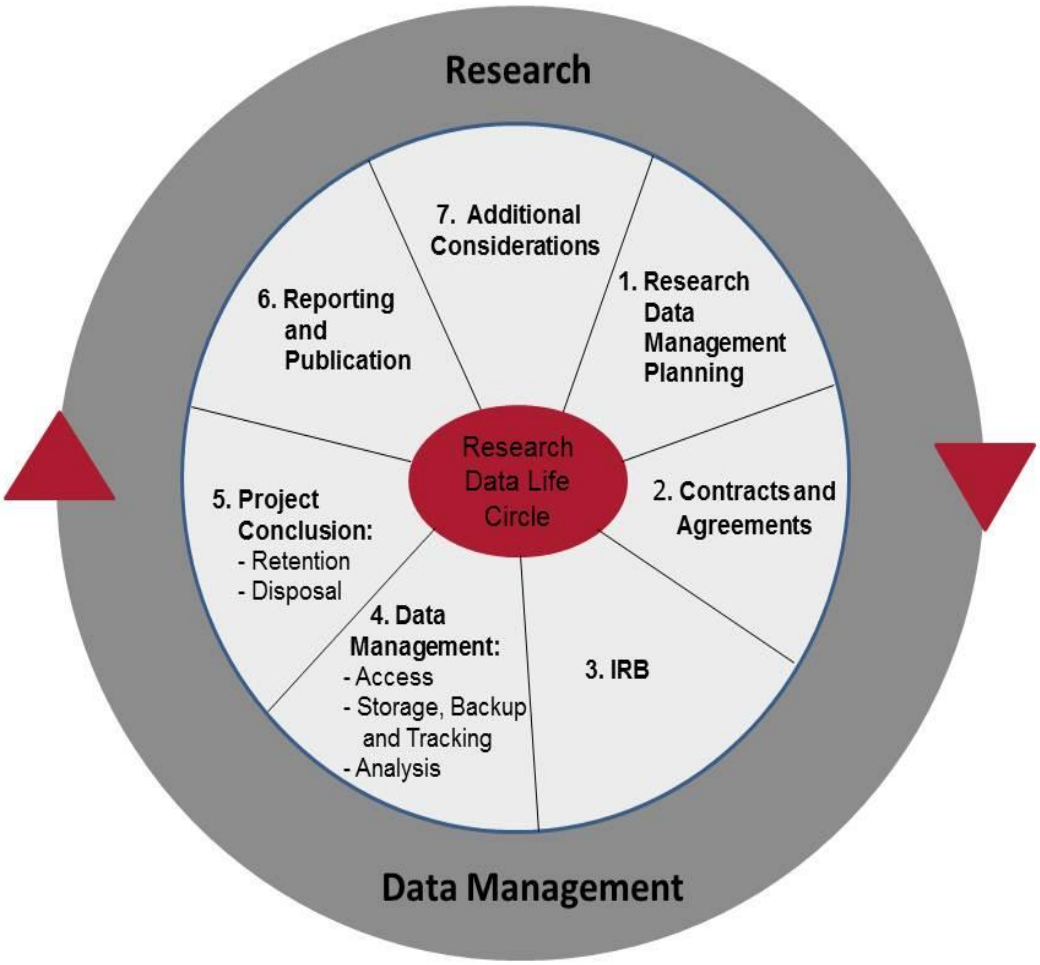
HARVARD CATALYST

THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

An Investigator's Guide to Research Data Management Practices



Adapted from Research Management Lifecycle [Online image] retrieved November 13, 2014 from http://guides.is.uwa.edu.au/content.php?pid=319161&sid=2616069

An Investigator's Guide to Research Data Management Practices

Table of Contents

Introduction

This **Investigator's Guide to Research Data Management Practices** ("Guide") addresses basic data management considerations for researchers who expect to work with confidential or sensitive information involving individuals in the course of a project.[1] Mapped along a Research Data Management Lifecycle, the Guide highlights key issues to consider and offers a roadmap for researchers to follow over the course of data collection, storage and use, dissemination, and final disposition/disposal.

Since data privacy and threats to internet, network, and enterprise security continue to evolve, the Guide attempts to provide a technology-neutral approach toward research data management, and is designed to alert researchers to the various types of issues, considerations, and agreements that are associated with data-use and sharing over the course of a study.

Where appropriate the introduction directs users to other useful resources, many of which are found on the Harvard Catalyst website.[2]

This document does not replace or supersede existing institutional research data management guidance.

