Data & Policy (2022), 4: e29
doi:10.1017/dap.2022.22

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

A journey toward an open data culture through transformation of shared data into a data resource

Scott D. Kahn[1,*] and Anne Koralova[2]

[1]Precision Medicine Now, Rancho Santa Fe, California, USA
[2]Type 1 Diabetes Program, The Leona M. and Harry B. Helmsley Charitable Trust, New York, New York, USA
*Corresponding author. Email: kahnteam@gmail.com

Received: 16 February 2022; **Revised:** 30 June 2022; **Accepted:** 05 August 2022

Key words: Data sharing; data privacy; GDPR; open data; open science

Abstract

The transition to open data practices is straightforward albeit surprisingly challenging to implement largely due to cultural and policy issues. A general data sharing framework is presented along with two case studies that highlight these challenges and offer practical solutions that can be adjusted depending on the type of data collected, the country in which the study is initiated, and the prevailing research culture. Embracing the constraints imposed by data privacy considerations, especially for biomedical data, must be emphasized for data outside of the United States until data privacy law(s) are established at the Federal and/or State level.

Policy Significance Statement

There is a gap between the objectives sought through open research data and the implementation of principles that encourage and promote data sharing by researchers and research teams. Funders of research must consider approaches that account for concerns including professional advancement and governmental regulations that often dissuade the sharing of data. Building on the concept of findable, accessible, interoperable, and reusable data, and through consideration of the impact of data privacy regulations, a series of implementation policies are shared as exemplars upon which for others to build. Remaining policy challenges are framed for future resolution.

1. Introduction

The allure of large aggregations of biological and medical data to improve the understanding of disease causation, disease progression, and therapeutic intervention has expanded the thinking in precision medicine over the last decade and has a driving influence toward the establishment of Data Commons (Brenner, 2007).[1] And while the intersection between the resultant Big Data and analytic methods such as the artificial intelligence (AI) subdisciplines of machine learning (ML) and deep learning is posited to enable advances across the health spectrum, there are surprisingly few examples where these aggregated

[1] The official portal for European data (https://data.europa.eu/euodp/en/data/).

data have proved adequate when interrogated by even the most technologically advanced analyses using AI methods (Rocher et al., 2019).

Most/all existing caches of research and clinical data exemplify the many paradoxical limitations on the application of analytic methods, inclusive of AI, to advance health knowledge due to lack of interoperability within an "aggregated" dataset. Information also tends to be siloed due to factors that either limit access to data or limit the use of data that are accessible. Accessibility may be affected by data use restrictions imposed by an institutional review board (IRB) or limited by the details of the informed consent employed. Accessibility is most often gated either by an investigator's desire to maximize publication opportunities for themselves and the role these publications play in professional advancement, or a need to keep data confidential to secure intellectual property for potential discoveries. And while accessibility challenges are significant, the challenges to making data that are accessible interoperable and (re)usable are even more vexing.

Accessible data from different studies or sources can be difficult to combine due to differences in the ontologies used to organize the data elements, differences in how individual data elements in the data dictionary are defined, and basic formatting differences between datasets. Addressing these challenges through person-driven curation sort is difficult to scale. Curation can also address systematic differences between datasets once these differences are discovered. Random errors in a dataset must be addressed on a case-by-case basis which does not scale to larger datasets or to additional datasets for aggregation. Unstructured data can be structured using natural language processing (NLP) methods once the NLP methods have been trained on exemplar data. And yet, even with all these approaches to harmonize datasets, many datasets still cannot be aggregated with others because they share no common data elements around which the dataset records can be combined.

The practice of collecting data as parsimoniously as possible, to both inform the study hypothesis while minimizing participant burden, leads to datasets that are hopelessly siloed. This is caused by a lack common data elements that render them poor candidates for aggregation into the rich datasets necessary for AI and ML exploration. This affects more datasets than one might imagine, especially those derived from clinical trials. Ideally, data should be enriched so that aggregation across studies within a disease category, and between disease classifications, can be used to explore potential relationships or comorbidities that would be discoverable by, for example, AI analytics. One straightforward enrichment strategy is to include genomics data for each sample to provide a universal descriptor that can be used to organize samples regardless of their source.

2. Methods

2.1. Encouraging a culture of data sharing

The Leona M. and Harry B. Helmsley Charitable Trust (Helmsley) has begun working with grantees to ensure that the necessary processes and infrastructure are put in place so that the data resulting from their grants are shareable with other researchers. By focusing on the many operational challenges in making their studies' data findable, accessible, interoperable, and reusable (FAIR), Helmsley has charted a course to systematically address each of the challenges that were introduced earlier. It was apparent that working through these challenges would support the broader community's efforts toward Open Data that are an important facet of the larger Open Science movement (e.g., Woelfle et al., 2011) and its drive to have Open Access to published articles, Open Data that adhere to the FAIR principles (Wilkinson et al., 2016), and Open-Source software and algorithms that result from research activities.

