Deloitte.

The Case for Modern Data Governance to Facilitate Participant-Centric Research for Childhood Cancer

This data governance framework improves the sharing of childhood cancer data and workflows to accelerate research and discovery.

Introduction

In childhood cancer research, the scarcity of research data has caused researchers to seek new and efficient ways to curate and share knowledge. The proposed data governance framework addresses this challenge by systematically coordinating research data management early on to increase the quality, availability, shareability, and reusability of childhood cancer data and workflows.

Overview

The data governance framework enables principal investigators (PIs) and research institutions to work together to enhance data sharing through the following principles:

01. Privacy and Security (e.g., privacy, consent, security, identity and access management)

02. Open Science (e.g., stakeholder agreement, open science values, accountability to agreements, return-of-results for participants)

03. Digital Object Management (e.g., code management, workflows, and data to facilitate data discovery and retention, proper management of metadata documentation and standards)

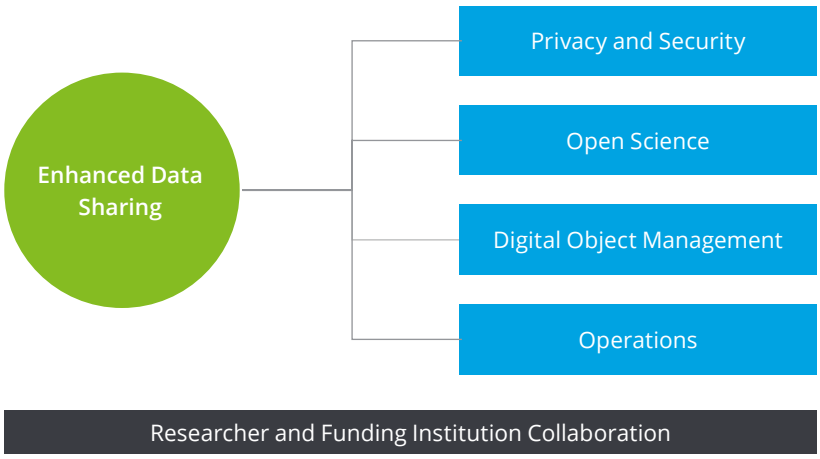
04. Operations (e.g., funding alignment with appropriate policy, technology platforms, and staffing)

Methodology

Stakeholders should adopt key data governance principles by working collaboratively with participants, research support teams, and data experts to manage data. In the sections below, we highlight concrete steps PIs and funding institutions can take to facilitate data governance and sharing for childhood cancer research.

01. Privacy and Security: Typically, participants prefer greater, rather than less, sharing of their data[1]. Given the limited availability of childhood cancer data, researchers must prioritize open-ended participant consent to share data. PIs should take an active role in developing consent forms that allow participants to understand and contribute their data to future research endeavors. Funding institutions should invest in the latest Identity and Access Management (IAM) technology to facilitate secure access to data. This will alleviate access burdens currently carried by researchers. To ensure compliance, funding institutions should follow up with periodic research audits to ensure security and consent requirements are met.

02. Open Science: PIs should take an active and early role to share data with participants and future researchers; funding institutions should mandate return-of-results to participants. To maintain participant-centric research, studies should always provide participants, guardians, and their respective point-of-care physicians with all relevant data, research findings, and recommendations. To ensure accountability by researchers, organizations must institute stronger rules that require data sharing and return-of-results to participants. The possibilities for open science fueled by cognitive computing, advances in machine and deep learning, and burgeoning data are endless and exciting,[2] , [3] therefore PIs and funding institutions should keep open science policies in mind to maximize the re-use of their data for future studies.



03. Digital Object Management: Digital objects should be part of an emergent ecosystem of data discovery. [4]Funding institutions and other stakeholders should define Findable, Accessible, Interoperable, and Reusable (FAIR) standards for childhood cancer data and workflows[5], while also creating policies that encourage data lifecycle management including provenance. PIs should, in turn, work collaboratively with their research support teams to ensure that all code, workflows, and associated metadata documentation are assigned a Digital Object Identifier (DOI) to facilitate reproducibility. A system promoting FAIR data at large will gain by improved efficiency, reproducibility, and the potential for new discoveries: efficiency, as there will be less need to reinvent the experiment[6].

