process? A dynamic design discussion would be valuable. The eventual goal is not only alignment with teams and NHLBI, but ensuring we're truly meeting user needs. Are there ways to obtain that validation while things are in the design phase?

Chip: Will think about a means of doing that and perhaps rethink how we populate the planning documents.

Rebecca Boyles: @Brandi- If you, or anyone sees something that needs more User Acceptance work, please highlight those

- Brandi: Yea, at a high level we want to be able to validate that the problem to be solved as understood is truly the problem contemplated, and then, that proposed implementation is expected to solve that problem

Brandi: UAT [user acceptance testing] happens too late, at least in my mind to catch solving the wrong problem

Becky: We need to capture that as an action item: cross pollination of Tools & Apps with FHIR

System Health Monitor is password protected to track usage. Dashboard will be visible to users but not the public. Check w/ Alastair. Does he want curated/certain information to appear to the public (from the more advanced grafana based system or does he want the dashboard to be public?)

Brandi to Everyone: Basically, I think we're moving to a more formalized consortia-wide Software Development Lifecycle and the next maturation step might be a formal sign off (in jira) at key design / implementation steps. It might be too heavy now but worth thinking how/if we build to that

Stephanie Gogarten: Can there be a connection between the data submission process for these cloud buckets and dbGaP submission? These are currently totally separate for everything but CRAM files, and it would be really great to integrate them.

Sweta Ladwa: we should be careful in stating "studies" vs. datasets; one study could have multiple datasets

- Ben Heavner: +1 @sweta - there's also projects/programs…. TOPMed is a big project, with data from many dbGaP studies, including many datasets…
- Yaling Tang: +1. From fellows side, I have already run into problems due to the updates happened recently.
- Josh Bis: I have come to realize that it is out of scope for BDC, but simplifying the aspects that Ben mentioned about the complexities of TOPMed as a collection of studies/projects/datasets with various permission structures seems like it will remain a long-term challenge
- Ben Heavner: And I think that seeking to share/collaborate on "cross-study" data sets will add complexity we have to confront.
- Stanley Ahalt: Yes, this complexity is a key issue!
- Josh Bis: yes, that's incredibly complicated (and my sense is that ability access, collaborate, and analyze cross-study data is an NHLBI goal, particularly for less computationally-sophisticated users. So it's a real conundrum!
- Jon Kaltman: I think part of the solution is teasing about what are required Policies and what are consortium policies that can "re-negotiated."
- Ben Heavner: This is a place where some TOPMed DCC staff have worked closely with Sue to develop current policies, especially @Quenna Wong and @Sarah Nelson
- Alisa Manning: To Jon's point, we can point to a couple examples in TOPMed, where a lot of work went into finding a solution to solve some of these problems:
- Ben Heavner: There's also @Jen Brody's work with the analysis commons.
- Alisa Manning: The TOPMed Imputation Server; BRAVO — two great examples from UMich / The Analysis Commons tackled the collaborative analysis aspect
- Josh Bis: The key to the Analysis Commons and collaborative cross-institution analyses was a Consortium Agreement between participating studies. And a lot of hands-on work to monitor permissions and trust between participating investigators.
- Alexander Bick: Re: Alisa, Josh, Jon - If we compare BDC & TOPMed to what else is in the 'marketplace' eg UK Biobank (and soon All of Us) - it will be important to think creatively about data access if we want people to continue to use the environment (and NHLBI data more generally). Currently researchers are willing to jump through a lot of hoops to get data access - but in the future if one application gives access to 500,000 genomes from UK Biobank and another application gives access to 1M genomes from All of Us, there will be less willingness on the part of users to jump through access hoops

- Stephanie Gogarten: The difference between TOPMed and those other programs is that they were designed with unified consent. TOPMed is a collection of pre-existing studies with a huge variety of different consents.
- Rebecca Boyles: Great reminder Stephanie. There are many studies like TOPMed and making that data more valuable is an important goal
- Josh Bis: for sure — solving the complexities of TOPMed is extremely challenging, but those lessons should apply broadly and will make bringing on new studies and datasets seem easy. Does BDC get us into a higher priority group?

Alisa Manning: Food for thought regarding Alessandro's feature updates: what can we do to decrease the burden on data ingest into BDCat. Potential idea: For data that would not be broadly shared within BDCat [more focused user communities] — could we enable a data onboarding path that looks more like "bring your own data" and use a researcher-driven data repo model?

External Expert Panel Retrospective

- David - Informative broad ranging discussion;
    - Gaps:
        - System access as identified by the Fellows; difficulty gaining and understanding navigating the system → separate engineering problems vs. academic/research problems; the latter cannot be solved in 3-month increments
        - WGs covering the right topics → consider narrowing the focus; prioritizing specific goals
        - Focus on executions -- moving from theoretical to practical
        - Governance model for APIs; introp important
        - Core tool sets delivered by BDCatalyst (focus on naive users) → allow for integration by more sophisticated developers
- Donna - David covered/summarized most of her points
- Mark - Echoed David's points
    - Project success to date of standing up a system with good functionality, lingering issues around learning how to use the system; emphasis on usability and growing the user-base
- Jason
    - Thanked the Fellows and encouraged them to continue providing feedback; constructive feedback is critical to further development and program success
    - Focus on user experience -- think about long term strategy of BDCatalyst (reproducibility); scalability of user growth → ask the question, ***what do we think we need to do in the future to achieve the usability goals*** -- communicate with BDC3, NHLBI, etc.

NHLBI Final Thoughts

- Alastair - Encouraged hearing the Fellows -- extremely successful component of the program;
    - Challenges will result in needed changes; the world has changed and we need to respond
- Jon K - Appreciative of everyone's effort and dedication to the program;
    - Quarterly meetings can be intimidating given the challenges surfaced -- need to prioritize what to engage/focus on without letting too much slide
    - Need to be flexible and responsive to external changes;
    - We want the Fellows to come back in March '21 and not have the same concerns they had today (Dec '20) → this would be an accomplishment
- Chip - Appreciative of development teams and user community; need to re-shift focus on users and their needs; challenge the development community to think how we can prioritize the user community

Recap and Next Steps

Day 2 Zoom Chat Link