[1] This document is not intended to be an exhaustive resource. It does not address research integrity, conflict of interest, institutional policies, quality assurance, or FDA submission criteria: 21 CFR Part 11.

[2] Harvard Catalyst Data Protection: http://catalyst.harvard.edu/programs/regulatory/data-protection.html

Managing Data along the Research Life Cycle

1a. Research Data Management Planning

A data management plan will help you manage and protect your data, meet funder requirements, and help others use your data, if shared. A well-structured project can help protect the confidentiality of patient and participant data. Carefully planned data management also allows for a better use of your time and resources.

Planning for the lifecycle of your research data should begin before data collection or creation. Regardless of whether you are prospectively collecting data as part of your research, or acquiring it from a third-party or another researcher, it is essential to think through all of the ways data will become incorporated into your project, including whether you have plans to make these data accessible to other users in the future. Your plans should help to assure appropriate use and access as well as the privacy, security, and confidentiality of the data. The ways you structure and collect the data have practical, legal and ethical implications.

Following the lifecycle elements, you will be able to develop a research data management plan customized to your research.

The first step in research data management planning is to clearly define the type of data that you will collect or produce. The Harvard Catalyst Data Privacy and Security Planning Checklist[3] is designed to assist in the data planning process by helping researchers describe their data with greater specificity for data protection, IRB submission, and regulatory compliance purposes. Advanced planning also facilitates accurate budgeting; data-intensive projects involving sensitive information may require security measures with costs that exceed routine departmental support, and must therefore be charged to the project. Talk with your grants management or clinical trials offices about how to develop appropriate budgets. The Harvard Catalyst Top Ten Research Data Security Tips offers quick tips, which researchers can implement to enhance research data security.

[3]Data Privacy and Security Planning Checklist:
http://catalyst.harvard.edu/pdf/regulatory/DataPrivacyandSecurityPlanningChecklist.pdf

1b. Elements of a Research Data Management Plan

Many sponsors and data providers will ask you to submit a data management plan (e.g., the National Science Foundation (NSF) Data Management Plan (DMP) requirement).[4] Check with your funder to determine which data management plan elements apply, if any. Your IRB, IT security, and/or contract reviewers may review the description of your data and your management plan.

When crafting a data management plan, consider:

- Any contracts or agreements that my be needed
- Documentation
- Storage and back-up
- Sharing and reuse
- Retention and disposal

Some data require special management considerations. Consult your research data protection contact[5] if any of the following are true:

- The data will be stored on a secure, password protected, server behind a firewall
- The data will be stored on any mobile computing device (laptop, PDA, iPod) or removable media (flash drive, CD/DVD) for any part of your study
- You will be using a cloud vendor, commercial service, or other third party platform for storage, backup, access, analysis, de-identification, re-formatting, or other service; each vendor or commercial solution may have different requirements, encryption, or fees
- Your data will be transferred or transmitted to vendor, contractor or otherwise be manipulated or processed by a third party for linking, correlation, analysis, etc.

To ensure that your research data management plan is successful, consider designating one person who is responsible for making sure the project adheres to the plan, and carefully document any updates and amendments as they occur.

[4] NSF Data Management Plan requirement: http://www.nsf.gov/eng/general/dmp.jsp

[5] Harvard Catalyst Data Protection Website: http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=51

2. Contracts and Agreements
When receiving, sharing, acquiring, or generating data, you may need to enter into various contracts and agreements. You and your study team should be alert to research data terms and conditions contained in documents such as follow:

Agreements that may apply to **confidentiality**:
- Non-Disclosure Agreements (NDA)
- Confidentiality Agreements
- Material Transfer Agreements
- Clinical Trial Agreements
- Certificates of Confidentiality

Agreements that may permit or prevent unspecified **future uses of data or unanticipated secondary uses of data**:
- Data sharing agreements (waivers, consents, etc.)
- Clinical Trial Agreements (waivers, consents, etc.)
- Data Use Agreements
- Business Associate Agreements

Agreements related to **third party sources** such as an outside vendor, external researcher, or government agency. Your institution may have policies about what is appropriate to "put in the cloud," or require additional approvals for collaborations. Applicable agreements may include:
- Service Provider or Data Storage Agreements (i.e. Amazon cloud agreements)
- Cooperative Research and Development Agreements (CRADAs)
- Memoranda of Understanding (MOU)
- Clinical Trial Agreement
- Consortium Agreement
- Data Use Agreements
- Business Associate Agreements

