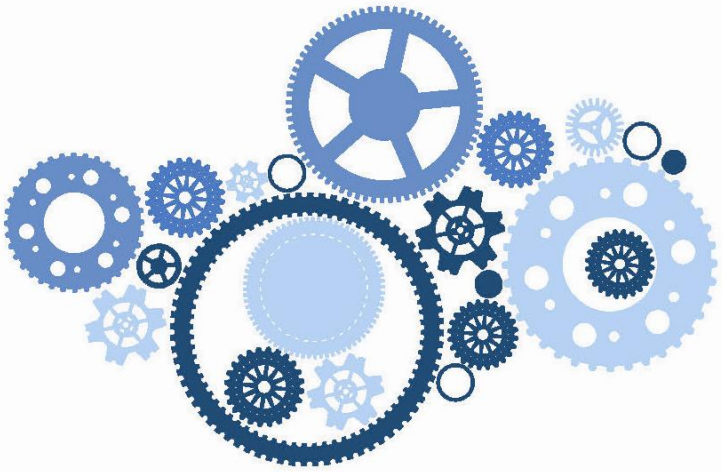
FasterCures
A CENTER OF THE MILKEN INSTITUTE



Data Sharing

THE *FASTERCURES* CONSORTIA-PEDIA REPORT

About the *FasterCures* Consortia-pedia project:

FasterCures initiated the Consortia-pedia project to better understand the breadth and scope of approaches that a wide range of consortia have adopted to bring together non-traditional partners with a shared R&D goal. Since 2012, our analysis of more than 350 biomedical research consortia has been aimed to better understand how different stakeholders are using this model of partnership to address shared unmet needs.

To better understand consortia models, *FasterCures* analyzed 21 efforts that represent the diversity of models used to bring together non-traditional partners to accelerate biomedical research. We present our analysis under seven partnership components.

1. Governance
2. Financing
3. Human Capital
4. Intellectual Property
5. Data Sharing
6. Patient Participation
7. Measurement of Impact

Each component is a chapter in the Consortia-pedia report and can be downloaded at: www.fastercures.org/consortiapedia.

UPDATED May 2014

KEY POINTS – DATA SHARING:

- Address **data-sharing rules and conditions** early with pre-negotiated agreements.
- Find mechanisms to **enforce data-sharing agreements** since many participants may be reluctant to share their own data.
- Develop a **common and standardized data model** to ensure effective data sharing.
- **Understand data-user needs** – making data accessible does not mean people will use them.

The rapid transition of research data into an electronic and portable format continues to revolutionize biomedical research not only by empowering emerging disciplines such as bioinformatics, but also by creating a currency that can be readily exchanged among partnering organizations. It's widely accepted that sharing data across organizations is the right thing to do to advance research, and the consortium model provides an infrastructure to create, test, and optimize an environment of trust and collaboration. The intended outcome for many consortia is to create a data-exchange framework to support both its immediate strategic research objectives as well as sustain cross-organization partnerships outside of the consortium's formalized boundaries.

But the first step, establishing data-sharing agreements among the organizations, is one of the biggest challenges for many consortia, particularly those with research objectives that depend on aggregating and making data available across researchers. *FasterCures'* interviews unveiled numerous stories of "data promised but not delivered" and more stories of how data were contributed in a format that is "useless" to the consortium. Even when data were equally contributed, the transactional costs and resources needed for these efforts were generally much more than anticipated.

The following are some considerations that many consortia shared with us during our interviews:

- How will data be shared, and how often? Examples include informal and infrequent, to systematic and milestone-bound
- What type of data and in what format? Examples include raw pre-clinical data, individualized data from a clinical trial, or summaries of data provided as a publication or presentation
- Will there be measures to ensure a certain quality of data? Examples include analysis and control measures to reduce concerns of "garbage-in, garbage-out"
- Will the data have additional annotations? These include additional descriptions on how the data were collected and aggregated
- Will the data follow standards and interoperability requirements to ensure equal contribution and access?
- What rules govern data-access for research partners and/or the public?
- What privacy, current policy, and ethical and legal concerns apply, particularly if patient-level data will be shared?

Many consortia describe their data-exchange policies within their intellectual property and data-sharing agreements. These detail the requirements, restrictions, and timelines for contributing data, as well as include data-access processes for partners and the public. For example, some consortia have data-access rules based on the amount of data contributed by a partner, or, if consortium-based data are intended for release to the public, may base the level of access on an external researcher's qualifications and objectives.

