RDCRN Guidance for the Development of Data Sharing Policies by Individual Consortia – v1.2 23APR2024

Data sharing is a key aspect of the RDCRN framework and a priority for many stakeholders involved with the network. Some consortia groups within the network have advanced data sharing practices in place, while other consortia are in the early stages of developing data sharing policies and plans. The goal of this document is to offer non-binding guidance to consortia on general principles for data sharing and key factors to consider when developing data sharing policies.

The NIH has established a Final NIH Policy for Data Management and Sharing, effective January 25, 2023: https://www.federalregister.gov/documents/2020/10/30/2020-23674/final-nih-policy-for-data-management-and-sharing-and-supplemental-information. This policy should serve as the basis for individual consortium data sharing policies, along with other NIH institute-specific policies as applicable. A consortium's cooperative agreement with the sponsoring NIH institute(s) should outline the specific data and data type to be shared. After the cooperative agreement end-date, consortia that "graduate" from the RDCRN may continue to have their own data sharing policies independent from the RDCRN. Data sharing performed directly by the consortia is encouraged, regardless of funding status.

Guidance from NIH/NCATS:

The NCATS RDCRN Data Repository (RDCRN-DR) is an NCATS funded data sharing resource containing clinical research data from individuals with rare diseases who are enrolled in RDCRN-sponsored protocols. Data types in the RDCRN-DR will reside on an NCATS Federal Government server and will be harmonized to published data standards where feasible to facilitate meta-analyses and the merging with additional data sets. The RDCRN-DR is a highly interoperable, secure, clinical data research environment that will harmonize clinical and patient data. RDCRN-DR data use and transfer agreements developed by NCATS will contain terms and conditions consistent with NIH data sharing policies and US Federal Government statues and laws. In those instances where RDCRN consortia specific data sharing policies and practices may conflict with NCATS and NIH data use and transfer agreements terms and conditions, the NCATS agreements shall prevail.

Informed Consent:

The foundation of data sharing is the informed consent process and researchers should build language into the consent to allow for data sharing, future research studies, and a potentially broad audience for the data. See Appendix 1 for additional guidance.

- Informed consent (IC) language should allow for the sharing of patient data, ideally in the form of what the Health Insurance Portability and Affordability Act (HIPAA) names a limited data set. Limited datasets are coded, with direct personal identifiers removed, and may contain dates; they are particularly useful for downstream analysis. Alternatively, de-identified data sets can be shared. Keep in mind that it may be easier to identify participants with rare diseases, special care should be taken to ensure that the risk of re-identification is minimized. The IC language

should also outline sharing of data with the NIH, RDCRN DMCC, and other researchers, in accordance with an approved IRB protocol, institutional policies, and any applicable laws.

- **Aggregate Data Set**: Any summary of individual level data, including regression output, simple univariate frequencies and cross-tabulations, and means or other summary statistics presented on grouped data (https://research.kpchr.org).

- **Coded Data Set**: Identifying information (such as name or social security number) that would enable the investigator to readily ascertain the identity of the individual to whom the private information pertain has been replaced with a number, letter, symbol, or combination thereof (i.e., the code); and a key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens. (https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html)

- **Anonymized Data Set**: Data recorded in such a way that subjects cannot be identified or re-identified (https://en.wikipedia.org/wiki/Data_anonymization)

- **De-identified Data Set**: According to HIPAA, a data set can be designated as de-identified based on one of two methods: First, an expert may determine that the dataset has a very small probability of leading to identifying an individual ("expert determination method"). Alternatively ("safe harbor method") the data set must be stripped of 18 direct identifiers (name, social security number, etc.), and contain no elements of a date except for the year, and no territorial aggregation below the state, except for the first three digits of a ZIP code identifying an area with 20,000 or more inhabitants. (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard)

