DATA SHARING

What types of data is RDCA-DAP including?

RDCA-DAP will integrate data from existing rare disease data collections including registries, clinical trials, observational studies, and natural history studies. We are currently pursuing partnerships with many groups around the world collecting rare disease data. We are interested in any type of clinical or patient report data, such as genomic data, imaging data, safety biomarkers etc. However, RDCA-DAP does not collect new prospective data or store biospecimens.

Can any organization with data to share participate in RDCA-DAP, even if they are not currently part of IAMRARE?

RDCA-DAP is happy to work with any registry system or data collector. If you are interested in contributing data, please contact rdcadap@c-path.org so that we may arrange a follow-up conversation. RDCA-DAP can integrate data from all rare diseases from any source. IAMRARE(TM) databases are one source of data that can be shared conveniently under an existing agreement between NORD and C-Path, but we are happy to accept data from any platform.

Can an organization contribute data from an ongoing study, where prospective data is still being collected?

Data can be contributed throughout a study. For ongoing studies, RDCA-DAP will set up a system by which data can be transmitted at regular intervals (e.g. every two years) to keep the database up to date.

What is the best way for a drug development company or academic with interest in a disease area to initiate a new project or join an existing one?

Please contact rdcadap@c-path.org and we will guide you through the process. The RDCA-DAP team is always willing to discuss how companies or other researchers can engage with the initiative, what data you may be willing to contribute and what your data and analytics needs are.

What are the biggest barriers to data sharing and how can patient or patient groups help to make data more accessible?

Data is not currently shared as widely as would be ideal to accelerate research, as there remain many small data silos. Most concerns of the community can be resolved by RDCA-DAP's data sharing processes, however. Researchers express a number of concerns about data sharing, including lack of interest in sharing, not prioritizing data sharing as it may take time/resources away from other things, concerns about whether individuals in their studies are willing to share data and competitive concerns with other researchers. In some cases, data custodians are reluctant to share as they rely on the value and collaborative partnerships that existing data can bring to the sustainability of the data collection and to support future data collection efforts. These concerns can be addressed in the terms of a data contribution agreement with an individual organization, and the original contributor retains ownership of the data in all cases. In other cases, data cannot be legally shared as the study design and protocol did not include consent that allowed for the custodian to use the data for purposes beyond those anticipated at the time. In this case, a process of getting in touch with everyone in the study to ask if they are willing to share data is very time consuming, and few researchers are willing to

undertake the effort. In general, patients are very interested in sharing data so long as the data is shared in a safe way that protects their privacy. Patients' championing of data sharing and consent for sharing prior to giving data can be extremely effective at improving the degree to which data is shared.

Who owns the data contributed to RDCA-DAP?

The original custodian of the data (the person or institution that shares the data with RDCA-DAP) retains ownership. They will dictate how RDCA-DAP may use the data through a legal agreement signed with the platform called a data contribution agreement.

Which IRBs are used? A central IRB, or can I access data if my local IRB has approved my study?

RDCA-DAP is not collecting new data, just integrating existing data in a de-identified format. Therefore RDCA-DAP itself does not require IRB approval. Individual studies that share data with RDCA-DAP should have been approved by an IRB or equivalent ethics committee, but we do not stipulate which one.

Can we see the data use agreement?

If you want to contribute, or access data, or partner with the initiative in other ways, please send an email to rdcadap@c-path.org and we will set up a call to discuss the next steps as well as share the data contribution agreement, data governance structure and terms and conditions for use of the data**.**

Does the rare disease have to be approved by the RDCA-DAP before it will take any data for that disease? Will RDCA-DAP take data from a study in which it is the only data set in the RDCA-DAP for that disease?

RDCA-DAP will accept data for any rare disease. The platform may actively seek data in specific disease areas identified as areas of need for analytics, but it will accept, integrate and make available any data that is shared with the platform. If only one dataset is available, and the custodian is willing to share it, then that data will be available as the only study for that disease. Over time our goal is to have multiple datasets for each disease. In the meantime, single datasets can be used for certain analyses and cross analyses with datasets in related diseases may be possible. If the custodian does not want their dataset to be identifiable, RDCA-DAP will not share the data externally until another dataset in the same disease area has been brought in.

BEST PRACTICES FOR DATA

Should new patient organizations without a natural history study engage early with RDCA- DAP? Who helps foundations design natural history studies and comply with best practices and data standards?

We encourage you to talk to experts in natural history data collection, including RDCA-DAP, prior to starting a new collection of data to ensure that you are asking questions that are standardized, structured and match other registries and data collections in your disease area. RDCA-DAP encourages use of common data elements (e.g. defined by NINDS at https://www.commondataelements.ninds.nih.gov/) if they exist, and standard data structures and ontologies (e.g. CDISC, OMOP) as appropriate. We recommend that you contact NORD's IAMRARE team (see https://rarediseases.org/iamrare-registry-program/) if you are interested in establishing a patient registry.