Interoperability and data-sharing standards

The type of data shared within a consortium could range from historical data that have already been collected independently to data that are actively collected by researchers within different sites. There is also a diversity of methods used to pool, standardize, and contribute data, creating a challenge for many of these consortia, as datasets often differ in structure, quality, and content. The diversity of results is only amplified with the varying methods of analysis and interpretation.

For data-heavy consortia, it becomes important that partners use interoperable platforms to collect, store, and analyze the data, along with annotations that describe how the data were collected and processed. In addition to technical infrastructure, participants should agree to use a set of data ontologies, which are vocabularies that define relationships among datasets. These steps ensure that the collection and analysis of data are conducted uniformly across participants, as well as provide the opportunity for any partner to query, scrutinize, and validate all of the data aggregated within a consortium's research activities.

Types of data shared within a consortium include:

- Individual or aggregated data
- Raw, annotated, or interpreted pre-clinical data (such as molecular analysis – genomic, proteomic, etc.)
- Raw, abstracted, or summarized clinical trial data
- Anonymized and patient-identifiable clinical data (observational, phenotypic)
- Data generated by mobile applications and platforms
- Publications and presentations

There are several consortia that proactively address standardization and technical concerns prior to making the data available to the other consortium partners. For example, the Innovative Medicines Initiative's (IMI) eTOX project, a consortium that aims to share toxicity data to create tools for patient safety, requires that all data are initially sent to a third-party contract research organization for standardization and quality control prior to being shared within the data portal. eTOX also developed ontologies to help simplify data contribution and access, such as those that standardize descriptions of anatomical location, pathology, clinical chemistry and toxicology, cell and tissue type, observational findings, species and strain of animals used in studies, and study design. Abiding by these ontologies and quality control measures inherently increases the value of both the aggregated data-set and the methods used for their collection and analysis to the broader scientific community.

It would be detrimental if each consortium created its own data standard, and fortunately this is not the case for many partnerships that depend on sharing clinical trial data. TransCelerate BioPharma, Project Data Sphere, and the IMI are some examples of consortia that have efforts linked with the Clinical Data Interchange Standards Consortium (CDISC) to standardize the language and protocol used for their data collection. CDISC and the Critical Path Institute have also partnered to create the Coalition for Accelerating Standards and Therapies (CFAST) to streamline methods used to collect, store, and evaluate data from multiple clinical studies for specific diseases. In collaboration with National Institutes of Health's (NIH) National Institute of Neurological Disorders and Stroke, this consortium developed the CDISC Therapeutic Area Data Standard for Parkinson's Disease, with similar "therapeutic area data standards" being developed for other diseases in collaboration with TransCelerate BioPharma, NIH, and the Food and Drug Administration. Leveraging the expertise and resources of another consortium not only helps maintain a mission focus, but also helps ensure that the data and collection methodology developed by any of these efforts have greater value to the scientific community because cross-consortium clinical data can be readily aggregated and analyzed. Table 1 provides examples of platforms for sharing different types of data.

Table 1: Examples of platforms used to share data

Consortium	Type of data	Platforms for sharing
Myelin Repair Foundation	Actively collected research data	Online cloud, Web conferencing, and face-to-face meetings
Health and Environmental Sciences Institute (HESI)	Actively collected research data	SharePoint, email, and face-to-face meetings. Plans to set up more rigorous system for complex data.
Project Data Sphere	Historical and anonymized participant-level clinical trial data	Secure and access-controlled database on SAS infrastructure. Patient-level data are automatically anonymized.
eTOX	Actively collected research data	Data first sent to a contract research organization for standardization. Different levels of engagement and contribution to the eTOX project and access to data are based on the terms of the agreement between the partner and the consortium.

Data quality

Even if the data are collected, annotated, and pooled following standardized procedures, their utility by the broader community is directly related to the quality of the methods used to collect the original dataset. Quality is dependent on multiple factors that can include: implementation of proper QA/QC measures at different stages of collection and analysis, quality of resources, institutional infrastructure, and investigator expertise.

But it's not just about bits and bytes. One example of a concern expressed by consortia working with molecular-based data is the variable quality of biospecimens used for analysis, particularly if they were not collected or handled uniformly. To reduce this type of variability, consortia such as the Parkinson's Progression Markers Initiative (PPMI) send all of their collected biospecimens to one research site for quality control, coordinating this work with a verification study site that performs the molecular analysis. Another approach is used by the Multiple Myeloma Research Consortium (MMRC), which operates its own tissue bank to ensure that biospecimens are uniformly collected and stored in an environment that is supported with weekly quality assurance reviews and audits.