- **Limited Data Set**: As specified in the HIPAA Privacy Rule, limited data sets differ from de-identified data sets as they may contain dates and city/state/zip code location information, but sharing of the information must be governed by a data use agreement (https://www.hopkinsmedicine.org/institutional_review_board/hipaa_research/limited_data_set.html)

- Informed consent language should also describe how the research team will protect the privacy, rights, and confidentiality of human research participants, which organizations or institutions will store and share the data, who will have access to the data, and how the data will be used now and in the future.
- Consortia should consider the adoption of Global Unique Identifiers (GUIDs), the process of which needs to be outlined in the informed consent. The use of GUIDs is recommended to facilitate linkage of data pertaining to the same individuals without using direct identifiers. The RDCRN DMCC recommends using the NINDS Centralized GUID solution, part of the Biomedical Research Informatics Computing System (BRICS) platform, but it is understood that some studies must use other GUID generators. In such instances, the DMCC recommends that, if at all possible, the study team generate the NINDS Centralized GUID in addition to the GUID they are

obligated to generate, so that a common GUID will be available for as many RDCRN protocols as is possible.

- For patient data covered by the European Union General Data Protection Regulation (EU GDPR), special considerations for language and privacy notices may apply. A sponsoring NIH institute may have additional requirements to demonstrate compliance with GDPR. Additional resources can be found on the Health and Human Services website: https://www.hhs.gov/ohrp/international/gdpr/compilation-of-gdpr-guidances-tables/index.html

Consortium-Specific Data Sharing Policies:

In establishing their own data sharing policies, consortia should consider guidance in the "Final NIH Policy for Data Management and Sharing" and be consistent with FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles and the data sharing policies of their sponsoring NIH institute.

- Read more about **FAIR** data principles: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/

- A data sharing policy should establish how data will be shared among the consortium sites and with the Administrative Core of the consortium, as well as with parties external to the consortium. If there are existing legacy data, the consortium should specify any differences in policy for new versus previously collected data. In writing the policy, the consortium should consider developing the following components:
    - A data stewardship statement, recognizing who is authorized to make final decisions about the sharing of data, and making reference to any subcontracts or other source documents establishing data stewardship.
    - A publication agreement and authorship policy defining the responsibility to publish, citation of funding, acknowledgement of consortium researchers, and priority of authorship.
    - A management solution for resolving potential conflicts of interest.
    - The recognition of different types of users requesting access to the data (ie. users internal or external to the consortium) and considerations for a particular category of user, particularly regarding the need for a Data Use/Transfer agreement.
    - An intellectual property (IP) statement outlining who makes decisions for and has the rights to intellectual property generated by the consortium.
    - A data access request procedure and related application form outlining how users will obtain access to the data.
        - The procedure should specify any considerations for different types of data releases, for example, coded limited data sets, coded de-identified data sets, anonymized data sets, and aggregate data sets.

        - Plans for how to document released data should be outlined; such documentation should include descriptions of who the data were shared with and any necessary authorizations that were completed.
        - The procedure should describe how the data can be used upon release.
        - A process for sharing back the resultant data of ancillary studies with the consortium and whether/how to incorporate those results into the database may be outlined.
        - The procedure should define any financial obligations for data sharing and data use (ie.downloading costs from a cloud environment, data services costs to prepare/aggregate the data, etc.). In particular, it is important to clarify if there are any restrictions or limitations for data release to third parties (particularly in light of the use of federal funds to support the research), if sponsor approval is needed to release the data, and if administrative fees may be applied.
    - A plan including any necessary actions to comply with EU GDPR, if the consortium gathers data from participants covered by the EU GDPR.
    - A continuity plan that details a continued strategy for sharing data with relevant stakeholders after the cooperative agreement end date.

RDCRN Data Repository:

Starting in RDCRN cycle four, under RFA-TR-18-020, RDCRN consortia are required to share data with the RDCRN DMCC for the purpose of establishing a federal data repository. NCATS is building the RDCRN Data Repository as a data sharing resource containing clinical research data from individuals who are enrolled in RDCRN research. Consortia may have additional data sharing obligations designated by their specific NIH ICO; this guidance does not interfere with NIH ICO requirements. In such circumstances, data availability will be coordinated between the RDCRN Data Repository and ICO-specific repository. Consortium internal policies for data sharing should be consistent with the goal of releasing data to a federal repository.