04. Operations: Funding institutions need to provide adequate monies that allow for proper data management, as well as access to technology platforms for researchers and participants. Past funding by institutions like NIH has been invaluable to "discovery science", leading to advances in precision and participatory medicine[7]. Likewise, PIs can actively participate in providing feedback to funding to institutions regarding technology needs and data challenges. They also must staff their research support teams with data experts, security engineers, and technical specialists.

1. Meyer, M. (2018). Practical Tips for Ethical Data Sharing. Advances In Methods And Practices In Psychological Science, 1(1), 131–144. doi: 10.1177/2515245917747656
2. Klenk, J., Payne, P., Shrestha, R., & Edmunds, M. (2019). Open Science and the Future of Data Analytics. Consumer Informatics And Digital Health, 337–357. doi: 10.1007/978-3-319-96906-0_18
3. Grossman, R., Heath, A., Murphy, M., Patterson, M., & Wells, W. (2016). A Case for Data Commons: Toward Data Science as a Service. Computing In Science & Engineering, 18(5), 10–20. doi: 10.1109/mcse.2016.92
4. Bourne, P., Bonazzi, V., Dunn, M., Green, E., Guyer, M., & Komatsoulis, G. et al. (2015). The NIH Big Data to Knowledge (BD2K) initiative. Journal Of The American Medical Informatics Association, 22(6), 1114–1114. doi: 10.1093/jamia/ocv136
5. Clarke, D., Wang, L., Jones, A., Wojciechowicz, M., Torre, D., & Jagodnik, K. et al. (2019). FAIRshake: toolkit to evaluate the findability, accessibility, interoperability, and reusability of research digital resources. doi: 10.1101/657676
6. Bonazzi, V., & Bourne, P. (2017). Should biomedical research be like Airbnb?. PLOS Biology, 15(4), e2001818. doi: 10.1371/journal.pbio.2001818
7. Bui, A., & Van Horn, J. (2017). Envisioning the future of 'big data' biomedicine. Journal Of Biomedical Informatics, 69, 115–117. doi: 10.1016/j.jbi.2017.03.017

Results

To foster better research collaboration and efficiency, we have proposed a comprehensive data governance framework for childhood cancer that can accommodate secure and highly advanced technologies to track, manage, and analyze diverse data in a reproducible manner. There are many documented examples of how improved data sharing and access can lead to accelerated research discovery. In one study, the flexibility to query an extremely complex patient phenotype from medical records, when coupled by methods for representing genotypes, demonstrated how researchers can benefit from increased data access and integration.[8] In another study, automating the ability to reason across integrated data sources and providing users who pose inquiries with a dossier of translated answers coupled with full provenance and confidence in the results is critical to accelerate clinical and translational insights, drove new discoveries, facilitate serendipity, improve clinical—trial design, and ultimately improve clinical care.[9]

To realize the benefits of data sharing, principal investigators (PIs) and institutions must work collaboratively to govern data from the beginning of the data lifecycle, ensuring data sharing is a key priority. Fostering the four governance principles into a predefined framework for sharing data will streamline future data access, curation and management efforts, resulting in accelerated research discovery.

Evidence

There are many documented examples of how improved data sharing and access can lead to accelerated research discovery, as well as published consensus in the scientific community regarding the efficiency gains of improved data sharing. Please see below for detailed evidence and citation:

01. "Even when some of these considerations are not met, it is important to balance concerns about data privacy and data repurposing with the recognition that many participants prefer greater, rather than less, sharing of the data they contributed to science. Participants typically volunteer for research with the expectation that all reasonable efforts will be made to ensure that the results are correct, and data sharing for reanalysis and replication purposes helps to meet that objective. Also, participants who are members of groups that traditionally have been underrepresented in research may have a particular interest in having their data used widely (although their data may, for similar reasons, be more vulnerable to re-identification than other participants' data are)—"Meyer, M. (2018). Practical Tips for Ethical Data Sharing. Advances In Methods And Practices In Psychological Science, 1(1), 131–144. doi: 10.1177/2515245917747656