Agreements that may apply to **where and how data is to be retained or archived** include:
- Cooperative Research and Development Agreements (CRADAs)
- Clinical Trial Agreements
- Data Use Agreements
- Business Associate Agreements
- Disposal contracts
- Lease agreements that provide for return of equipment/media containing research information

To navigate the contractual obligations and considerations that may apply to research data use and production, consult the departments in your institution responsible for negotiating such arrangements; these offices might be called technology transfer office,[6] grants and contracts,[7] clinical trials office[8] and/or your IRB.[9] Contact your institution's offices through the Harvard Catalyst Regulatory Atlas.[10]

[6] Technology Transfer Offices: http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=23
[7] Grants and contracts: http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=22
[8] Clinical Trial Office: http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=18

3. Institutional Review Board (IRB) protocol development and review

If your data involves human subjects or identifiable information about human subjects, the project will require IRB review. The regulations define human subjects research as relating to a living individual about whom an investigator obtains:

1) Data, through intervention or interaction with the individual; OR
2) Identifiable private information.

Research data may be identifiable or de-identified. The data collected should include the least possible information that allows you to meet the aims of your study.

Some common data-related elements to address in your IRB protocol include:

- A list of all study staff members and a description of their access to the data
- A list of any collaborators with whom you anticipate sharing data as well as those who will be collecting or reporting data, or have access to personally identifiable information about subjects, including individuals or entities to whom you may transfer data for statistical analysis or de-identification.
- The source of each of the datasets; your IRB may request information about other approvals associated with the collection and access to research data
- If your data is determined to involve human subjects, the subjects' authorization and informed consent will most likely be required; ensure that informed consent documents and authorizations permit the types of data sharing you anticipate (e.g., sharing outside of your institution, publication, or posting in publicly accessible repositories; NOTE: Some forms of research are eligible for waivers of the authorization requirements. Consult your IRB for guidance).

Additional issues to consider:

- If using more than one dataset, consider how linking data impacts identifiability and risk
- If you wish to change the design or conduct of your study after collecting initial data, you must submit modifications to the IRB for review and approval
- If data will be made available through a registry, e.g., dbGaP, or if future open access of data is planned or likely, indicate how data will be released
- Ensure that you understand whether any agreements entered into permit or prevent unspecified future uses of data or unanticipated secondary uses, as these also require IRB review and consideration
- Work with your IRB to determine if a Certificate of Confidentiality is appropriate for your study; this allows the investigator and others with access to research records to refuse to disclose identifying information on individual participants in civil, criminal, administrative, legislative, or other proceedings at the federal, state, or local level; the NIH grants such certificates but they have important limits and are not a "magic bullet."

While the IRB functions independently, it might coordinate with other committees (e.g., with IT for certain levels of data protection); contact your IRB for more information**.**

---

[9] IRB: http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=4

[10] Harvard Catalyst Regulatory Atlas: http://connects.catalyst.harvard.edu/regulatoryatlas/

4. Data Management

Decisions about who will access the data, where the data will reside, and how it will be backed-up may affect data use and access. Consult your institution's research data protection[11] resources and contacts through the Harvard Catalyst Regulatory Atlas.[12]

Access and Collaboration

Be able to describe with specificity who has access to the data and the manner of access. Adhere to the minimum necessary principle (also known as least privileges), meaning only those with a legitimate research, business, or operational need should have access. For example, a research partner who is only reviewing output and co-authoring a paper might not need access.

The following elements may need to be considered:

- The number of people who will collect or work on the data, and whether any are external to your institution
- Whether data must be accessed remotely
- Whether a data sharing agreement imposes restrictions related to the use or sharing of the data
- Whether plans are needed to make the data accessible to other users in the future; if so, what measures will be taken to meet assurances of privacy, security and confidentiality (e.g., if you plan to provide data and images on your website, will the website contain disclaimers, or conditions regarding the use of the data in other publications or products?)