How will RDCA-DAP interact with stakeholders engaged in ongoing data collection to advise on best practices?

We are establishing a feedback loop to engage with data contributors to help them optimize data collection over time. As we learn more about common challenges across data collection efforts, we will share best practices with the community with respect to study design, data elements, data structures etc. For now, we recommend

looking at CDISC data structures and nomenclature, common data elements if they exist for your rare disease and working with existing data collection efforts for best practices and common protocol guidance. Using validated patient reported outcome or quality of life instruments is advisable where they exist. As RDCA-DAP reviews data coming into the platform, we plan to share learnings from the data with the data custodians and broadly with the rare disease community.

Which standards and ontologies are you focused on or does that depend on the type of data you have (e.g. CDISC, human phenotype ontology, disease ontology, UniProt for proteins, etc.)?

Different ontologies will be needed depending on the type of data. The RDCA-DAP team has broad experience with CDISC, but this structure is most appropriate for clinical data. RDCA-DAP is very interested in exploring new ontologies for new data types and have entered into an agreement with the European Joint Programme for Rare Diseases to learn more about the rare disease ontologies that have been developed in Europe, as well as exploring others (OMIM and Monarch, for example). You can learn more about the project team's current thinking about standards and ontologies in these video resources: Standards and Ontologies and Standards and Integration.

Is the data in RDCA-DAP de-identified? Can it be re-identified for future more in-depth analysis?

RDCA-DAP requires that all data is de-identified prior to sharing with the project to protect patient privacy. Our terms and conditions require that the user does not attempt to re-identify individuals in the database. In very small patient populations, however, it is difficult to fully de-identify the data, but by integrating worldwide data without standard identifiers (name, address, date of birth, etc.) we reduce the likelihood of identification. Any information in a dataset that could be identifiable will not be shared. If you need identifiable data, you will need to work with the person or group that collected the data originally.

How are you connecting or linking one patient's records across multiple data sources? Do you use a Global Unique Identifier (GUID)?

RDCA-DAP recognizes the likelihood that individual patients may be represented in multiple datasets that are integrated into the database. In anonymized datasets, if the contributed data have not adopted global identifiers, it is not possible to recognize potential duplicate patients across datasets. However, the potential bias can be adequately controlled by standard statistical and mathematical methods. Other quantitative drug development tools built by C-Path using databases with this same issue have accounted for this bias within the development of the models and tools and this has not affected the utility of the tools developed. Moving forward, RDCA-DAP strongly supports the use of global unique identifiers (GUIDS) that can be used in all studies to allow us to link records from individual patients without identifying the person in question.

Can you please clarify what you mean by "data harmonization"?

When we refer to data harmonization, we are talking about the process by which we look at different datasets from different investigators and locations and make them interoperable. That is, we make sure that unrelated elements with the same name are the same thing (e.g. that two studies using a 6-minute walk distance were, in fact, performing the test the same way), units are equivalent in all datasets (e.g. meters vs. feet), measurements that are performed the same way can be identified (e.g. measuring a protein level using one technique is identifiable from a measurement of the same protein using a different method), elements that may have several names are unified under a common name (e.g. "age" and "time since birth" are the same) etc. This does not change the meaning of the data but allows the datasets to be aggregated and compared.

What is the difference between a "data lake" and a "data warehouse"?

Data may be stored in either a data lake or a data warehouse and may move between the two. A data lake is a term to describe a relatively unstructured set of data, where data points may be accessed through search terms or common metadata tags (labels). A data warehouse is a more structured dataset that has been harmonized to a greater extent and is analysis ready. RDCA-DAP anticipates having data in both formats, depending on the structure of data as it is shared with us. All data will be searchable and useful for analysis, and unstructured data may become more structured as it is used, and therefore may move between the lake and the warehouse.

DATA ACCESS

Will this normalized data be made available for federations with other data sets that are not normalized into RDCA-DAP?

We are looking into gaining federated data access to outside datasets so that they may also be used within the platform in concert with data in RDCA-DAP. In addition, if the contributor of the data to RDCA-DAP gives permission for external access to the data then RDCA-DAP data will be able to be accessed through other federated data platforms. This is managed on a dataset-by-dataset level adhering to the requests of the data owner, as written in individual data sharing agreements. As RDCA-DAP promotes sharing of data, we hope that data will be able to be shared and accessed in this manner.

Are natural history studies that are submitted to RDCA-DAP preserved as natural history studies?