Helmsley's efforts to develop policies to promote data sharing in the studies it funds, pioneered within the Type 1 Diabetes (T1D) Program, was motivated by a desire to create an environment where the broadest research community could ethically use data from Helmsley-funded projects for the greatest

https://doi.org/10.1017/dap.2022.22 Published online by Cambridge University Press

impact. The data sharing initiative to promote data sharing by its funded grants required a year to develop and was built upon four principles:

1. a dedication to improving the lives of people by making relevant data available in the community in a timely manner to maximize impact,
2. the protection of a study participant's data by providing the necessary safeguards that will protect the identity and confidentiality of each participant,
3. providing for the recognition of attributions that honor and respect stakeholders' own incentives, and that recognize the attributions of the investigators and their collaborators, and
4. ensuring the appropriate accountability and responsibility such that all data will be collected, stored, and shared in a well-defined and consistent process that follows data standards that guarantee the quality, security, and equitable access to the data.

2.2. Development of processes and policies

Implementing a formal data sharing plan for Helmsley's grantees required updates to the grant application process, and to the information requested from the applicants. There was a need to discuss the inclusion of data sharing language in any informed consents prior to IRB approval to ensure that broad access to the data was supported by a study's participants and approved by the IRB. Contractual terms were inserted within the grant agreement that defines a period of embargo for the study data by the project's investigators typically not to exceed 12 months, and a data management plan that summarizes how data are managed and protected during the project was requested. Finally, the longer-term plan for storage of the data in a public registry has been requested from the grantee and should include a sustainment plan for the data resource. And while this last aspect of the data management plan remains a work in progress as open data networks around the globe continue to mature, Helmsley has supported the open data platform developed by Vivli[2] that can organize and archive observational and clinical trial data.

2.3. Development of relevant data standards

The standardization of clinical data is difficult for a variety of social/cultural and procedural (e.g., IRB or ethical review board) reasons and is exasperated by the state of data standards that span data collection through to data analytics and reporting of results. Many data standards exist for the description of disease and disease management; standards developed by the Clinical Data Interchange Standards Consortium (CDISC)[3] are employed by the FDA for the submission of clinical trial data for drug approvals.

While nascent standards for diabetes exist within the CDISC, they lack the definition of the clinical environment for T1D patients. Fortunately, the CDISC has a strong foundation and process infrastructure with which to revise and extend these standards to address gaps in their coverage of the necessary terms and relationships. The improvement of available standards has been targeted by Helmsley through investments in four areas of standards enrichment that will directly impact many T1D studies: pediatrics, screening, and monitoring (i.e., identifying individuals at risk and providing follow-up care), devices, and exercise and nutrition.

In the development of each set of standards, a group of subject matter experts (SMEs) work to gain consensus among data modeling experts within the CDISC before a proposed standard is released for community review. Many of the SMEs involved in the standards development are investigators that have used existing CDISC standards in their projects; each has structured their data to facilitate the adoption of the standards under development once they have passed community review and final release. While the T1D data requirements to describe screening and monitoring and to capture device data from continuous glucose monitors and insulin pumps are specific to this disease, the standards definition and extensions around exercise and nutrition can have a much broader application.

[2] Vivli center for global clinical research data (https://vivli.org/).
[3] CDISC portal (https://www.cdisc.org/).

https://doi.org/10.1017/dap.2022.22 Published online by Cambridge University Press

Due to the FDA requirements to use CDISC standards for trial data submitted for regulatory review, pharmaceutical and device companies have in-house expertise and experience with implementing these standards. However, academic investigators face significant challenges around the use of the standards they are unfamiliar with and that often require an expertise in data science and informatics that may be in short supply for many academic research teams.

2.4. Implementation challenges

The practical implementation of these data sharing policies has focused on several key items that remain a work in progress. The first implementation challenge has been to understand a project's informed consent to ensure that the requested consent enables broad data sharing, and that the informed consents were consistent across a project's participants. The assessments of informed consents are left to the grantee's organization to attest compliance. This is especially important for informed consents in multiple and non-English languages. The next challenge involves alignment on a data embargo; Helmsley prefers an embargo that does not exceed 12 months. A more vexing challenge involves the adoption of data standards to avoid the many curation challenges that historically plague shared datasets. The practical issues center on the local data models that each investigational group has adopted (and has used in previous research) to collect and analyze clinical data, and the general lack of familiarity and expertise with any standard data model. A solution is to employ local data models to conduct the research and to extract, translate, and load these data into data models such as the CDISC prior to archival submission. The final implementation challenge, concerning the embellishment of a dataset so that each specimen is characterized by transferrable data elements that are shared with external datasets, requires funding to generate the additional data and a modification of the study protocol to support a collection of ancillary data that do not relate per se to the study hypothesis.