02. "Increasingly, open science also means full appreciation of the value of data and opening virtual floodgates that allow massive amounts of data to flow freely within and across the health sector and healthcare systems. Data-intensive computing and analytics allow researchers to generate new insights and discoveries in ways that were almost unimaginable until recently. The possibilities for open science fueled by cognitive computing, advances in machine and deep learning, and burgeoning data are endless and exciting—"Klenk, J., Payne, P., Shrestha, R., & Edmunds, M. (2019). Open Science and the Future of Data Analytics. Consumer Informatics And Digital Health, 337–357. doi: 10.1007/978-3-319-96906-0_18

03. "As the amount of scientific data continues to grow at ever faster rates, the research community is increasingly in need of flexible computational infrastructure that can support the entirety of the data science life cycle, including data exploration and discovery services to support data analysis and reanalysis as new data is added and scientific pipelines are refined... Across these case studies, several common requirements emerge, including the need for persistent digital identifier and metadata services, APIs, data portability, pay for compute capabilities, and data peering agreements between data commons—"Grossman, R., Heath, A., Murphy, M., Patterson, M., & Wells, W. (2016). A Case for Data Commons: Toward Data Science as a Service. Computing In Science & Engineering, 18(5), 10–20. doi: 10.1109/mcse.2016.92

04. "NIH Big Data to Knowledge (BD2K) aims to have these digital objects exist, not in isolation, but rather as part of an emergent ecosystem that is shared with the biomedical research community at large—"Bourne, P., Bonazzi, V., Dunn, M., Green, E., Guyer, M., & Komatsoulis, G. et al. (2015). The NIH Big Data to Knowledge (BD2K) initiative. Journal Of The American Medical Informatics Association, 22(6), 1114–1114. doi: 10.1093/jamia/ocv136

05. "The Findable, Accessible, Interoperable, and Reusable (FAIR) guiding principles have prompted many stakeholders to consider strategies for tackling this challenge by making these digital resources follow common standards and best practices so that they can become more integrated and organized—"Clarke, D., Wang, L., Jones, A., Wojciechowicz, M., Torre, D., & Jagodnik, K. et al. (2019). FAIRshake: toolkit to evaluate the findability, accessibility, interoperability, and reusability of research digital resources. doi: 10.1101/657676

8. 8.Murphy, S., Avillach, P., Bellazzi, R., Phillips, L., Gabetta, M., & Eran, A. et al. (2017). Combining clinical and genomics queries using i2b2 – Three methods. PLOS ONE, 12(4), e0172187. doi: 10.1371/journal.pone.0172187
9. 9.The Biomedical Data Translator Program: Conception, Culture, and Community. (2018). Clinical And Translational Science, 12(2), 91–94. doi: 10.1111/cts.12592

06. "The potential of FAIR —Findable, Accessible, Interoperable, and Reusable—scientific content will help, for inherent in the model is a reliance on standards. Moreover, if implemented correctly, the provider of scientific content will gain from metrics that describe how much that content is used and also have an easy pathway to meeting the sharing requirements from funders. Users of content will be able to find and use relevant content. The system at large will gain by improved efficiency, reproducibility, and the potential for new discoveries: efficiency, as there will be less need to reinvent the experiment—even those from your own laboratory; reproducibility, by making available the complete workflow; and new discoveries, from the aggregation of multiple types of data—"Bonazzi, V., & Bourne, P. (2017). Should biomedical research be like Airbnb?. PLOS Biology, 15(4), e2001818. doi: 10.1371/journal.pbio.2001818

07. "The BD2K program's charge has been to make large-scale data usage commonplace – streamlining its synthesis, exploration, and its ease of analysis...Indeed, the NIH's investment in BD2K has already been unprecedented, laying the foundation for future advances in precision and participatory medicine. Without the support of multiple NIH institutes, a far-reaching strategy to evolve the "discovery science" of biomedical big data from would not be attainable—"Bui, A., & Van Horn, J. (2017). Envisioning the future of 'big data' biomedicine. Journal Of Biomedical Informatics, 69, 115–117. doi: 10.1016/j.jbi.2017.03.017