Storage and Backup

The type of access required, and the number of people accessing the data may help determine the manner of storage and the level of security controls. Some data require special management considerations. For example, Protected Health Information (PHI) is subject to several restrictions. You should think carefully about the following:

- How any data sharing will be tracked or documented
- *Where* the data will be stored (In the cloud? Accessed from a secure server? Both? Would computers not connected to network be better?)
- *How* the data will be stored and protected (On a secure, password protected, server behind a firewall (e.g., on a P: drive, not a C: drive)? How protected? Password or encryption? How many people get the password? Who may access?)
- If using mobile computing device (laptop, PDA, iPod) or removable media (flash drive, CD/DVD) for any part of your study, determine how the data containing PHI will be stored
- Whether PHI will be stored by a collaborator or vendor-owned platform (What devices and safeguards will be implemented? How will remote access to the system be secured?)
- Work with your institution to determine if it is appropriate to contract with a third-party vendor to store and back-up data; Identify the various considerations, especially if cloud storage will be adopted across institutions
- A procedure for backing up the data (How secure is the backup system? Who has access? How long are back ups kept?)
- Estimated size of datasets that will be collected and produced, and whether the amount and/or formats of data will change over time; IT departments may need to be informed of anticipated large data sets in order to support back-up.

---

[11] Harvard Catalyst data protection contacts: http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=51

[12] Harvard Catalyst Regulatory Atlas: http://connects.catalyst.harvard.edu/regulatoryatlas/

Analysis and Processing of Data

- Seek out institutional resources to create secure research computing environments if your research involves sharing, electronic data transfers, or multi-site analysis of your data with external collaborators.
- Describe your statistical method and analysis plan, including sample size and its scientific rationale

There may be various means of extraction, transformation, and loading of external data sources that will become a part of your study data. Extracting data from your internal computer/server environment or otherwise transforming the raw data to send to a collaborator at another institution for analysis could significantly change the nature of the data. For example, institution B may require that identifiable data from institution A be de-identified by a reliable third party before being uploaded into institution B's research enterprise environment. Be sure to consider such changes in your management plan.

Tracking

- Describe how any data sharing will be tracked or documented
- Consult your local research compliance resources to determine if local systems can track your data, including audit and monitoring functions; be prepared to explain the file format(s) of your data (i.e. jpeg, .doc, sas, etc.) and consistent naming conventions are key factors in documenting and tracking your research data.

5. Project Conclusion

A project constantly evolves and is often comprised of many related projects. Project close-out is typically thought of as merely an administrative exercise at the end of activity supported by a particular sponsored research award. However, the conclusion of any discrete project should be considered an occasion for important data management housekeeping. The disposition of the data at the end of the project should be documented and become part of the research record.

Retention

Consider what data should be retained at your institution, lab, in archives, or in other local repositories. Institutional library and data specialists may assist in planning or establishing processes for archiving data, including aiding in the selection of formats and media. Familiarize yourself with publication requirements and institutional guidelines for data retention.

Identifiable data should be held for the minimum amount of time necessary to conduct the research (and meet any access requirements). For example, data that is collected in a corporate sponsored clinical trial might have contractual obligations regarding how long the data must be retained. Data collected in federal or state funded projects or when using large health care data sets, may require public access to data and therefore may have specific requirements regarding retention, disposal and archiving. It is essential to understand such requirements and proactively plan so that, at the end of a project, data is properly retained, disposed of, shared, or securely archived.

Be sure to retain data documentation as part of the research record, including:

- Source of data
- Size of data set
- Number of records
- Variables
- Format of data
- Final disposition of data

Disposal: Returning, Destroying, and/or Archiving Data

If data sets were obtained through an agreement with an outside provider, or institution you may be required to return the raw or source data sets at the end of the project, or to destroy the data sets and document that you have done so. Being able to separate a raw data set from an analytical data set is an important parts of project documentation. If an underlying data set contains personally identifiable information about research subjects and a separate de-identified data set has been created, there may be an obligation to destroy the data set containing identifiers.

6. Reporting, Publication, and Public Access
You must take a variety of considerations into account in planning for the reporting and publication of research data, including technically how you plan to report or post the data. Consider whether your sponsor or journal publisher requires public access to the data and ensure that your agreements, consents, and approvals allow for the specific types of sharing, posting, or other secondary uses required. (See Appendix B)

If you are placing your data on a publicly available personal or project website, you may want to consider the impact on individuals if your data is combined with other publicly available sources.

You may also want to consider the impact on individuals if your data were to be obtained through Freedom of Information Act requests.