- The consortium cooperative agreement will guide the release of data to the RDCRN data repository; the extent and timeliness of data transfer should be incorporated into the work plan and milestones of each consortium and negotiated with NIH Program Officials.
- Once the data is deposited by the consortia into the RDCRN Data Repository, governance of data sharing is the responsibility of the NIH. The RDCRN DMCC will manage the Data Repository under NIH guidance. The NIH will solicit and consider input from all stakeholders in designing policies for the RDCRN Data Repository.
- The administrative core of a consortium should have the authority to manage data on behalf of the sites and for the consortium as a whole and to negotiate and execute a data sharing agreement with the RDCRN DMCC, allowing the DMCC to receive and manage the data, for the purpose of populating the RDCRN Data Repository. This may be accomplished through data sharing and ownership language in a Materials Transfer Agreement, Subcontract Award, Data Use Agreement, etc.

Data Standards:

Data collected under the Rare Diseases Clinical Research Consortia U54 mechanism should adhere to data standards developed by the RDCRN Data Standards Committee and approved by the RDCRN Steering Committee to ensure best data management practices and to support reuse of the data by approved researchers gaining access to the data through the RDCRN Data Repository. Standards should be incorporated with the design of the study and associated database(s).

- To the extent feasible, the RDCRN data standards should be applied to the data directly shared by the consortium according to the Consortium-Specific Data Sharing Policies.
- The RCRN DMCC will provide guidance on the preparation of datasets that each consortium will transfer to the RDCRN Data Repository.
- Quality control of datasets will be a joint effort between the research team, consortium leadership, and the RDCRN DMCC. For datasets deposited into the RDCRN Data Repository, quality control will be incorporated into the submission and approval process.

APPENDIX 1 – Example of Recommended Consent Language

The informed consent language outlined here is recommended but NOT required; users are welcome to modify the text to fit their specific needs. An alternative abbreviated paragraph is also included. We advise the research team to seek input from patients when designing consent/assent language.

Special Considerations:

- The NIH advocates for broad sharing of federally funded research data. Investigators should consider including consent language that permits this for all subjects and does not allow opting out of data sharing. However, the type of study and recommendations of the responsible IRB may warrant adding an option for the participant to opt in/out of future data sharing, especially if there is significant potential direct benefit to the participants through the study.
- Certain data types, such as genomic and imaging data, may require special consent language. For example, some genomic studies may need to include language that describes whether genetic results will be returned to participants. At a minimum, consent language should state what data types will be shared (e.g., genomic, imaging, etc.) and for what purposes (e.g., General Research Use).
- If a study collects data for participants covered under the European Union General Data Protection Regulation (EU GDPR), certain considerations, language (such as the right of revocation), and privacy notices may apply. We recommend that you consult with your institution's legal authority to ensure compliance with EU GDPR.
- Research teams should consider whether the data will be shared in the future as a coded data set, which can be linked back to identifiable information, or a fully anonymized data set, which cannot be linked back to personal identifiable information. This will help in the design of the appropriate consent language.

Sharing Data with the Rare Diseases Clinical Research Network (RDCRN)

The Rare Diseases Clinical Research Network (RDCRN) is an initiative of the National Institutes of Health (NIH) to advance medical research on rare diseases. A long-term goal of the network is to improve diagnosis and treatment of rare disease conditions. Knowledge and data sharing is an integral part of the RDCRN because it helps scientists understand commonalities among different rare diseases and facilitates rapid advancement of research.