Each study in the database may be analyzed as an individual dataset, but data from that study will also be available in aggregate/combined with data from other studies. The study type will be tagged with the type of source data (e.g. clinical trial, patient registry, clinician entered natural history).

Is there a way RDCA-DAP can capture what the data is being used for and what decisions it is driving for data users?

The request for data access form asks for a description of what research is proposed and why the user wants to use the data, in addition to what data is being requested. This means that RDCA-DAP and the review committee know how the data will be used, if it is an achievable project based on the data in the database, and if others have already done the same analysis or are planning to.

I would suggest that the "A" in the DAP (Analytics) is not getting enough attention. The statisticians and analysts of the program are really providing invaluable assistance in analyzing the data being shared so as to improve clinical outcome measures on our clinical trials.

The statistical tools that will be available in the platform are still under development. This supplementary video on the planned platform (link: RDCA-DAP Platform Preview) provides some insight as to what this will look like. We are open to discussion with potential users as to what specific analysis tools are needed. If you have ideas and recommendations, please contact rdcadap@c-path.org. More advanced tools may be able to be developed utilizing RDCA-DAP data (e.g. those presented in the meeting and in additional supplemental videos). These may be developed either independently by users of the platform or in collaboration with C-Path's internal Quantitative Medicine team. We will continue to provide updates on the analytics aspect as we continue to develop them.

What are the benefits of the platform analytics for participating groups?

There are many benefits for using the analytics platform – here are a few:

- Access to standardized data in one place;
- An easy-to-use interface to determine if the platform has the data you need and do simple statistical tests.
- Ability to download some data in CDISC format, appropriate for regulatory filing (dependent on the type of data in the database);
- Platform tools to help with simple statistical analysis and data exploration;
- Access to a platform with advanced pharmacometrics tools to develop more complex analyses, where you can add your own software as needed;
- Potential collaboration with the statisticians and pharmacometricians within C-Path's Quantitative Medicine team who have experience with these types of data and solutions for medical product development;
- Potential collaboration with C-Path's regulatory science experts who have experience developing regulatory-grade drug development tools such as outcome measures, biomarkers, natural history and disease progression models.

Is there a long-term funding plan for this project so that the information stored in this platform is available for many, many years to come?

RDCA-DAP is funded through an FDA grant to the Critical Path Institute (C-Path; cooperative grant agreement U18 FD 005320) and developed through a collaboration between C-Path and the National Organization for Rare Disorders® (NORD). While the current funding remains in place, RDCA-DAP will not charge for use of data. If federal funding were to stop, the partners would explore other grants to keep the platform accessible, and/or consider charging for-profit companies a fee to access data. Our goal would be to ensure continued access to the data.

What data does RDCA-DAP currently have available?

RDCA-DAP currently has 20 datasets representing 10 rare disease states and is actively working to bring in additional data. Not all data can be shared with the community at a patient level, per the preferences of the data custodians, but much of this data will be available when the platform opens in 2021. The current disease areas where we have data include: congenital hyperinsulinism, Duchenne muscular dystrophy, Friedreich's ataxia, GNE myopathy, phenylketonuria, polycystic kidney disease, pemphigus and pemphigoid, Prader-Willi syndrome and tuberous sclerosis.

GENERAL

How does a patient group seek FDA support to develop clinical outcome assessments (COAs) for their disease state?

FDA periodically posts requests for applications for grants. Details of the program referenced during the meeting can be found at: https://www.fda.gov/drugs/development-approval-process-drugs/cder-pilot-grant-program-standard-core-clinical-outcome-assessments-coas-and-their-related-endpoints. The standard core clinical outcome assessments program is outside the scope of the RDCA-DAP.

Is the "Rare Disease Patient Registries Unlock Cures" infographic available for outside organizations to share to encourage patient enrollment into registries?

Yes, please contact NORD at research-programs@rarediseases.org to request permission to use this graphic. It is also on the NORD website at Unlock Cures.

Since so many rare disorders affect children, does RDCA-DAP overlap with I-ACT (Institute for Advanced Clinical Trials for Children)?

I-ACT for Children is an independent non-profit organization that works specifically on pediatric disorders, aiming to accelerate clinical trials for children and make them more efficient. I-ACT for Children was created out of the Pediatric Trials Consortium, a C-Path Consortium, and several current C-Path Consortia collaborate with I-ACT for Children on specific projects.

Is there demand for tools by which patients could receive monetary value for the data they share?

Data shared with RDCA-DAP needs to be de-identified prior to sharing with the platform. This means that the original patient cannot be identified, so there would not be any way to identify and pay the patient for use of their data. RDCA-DAP also does not charge for use of the data by outside entities.