7. Additional regulatory and legal considerations (See References in Appendix)

Research data may be subject to many kinds of regulations, legal constraints and institutional practices. Your local IRB will be the best authority on which regulations and legal considerations you should be aware of for your particular research project.

a. State Law

Many state privacy laws impact biomedical research. These laws often times apply to health information concerning a specific disease or area of illness deemed to be particularly sensitive, such as mental health, substance abuse, HIV/AIDS, sexually transmitted diseases, mental retardation, and developmental disabilities. Such laws may require that extra precautions be taken to protect the privacy of individuals participating in research that could reveal their status with respect to these diseases, disorders, or conditions.

b. HIPAA

HIPAA applies to HIPAA “covered entities” and HIPAA “business associates”. In rare circumstances, a researcher may be acting as a HIPAA covered entity if he or she is providing health care and conducting certain electronic transactions for which the Department of Health and Humans Services has developed a standard such as payment claims (i.e. billing Medicare). A researcher may also be subject to HIPAA standards if he or she is using or disclosing protected health information (PHI) on behalf of a covered entity. The terms under which such uses or disclosure on behalf of a covered entity are customarily set forth in a business associate agreement.

See Appendix for regulatory references

Appendix

I. HIPAA References

a. **Privacy Rule Requirements** - When the researcher is using PHI protected by HIPAA, then rules in addition to the Common Rule may apply.[7] HIPAA governs uses and disclosures of PHI by a HIPAA "covered entity" which means a health plan, health care providers that electronically transmit data in a HIPAA transaction, and health care clearinghouses. (45 CFR § 160 and subparts A and E of § 164)

b. **Permitted uses and disclosures** - Covered entities can use or disclose PHI for research purposes in the following circumstances (per 45 C.F.R. § 164.512(i)):

- **Authorization**: The researcher obtained specific written authorization from the research participant. (45 CFR 164.508)

- **Preparatory Research**: The researcher asserts that the use or disclosure of PHI is "solely to prepare a research protocol or for similar purposes preparatory to research, that the researcher will not remove any [PHI] from the [CE], and representation that [PHI] for which access is sought is necessary for the research purpose."[8] (45 CFR 164.512(i) (1) (ii) of the Privacy Rule)

- **Documented Approval**: An IRB or Privacy Board approves a waiver of research participants' authorization for use/disclosure of information about them for research. (45 CFR 164.512(i))

- **Research of Decedents' PHI:** The research focuses solely on decedents' information. (45 CFR164.512(i)(1)(iii))

- **Limited Data Set:** The CE and researcher enter into a data use agreement, pursuant to which the CE may disclose only a limited data set to the researcher for research, public health, or health care operations. A limited data set excludes certain direct identifiers of the individual, relatives, employers, and household members. The covered entity providing the data and research must sign a Data Use Agreement that "(1) describes the permitted uses and disclosures of the information and (2) prohibits any attempt to re-identify or contact the individuals." [9] (45 CFR 164.514(e))

- **De-identified**: If PHI is de-identified the health information is no longer PHI or subject to the Privacy Rule. 45 CFR 164.514(a)-(c). A CE can always access, use and disclose for research purposes health information that has been de-identified in accordance with 45 CFR 164.502(d), and 164.514(a)-(c) of the Rule without needing to follow the Privacy Rule. Data can be identified either through (1) stripping certain specified elements from the data, or (2) having an expert determine through statistical analysis that there is a "very small" risk that an individual could be identified based on the data.

c. **HIPAA Waivers -** If your research involves PHI, it may be eligible for a waiver of the requirement of authorization but such waivers must be reviewed by an IRB or Privacy Board. Check with your IRB.

d. **HIPAA Security Rule** - The HIPAA Security Rule (45 CFR Part 160 and 164, subparts A and C) establishes national standards to "protect individuals' electronic personal health information that is created, received, used or maintained by a covered entity."

II. Other Regulatory References:

a. ClinicalTrials.gov: Federal law requires that certain trials (and their results) be registered on clinicaltrials.gov. Determine if you must register your study and find information about how to submit study data here.[13] Additional information is available here.[14]

b. Computerized Systems Used in Clinical Investigations [FDA Non-Binding Guidance] – This guidance on 21 CFR Part 11 compliance provides recommendations to sponsors, contract research organizations, data management centers, clinical investigators and institutional review boards regarding the use of computerized systems in clinical investigations."[15]

c. Genomic Data: If your study involves genotypic or phenotypic data, you should consider whether your data must be submitted to the **d**ata**b**ase of **G**enotypes **a**nd **P**henotypes (dbGaP). Data submission requirements can be found here.[16]

For NIH funded research that generates large scale human or non-human genomic data is subject to the NIH Genomic Data Sharing Policy.[17]

Information about the Genetic Nondiscrimination Information Act (GINA) may apply to research requests for genetic information. Researchers should consider including information about GINA into informed consent documents.[18]

d. Genetic association studies: Federally funded genetic association studies may require