Once a study has been initiated under the Helmsley data sharing policies described above, and the embargo has lapsed, the sustainment of the dataset (once the study's funding has been exhausted) can be one of the most challenging to solve because it is not typically part of the project planning process, and due to a general lack of global resources that have funding to operate as public archives/repositories. Additionally, the incentives for the researcher to comply with the data sharing policies shift from the contractual obligations connected to the funding to the community obligations around the promise to share study data in a timely fashion as well as aspirations for follow-on funding for the project.

2.5. Putting data sharing policies into practice on the global stage

Helmsley shaped the initial implementation of its data sharing policies by working with two large projects: one to study the prevention of autoimmunity associated with T1D, and one to study how exercise type and timing impact glucose management in people with T1D. During the pre-grant proposal phase that spanned months for each project, Helmsley's leadership and data sharing experts communicated the desire to incorporate data sharing language within the grant that was vetted with the respective project teams until consensus was reached around the four data sharing principles. The international project involved multiple participating sites, and a U.S.-based project performed remote recruitment by a coordinating center, to carry out multiyear research plans. And while both project teams were aligned with Helmsley's data sharing philosophy, there are still several challenges (i.e., both cultural and technological) on which Helmsley continues to work with each project team to fully operationalize the data sharing policies.

The Global Platform for the Prevention of Autoimmune Diabetes (GPPAD)[4,5] is a clinical trial platform designed to screen newborns for genetic risk of T1D and enroll at-risk children in clinical trials to test interventions to prevent autoimmunity. The platform spans multiple sites in Germany, Belgium,

[4] Global platform for the prevention of autoimmune diabetes portal (https://www.gppad.org/en/).

[5] Data can be obtained for further research by contacting the Helmholtz Institute via the GPPAD website. A Data Transfer Agreement may be required for data access outside of the European Union.

https://doi.org/10.1017/dap.2022.22 Published online by Cambridge University Press

Poland, Sweden, and the United Kingdom. To date, the GPPAD team has screened over 245,000 newborns and enrolled over 1,000 children in its first clinical trial. In GPPAD, genetic risk is assessed through consideration of 47 gene markers; a goal of Helmsley's data sharing policy is to broadly share these data while complying with the General Data Privacy Regulation (GDPR; Skendžić et al., 2018) that protects the data privacy rights of all EU citizens. The challenge in sharing *any* genetic data rests in the lack of specifics in GDPR and the need to perform a data protection impact assessment (DPIA) to assess the risk that the collected genetic data might pose to each individual combined with the likelihood that these risks would be manifest for an individual. No precedent existed when the project began (before the enactment of GDPR!) and the workshops to perform a DPIA were only held after a significant amount of genetic screening data was already collected. The DPIA was performed by a professional organization without the involvement of Helmsley. Fortunately, the desired data sharing was found to be consistent with the informed consent provided by each individual once a data transfer agreement is established between the (external) researcher and the institution acting as the data controller. It is worth noting that the sharing of biomedical data is called out within GDPR; each type of data being shared must be evaluated by a DPIA. The key lesson learned from this experience is to perform all applicable data privacy assessments before data are collected from participants to ensure that all possible protections and permissions can be established prior to project initiation.

The second project, the T1D Exercise Initiative,[6] is based in the United States and seeks to uncover how different exercise types, intensities, duration, and timing impact glucose response in people with T1D during and after exercise. The study recruitment is done by a public website and includes 600 adults with T1D. The objective is to pave the way for interventions that could inform clinical guidelines, decision support, and automated insulin delivery systems. The data collected have been consented in the broadest possible manner albeit the dataset lacks any transferrable characterization of a study participant that could be used to join their data record to other (external) studies. This was addressed by augmenting the dataset with genotyping information for each individual that can be used to make comparisons between genetically similar individuals with and without T1D, for example. The genotyping data[7] are not used within the study itself, rather its sole purpose is to provide a universal descriptor that enables more widespread use of the shared dataset in other studies.[8] And while the data model used within the study was based loosely upon the CDISC diabetes standard, researchers within the study used a bespoke data model for their analysis of the project data. As such, there were some challenges in transforming the data collected into the newly developed T1D CDISC standards prior to submission of the study data to a data archive. These types of data transformations can require many months and specialized personnel to implement.