08. The flexibility to query an extremely complex patient phenotype from the medical record, when coupled by the powerful methods for representing genotypes as described in this paper, show how the empowered researcher can access the data in an open and powerful query tool—"Murphy, S., Avillach, P., Bellazzi, R., Phillips, L., Gabetta, M., & Eran, A. et al. (2017). Combining clinical and genomics queries using i2b2 – Three methods. PLOS ONE, 12(4), e0172187. doi: 10.1371/journal.pone.0172187

09. "...our final assertion is that automating the ability to reason across integrated data sources and providing users who pose inquiries with a dossier of translated answers coupled with full provenance and confidence in the results is critical if we wish to accelerate clinical and translational insights, drive new discoveries, facilitate serendipity, improve clinical—trial design, and ultimately improve clinical care—"The Biomedical Data Translator Program: Conception, Culture, and Community. (2018). Clinical And Translational Science, 12(2), 91–94. doi: 10.1111/cts.12592

Client Examples

To further the case for a modernized, integrated approach to data governance, Deloitte clients at large public and private healthcare institutions have benefitted from such a modern approach to data management.

Use Case 1: Health Care Provider

- Developed a data strategy to tackle inconsistent data definitions and usage and address lack of policies and procedures for data management
- Created key processes to help enhance and maintain the data warehouse effectively
- Enhanced integrity and accountability of data by adopting a data quality framework for metadata management, change control, issue management, and report sharing
- Launched a data stewardship program to enable data quality capabilities throughout the organization

Use Case 2: Medical System

- Instituted an interoperability data governance effort streamlining the use of existing health standards and procedures to maintain patient data integrity during health information exchanges across Electronic Health Records (EHRs)
- Dictated a centralized common model for health standards in future data at the organizational level, facilitating seamless health data exchange
- Reduced organizational data silos through consistent reasoning and interaction with data

Use Case 3: Biotechnology Company

- Structured a cross-functional team working closely with stakeholders across departments to identify areas of opportunities and improvements
- Instituted repeatable processes for data domains
- Improved data quality with monitoring program
- Standardized definitions and vocabularies
- Established repeatable governance processes expanding across all data domains, globally

Related Resources

- The CDO Playbook, Deloitte Center for Government Insights: https://www2.deloitte.com/insights/us/en/industry/public-sector/chief-data-officer-government-playbook.html
- Deloitte Government and Public Services, Health and Social Care: https://www2.deloitte.com/global/en/pages/public-sector/topics/health-and-social-care.html?icid=nav2_health-and-social-care

About the Authors:

- Dr. Juergen Klenk is a principal at Deloitte Consulting LLP and a mathematician by training. He focuses on advancing Precision Medicine and Data Science in healthcare and biomedical research. He can be reached at jklenk@deloitte.com
- Dina Mikdadi, MPA, is a data scientist at Deloitte Consulting LLP, with a background in public health, analytics, data governance, and clinical research data. She can be reached at dmikdadi@deloitte.com
- Jessica Lo is a systems architect at Deloitte Consulting LLP, with experience doing data management, data analytics, and strategy consulting. She is also a certified Google Professional Cloud Architect. She can be reached at jessilo@deloitte.com
- Amina Jackson, MS, is a bioinformatician, computer scientist, and an expert in cybersecurity at Deloitte Consulting LLP, analyzing and interpreting data for the better understanding of genetics and the molecular mechanism of diseases. She can be reached at amijackson@deloitte.com

Contact us:

Juergen Klenk
PhD - Principal
Deloitte Consulting LLP
Email: jklenk@deloitte.com

Dina Mikdadi
MPA - Data Scientist
Deloitte Consulting LLP
Email: dmikdadi@deloitte.com

Jessica Lo
Systems Architect
Deloitte Consulting LLP
Email: jessilo@deloitte.com

Amina Jackson
MS - Data Engineer
Deloitte Consulting LLP
Email: amijackson@deloitte.com

Deloitte.