[13] ClinicalTrials.Gov: http://clinicaltrials.gov/ct2/manage-recs/fdaaa

[14] Harvard Catalyst Clinical Trial Registration information http://catalyst.harvard.edu/programs/regulatory/clinical-trial-reg.html

[15] Computerized Systems Used in Clinical Investigations: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070266.pdf

[16] NCBI: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.cgi

[17] NIH Genomic Data Sharing Policy: http://gds.nih.gov/03policy2.html

[18] GINA fact sheet for researchers: http://www.genome.gov/Pages/PolicyEthics/GeneticDiscrimination/GINAInfoDoc.pdf

that data sets be deposited in the GWAS Central[19] repository.

e. Investigational New Drug Applications and Investigational Device Exemptions - These regulations (21 CFR Part 312) and Investigational Device Exemptions [21 CFR Part 812] specify data collection and maintenance requirements when conducting a clinical investigation of products unapproved by the FDA.

f. NAID Requirements: Data collected in federal or state funded projects or when using large health care data sets, may require public access to data and therefore may have specific requirements regarding retention, disposal and archive (e.g.,see the NIAID requirements here).[20]

g. NSF Public Access: For NSF funded research, public access requirements apply (More Information).[21]

h. Publication. For clinical trial data, publications are requiring clinical trial registrations as a condition for publication. (e.g., the International Committee of Medical Journal Editors registration recommendations).[22]

i. Public Access: In February 2013, the Office of Science and Technology Policy issued a broad mandate to the major federal agencies supporting research to develop access plans for all *federally funded scientific research [to be] made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data* (More Information).[23]

j. PubMed: All NIH-funded investigators must submit to PubMed Central an electronic version of their final peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication. (More Information)[24]

k. Return of Results: If your research involves human subjects, your IRB may require that you make summary results of your research available to research participants; be sure to check with your IRB[25] regarding such requirements.

---

[19] GWAS: http://www.gwascentral.org/

[20] NIAID Public Access Requirements:
http://www.niaid.nih.gov/researchfunding/sop/pages/publicaccess.aspx

[21] NSF: http://www.nsf.gov/about/budget/fy2014/pdf/45_fy2014.pdf

[22] ICMJE: http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/

[23] Public Access Requirements for Federal Agency Funded Research:
http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[24] Public Access Resources:
https://www.countway.harvard.edu/menuNavigation/libraryServices/nihPublicAccess.html#steps

[25] Harvard Catalyst Regulatory Atlas:
http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=c&id=5

Attribution, Sharing and Adapting

We encourage you to:

- **request** — email us and request the materials
- **share** — copy, distribute, and transmit the work
- **adapt** — adapt the work to suit your needs

Under the following conditions:

- **Attribution**: In freely using the materials, we require that you acknowledge Harvard Catalyst as the publisher and that you give appropriate credit to any named individual authors.

- **Suggested citation:** *This material is the work the Harvard Catalyst Data Protection Taskforce and subcommittee of the Regulatory Foundations, Ethics, and Law Program. This work was conducted with support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award 8UL1TR000170-05 and financial contributions from Harvard University and its affiliated academic health care centers). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, or the National Institutes of Health.*

With the understanding that:

- **We might contact you**: We are interested in gathering information regarding who is using the material and how they are using it. We may contact you by email to solicit information on how you have used the materials or to request collaboration or input on future activities.
- **When reusing or distributing, make clear the above terms**: For any reuse or distribution, you must make clear to others the terms of this work. The best way to do this is with a link to the web page containing this guide.
- **When adapting**: Please share improvements to the tool back with us so that we may learn and improve our materials as well.

Contact Us

Copies of all materials are freely available. Please send your requests, questions and comments to *regulatory@catalyst.harvard.edu* and visit the Harvard Catalyst Data Protection Subcommittee page here.

CORE WRITING GROUP

Last Name	First Name	Company/ Affiliation
Bolt	Kris	Harvard University
Zurba	Joe	Partners HealthCare
Edmiston	Scott	Harvard Catalyst
Winkler	Sabune	Harvard Catalyst
Bierer	Barbara	Harvard Catalyst

CONTRIBUTORS

Recognizing those that contributed specific content, templates or examples that are included within this guidance document:

Last Name	First Name	Company/ Affiliation
Steve	Berry	Beth Israel Deaconess Medical Center
Jason	Rightmyer	Hebrew Senior Life
David	St. Clair	Boston Children's Hospital
Fernandez-Lynch	Holly	Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics
Cohen	I Glenn	Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics
		Harvard Catalyst Data Protection Committee