3. Summary of Helmsley's Data Sharing Journey

This report has described the development of a data sharing policy that has been implemented for T1D projects funded by the Helmsley Charitable Trust over a multiyear time frame. It highlights the importance of dissecting the many challenges in making data shareable AND reusable, the importance of building on standards such as the CDISC, and the ability to transform standalone data into a data resource by adding in transferable specimen descriptors such as genotyping data. Our experiences with the European GPPAD project and the Exercise study in the United States have highlighted several challenges in making data sharing a reality. Of these challenges, the importance of addressing the requirements of data privacy upfront is one that is mostly applicable in Europe due to GDPR, but that will likely be an ongoing issue in the United States and globally as data privacy regulations are enacted by governments. This study has focused on biomedical data sharing; however, the broad regulations around data privacy are applicable beyond this class of data. An additional challenge involves the transformation of project data from the

[6] Type 1 diabetes exercise initiative portal (https://www.jaeb.org/t1dexi/).

[7] Genotyping arrays are low cost, and the expertise to generate such data is widespread.

[8] One must also account for any data privacy considerations that accompany the generation of genetic data for study participants.

https://doi.org/10.1017/dap.2022.22 Published online by Cambridge University Press

formats used to collect and analyze the data by the project team(s) into a standards-based format that can promote widespread data sharing. Both challenges, once surmounted, address some of the more vexing issues in implementing FAIR data. One that remains involves the creation of archival sites that can present a critical mass of data that will enable novel investigations via data reuse. And once the dataset(s) are available in a standard data model on an archival site, the measure of success of data sharing will be evaluated by the quantity and types of data reuse that are achieved.

The journey toward an open data culture is an ongoing activity, and one that will have additional challenges to overcome. Some of the cultural barriers are being addressed alongside the evolution of data privacy laws and how they apply to health-related data.[9] Still, making datasets truly reusable—effectively transforming shared data into a data resource—continues to require deeper planning and consideration. However, a commitment to open data by principal investigators and their institutions, by study participants and their informed consents to share their data, and by the funding organizations, such as the Helmsley Charitable Trust, continues to gain momentum and that holds promise to unlock new discoveries that would have been impossible without the sharing and reuse of scientific data.

Acknowledgment. The authors would like to acknowledge the vision and leadership of David Panzirer and Gina Agiostratidou in the development of the Helmsley Trust's data sharing policies.

Funding Statement. The preparation of this article received no funding.

Competing Interests. The authors declare that they have no competing interests.

Author Contributions. Conceptualization: S.D.K.; Writing—original draft: S.D.K.; Writing—review and editing: both authors.

Data Availability Statement. No data are directly associated with this article, but access to the project data discussed will be deposited in a public repository (https://vivli.org) or via the data controller (Helmholtz Institute, Munich, Germany) as described.

References

Brenner S(2007) Common sense for our genomes. *Nature 449*, 783–784. https://doi.org/10.1038/449783a

Rocher L, **Hendrickx JM**, **de Montjoye YA** (2019) Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications 10*, 3069. https://doi.org/10.1038/s41467-019-10933-3

Skendžić A, **Kovačić B**, **Tijan E** (2018) General data protection regulation—Protection of personal data in an organization. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE. pp. 1370–1375. https://doi.org/10.23919/MIPRO.2018.8400247

Wilkinson MD, **Dumontier M**, **Aalbersberg IJ**, **Appleton G**, **Axton M**, **Baak A**, **Blomberg N**, **Boiten J-W**, **da Silva Santos LB**, **Bourne PE**, **Bouwman J**, **Brookes AJ**, **Clark T**, **Crosas M**, **Dillo I**, **Dumon O**, **Edmunds S**, **Evelo CT**, **Finkers R**, **Gonzalez-Beltran A**, **Gray AJG**, **Groth P**, **Goble C**, **Grethe JS**, **Heringa J**, **'t Hoen PAC**, **Hooft R**, **Kuhn T**, **Kok R**, **Kok J**, **Lusher SJ**, **Martone ME**, **Mons A**, **Packer AL** , **Persson B**, **Rocca-Serra P**, **Roos M**, **van Schaik R** , **Sansone S-A**, **Schultes E**, **Sengstag T**, **Slater T**, **Strawn G**, **Swertz MA**, **Thompson M**, **van der Lei J**, **van Mulligen E**, **Velterop J**, **Waagmeester A**, **Wittenburg P**, **Wolstencroft K**, **Zhao J**, **Mons B** (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Woelfle M, **Olliaro P**, **Todd MH** (2011) Open science is a research accelerator. *Nature Chemistry 3*(10), 745–748. https://doi.org/10.1038/nchem.1149

[9] See for example the ongoing discussion of the European Health Data Space at https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en.

Cite this article: Kahn S. D and Koralova A (2022). A journey toward an open data culture through transformation of shared data into a data resource. *Data & Policy*, 4: e29. doi:10.1017/dap.2022.22