Other consortia may not have the resources or infrastructure to do the pre-analytical preparations exemplified by PPMI and MMRC. At a minimum, many require that their research partners adopt principles of good laboratory practices (GLP), which standardize the procedures used to collect, handle, and analyze biospecimens, and also include a requirement to track and trace the data back to the original analytical source. If these processes cannot be standardized, it becomes increasingly important to annotate data with descriptions on the pre-analytical processes so that others can reproduce and validate the original dataset. A more in-depth analysis on the challenges and solutions for the molecular analysis of patient biospecimens can be found in the *FasterCures* publication "Banking on Trust."

Public dissemination of data

Most consortia have a mission to share their research findings with the public, and our interviews found that the timing and manner in which this information is disseminated vary among many consortia (see Table 2). For example, very few groups have a completely open-access policy, and many have different levels of control for data dissemination. As mentioned earlier, these policies are described in the intellectual property and data-sharing agreements, and are often enforced through rules that limit the access of data to the participants and the public.

Table 2: Examples of the ways in which data are released to the public

Method for public release of data	Examples
Unconditional and immediate release of data/tools	• Sage Bionetwork's Clinical Trial Comparator Arm Partnership • Foundation for the National Institutes of Health's (FNIH) Observational Medical Outcomes Partnership
Immediate release of data to external researchers with proper qualifications and objectives	• FNIH's Alzheimer's Disease Neuroimaging Initiative and I-SPY 2 • Michael J. Fox Foundation's Parkinson's Progression Markers Initiative • Project Data Sphere • Critical Path Institute's Coalition Against Major Diseases
Release of data after sponsor-exclusive time period	• Multiple Myeloma Research Foundation's CoMMpass Study • Some programs from Biomarkers Consortium
Publications and presentations	• HESI • TransCelerate BioPharma

One concern preventing many consortia from adopting policies that allow for the unconditional release of data to the public is the potential for its irresponsible or inappropriate use. Many consortia structure their data-access rules to minimize this possibility. For example, many consortia provide clinical research data to the public only after reviewing the requestor's qualifications and experimental plans. Even after the formal application process, permission may come with provisions that limit the use of data and the venues where results can be published. Most of this due-diligence and policy work is done by data access and publications committees, typically composed of consortium partners and external experts. For example, Alzheimer's Disease Neuroimaging Initiative (ADNI) has a data and publications committee that develops and enforces policies describing access to ADNI data, approves access for external researchers, reviews manuscripts that emerge from ADNI data use, and tracks publications.

There are also ethical concerns where patients may not have provided consent for the reanalysis of their previously collected data, particularly if their data are going to be analyzed for a different purpose. The datasets themselves may also require additional work by the consortium to address patient privacy concerns, which can be particularly time consuming if the data come from multiple sources. One solution is being implemented by Project Data Sphere, which incorporates a streamlined data preparation step that uses a team of experts who are able to prepare and de-identify datasets for shared use.

Utility of the accessible data

Consortia may still encounter additional challenges for collecting and disseminating their data. One example is the assurance that the collaboration's data have utility to the outside community. The format of the data, such as raw data, may not appeal to outside researchers, and some of the consortia that we interviewed went through significant effort to curate their datasets for the public. Some of the data-intensive consortia recommended the inclusion of potential users as part of their steering committees to provide an understanding of the external research community's data requirements, both technical and use-cases.

Sage Bionetworks leads one example to make data more readily usable to the public through its Synapse platform, an open-data system that intends to support collaborative research. In partnership with several consortia, this platform is being leveraged to make datasets interoperable within an open science framework.[1] Partners include the National Cancer Institute's The Cancer Genome Atlas as well as data collected by several consortia focused on Alzheimer's disease and rheumatoid arthritis. Sage Bionetworks also hosts the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project to incentivize and engage the broader scientific community to use their available datasets and the Synapse platform, using a series of challenges that ask the community to develop predictive models for specific diseases.

[1] http://sagebase.org/2013/09/26/synapse-enables-18-novel-cancer-publications/, Accessed on 12/30/2013

As described earlier, consortia such as ADNI track the requests and usage of their data and publish these figures on their "Data Usage Stats" Web page. These analyses divide data requestors by geography and sector with statistics highlighting the publications that emerge from the use and interpretation of ADNI data. In addition to these active statistics, ADNI's leadership also published a comprehensive and retrospective review in *Alzheimer's and Dementia*,[2] applying the advances supported by this consortium to the Alzheimer's disease research landscape.

For more information and the latest updates on the *FasterCures* Consortia-pedia, visit www.fastercures.org.

[2] Weiner, MW et al., The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception, Alzheimer's and Dementia, 8 (2012), S1-S68.