Your clinical information, including clinical exam results and other data (referred to as "your data") [add other types of data such as genomic, medical images, histopathology, etc.] collected for this study may be stored in multiple locations. Your data will be stored within the National Center for Advancing Translational Sciences (NCATS) RDCRN Operational Cloud Environment at NIH and managed by the NIH-funded RDCRN Data Management and Coordinating Center (DMCC) and also will be also part of a Federal data repository hosted by the NIH. [Add other locations as necessary] The NIH will be

responsible for your data kept in the federal data repository. They will care for your data and make decisions about how they are used. Your clinical data may be stored in perpetuity in these locations.

The RDCRN DMCC uses several layers of protection including: password protected access, data encryption, and constant network monitoring to ensure the security of the stored clinical data. The DMCC systems comply with all applicable guidelines to ensure confidentiality, data integrity, and reliability. We will protect the confidentiality of your information to the extent possible. Your name and other identifying information will be kept locally at the clinical site you attend, in order to contact you, and will not appear in the data stored in the clinical research database or in the federal data repository. At those locations, the data will have a code that links to your identifying information. The code key will be kept in a locked location separate from your health and research information. The code key can only be accessed by people on the research team who have permission from the site investigator. [If a study utilizes e-Consent, the location of identifying information should be outlined].

You may be assigned a code number called a Global Unique Identifier (GUID) using an NIH GUID system. The GUID is a unique code made up of letters and numbers that allows researchers to share data from other studies in which you have participated without letting others know who you are. A GUID does not contain direct identifiers, and you cannot be identified using only the GUID. To generate the GUID, we will ask you for your full date of birth (day, month, year), first name at birth, last name at birth, middle name at birth (if applicable), gender at birth, city/municipality of birth, country of birth. This personally identifiable information (PII) will be processed using the NIH Centralized GUID generator software program. Once the GUID is produced, there is no way to get back to your PII. The software will not keep your PII, but will have enough information to determine if you already have a GUID assigned in the system. If you participate in another project and provide the same PII, you will be assigned the same GUID. Your GUID will be part of our research records.

We would like to make your data, without direct personal identifiers, available for other research studies that may be done in the future. Our goal is to make more research possible to learn about health and disease. Future research may be about similar diseases or conditions to this study. However, research could also be about unrelated diseases, conditions, or other aspects of health. These studies may be done by researchers at other institutions, including commercial entities and they may be from anywhere in the world. They may work at universities or hospitals. They may work for a government. They may work for companies to make new medicines or products, which may generate profit. You may not benefit directly from allowing your information to be shared. You will not be paid for the future sharing or future use of your data. There will be an approval process for researchers who want to work with study records that might identify you. They will have to tell the NIH, through a data access request and application process, about the research they want to do. They will have to do ethics training and their study needs to have IRB approval. They will have to sign a legal agreement stating they will not try to find out who you are.

Participating in this study means you agree to share your data. You can change your mind later, but researchers may still use your data that have already been shared. If you do not want your data used for other research projects, you should not participate in this study.

Alternate abbreviated paragraph:

This research is part of the Rare Diseases Clinical Research Network (RDCRN), an initiative of the National Institutes of Health (NIH) to advance medical research on rare diseases. A long-term goal of the network is to improve diagnosis and treatment of rare disease conditions. The clinical information collected for this study will be stored within the National Center for Advancing Translational Sciences (NCATS) RDCRN Operational Cloud Environment at NIH and managed by the NIH-funded RDCRN Data Management and Coordinating Center (DMCC) and also will be also part of a Federal data repository hosted by the NIH. The NIH and the data management center uses several layers of protection for the clinical data stored there. It meets all of the local and federal security requirements for research datacenters. The NIH may make your data, without direct personal identifiers, available for other research studies in the future. Future research may be about similar diseases or conditions to this study, but could also be about unrelated diseases, conditions, or other aspects of health. Researchers who want access to your data will have to tell the NIH, through a data access request and application process, about the research they want to do. They will have to do ethics training and have IRB approval to do the research. They will have to sign a legal agreement stating they will not try to find out who you are.