|  |  |  |
| --- | --- | --- |
| ***HCSRN Project Policy Best Practices:* Data Sharing for De-identified Data Sets** | | |
| **Accepted by:** | Asset Stewardship Committee | **Last Revision:** 12-26-2013 |
| * [Scope and Purpose](#Scope_Purpose) * [Definitions List](#Definitions) * [Potential Policy Content Areas](#Topics_Sections) * [Identified Best Practices](#Best_Practices) * [Other Data Sharing Resources](#More_HMORNResources) * [Other HCSRN Project Policy Topics](#More_PolicyTopics) * [Submitting Updates and Feedback](#Updates_Feedback) | | |

### Scope and Purpose

Large and complex multicenter HCSRN consortium projects may develop project specific policies as part of their infrastructure and governance. Such guidance documents provide structure, clarification of responsibilities, and pre-defined practices for these projects.

This document is meant to provide general guidance for HCSRN consortium projects that are developing or updating such policies. Best practices from existing HCSRN project policies are synthesized and presented herein. This document and its companion resources are intended to increase efficiency, avoid duplication of effort, encourage consistency and facilitate transparency while ensuring that projects in the HCSRN maintain control over their own policies.

This policy covers the sharing of both general and restricted de-identified research data sets for either primary or secondary analyses, with organizations both inside and outside the HCSRN. Such physically transferred data sets may also include information about the originating health plan(s) (e.g., disease and diagnosis rates, diagnostic counts), as long as the explicit identity of the health plan has been removed. This policy does not cover instances where de-identified data tables are queried remotely, using HCSRNnet or similar query tools, as those instances do not involve dataset transfer.

Sharing of de-identified research data may be either:

• for analytic use by the primary investigative team - when protected health information (PHI) is not needed to support the analyses, or

• in accordance with the NIH Data Sharing Policy for secondary data use, which advises that de-identified research data should be shared when the main outcome paper is accepted for publication.

A goal of many of these multi-site projects is to maximize the sustained value of these data sets by addressing additional important research questions beyond the original project’s specific aims.

De-identified data sets contain no protected health information. They are not limited data sets. When de-identified research data is shared outside of a covered entity, the data recipient does not have access to a crosswalk file or have any other means of re-identifying patients or linking PHI.

It is extremely important to protect privacy, confidentiality, and often anonymity of patients, members, research subjects, clinics, and health systems. Prior to data transfer (physical or electronic), protocols at the data sharing site must be followed to certify the research data set has been properly de-identified, and the risk of re-identification is minimal. The specifics of these protocols are determined by each HCSRN member organization, in the light of applicable federal and state laws.

A Data Use Agreement (DUA) is not required for sharing de-identified research data sets. One condition is that recipients of de-identified data must not attempt to re-identify it, by linking it to identifiable data or PHI. Accordingly, HCSRN member organizations may want to execute a Data Sharing Agreement with recipients of their de-identified data, in order to enforce this requirement, and to further restrict its use.

There are other issues to consider in the development of inter-site data agreements. With regards to project data, the HCSRN has a set of shared principles on: a) When “small cells” constitute a re-identification risk; b) who determines that a small cell issue constitutes a re-identification risk.; and c)when a data use agreement is appropriate. A document capturing these principles is in development and should be published in the near future, with a checklist for those using the “expert determination method” regarding determination of low re-identification risk and the accompanying documentation required by regulations.

The practices and definitions detailed below are intended to ensure the most productive scientific use of the pooled data while ensuring that all legal and regulatory requirements as well as ethical imperatives are met.

**Definitions List**

| **Term** | **Definition** |
| --- | --- |
| **Covered Entity** | An organization to which the HIPAA Privacy Rule applies. Nearly all HCSRN member organizations based in the US are HIPAA Covered entities. Each Kaiser Permanente region that is a member of the HCSRN is considered a separate covered entity for the purposes of this document. |
| **Data Disclosure** | The intentional release, transfer, granting of access to, or divulging of identifiable research data (e.g., limited data set, protected health information) outside the entity holding the information. |
| **Data Ownership** | Each site has ownership over its contributed data. Therefore, it is important to recognize and respect each site’s contributions to creating a pooled analytic data set. |
| **Data Sharing** | The intentional release, transfer, granting of access to, or divulging of research data outside the entity holding the information. |
| **Data Use** | Using, applying, examining, or analyzing protected health information (PHI) or a data set containing PHI or de-identified information. |
| **Data Sharing Agreement (DSA)** | Written agreement between a covered entity and an outside party restricting the use of a de-identified research data set. |
| **Data Use Agreement (DUA)** | A Data Use Agreement (DUA) is required for disclosing limited data sets, as defined by HIPAA standards. Written agreement between a covered entity and an outside party requesting a disclosure of PHI contained in a limited data set. Before a limited data set is released IRB approval must be obtained, and both the DUA and subcontract agreement must be fully executed.  A data use agreement must:  1. establish the permitted uses and disclosure (and may not authorize the data recipient to use or further disclose information that would violate HIPAA privacy requirements for an entity)  2. establish who is permitted to use or receive the limited data set, and  3. promise that the data recipient will:  A. not use or further disclose the information other than as permitted by the agreement or as otherwise required by law,  B. use appropriate safeguards to prevent use or disclosure of the information other than as provided for by the data use agreement,  C. report to data provider any use or disclosure of the information not permitted by the agreement of which the recipient becomes aware,  D. ensure that any agents, including a subcontractor, agree to the same restrictions and conditions that are in the agreement, or  E. not identify the information or contact the individuals. |
| **De-Identified Research Data** | Per the HIPAA Privacy Rule, information that is generated, collected, and/or reported for the purpose of research according to standard research rules, and has been stripped of patient identifiers, including all elements of protected health information (PHI).  De-identified research data is anonymous. Aggregate data (e.g., counts of people, events, etc.) is nearly always de-identified, but not all de-identified data is in aggregate form. The health plan(s) from which it originates may or may not be identified, and if identified should not be explicit or identifiable.  De-identified research data is provided by or on behalf of the covered entity from its patients’ or members’ medical records and other information, including but not limited to health care delivery practices, utilization data, pharmacy data, membership or other clinical or health plan information.  Recipients of de-identified data must not attempt to re-identify it. That is, they may not link it to identifiable data, PHI or otherwise re-identify individuals contributing to the data in any way. |
| **Limited Data Set (LDS)** | A limited set of identifiable patient information, as defined under HIPAA, that may be disclosed to an outside party without a patient’s authorization if: (a) the purpose of the disclosure is only for research, public health or health care operations; and (b) the data recipient signs a data use agreement.  Identifiers that must be removed in order for health information to be a LDS is defined as follows:  Limited Data Set, as defined in the Privacy Rule at 45 CFR Section 164.514(e), is PHI that can include specific identifiers and must exclude others considered to be PHI. A Limited Data Set may include: 1) dates (e.g., admission, discharge, and service dates, dates of birth and death); and 2) five-digit zip codes and state, county, city, and precinct, but not any other postal address information. A limited data set must exclude the following direct identifiers of an individual and his or her relatives, employer(s), and household members: name; postal address information (except town or city, state and zip code which are permitted); telephone numbers; fax numbers; electronic mail addresses; Social Security numbers; medical record numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; license plate numbers and other vehicle identifiers and serial numbers; device identifiers and serial numbers; URLs; Internet Protocol (IP) address numbers; biometric identifiers including finger and voice prints; and full-face photographic and any comparable images. In the event of any conflict between this description and the definition in the Standards for Privacy of Individually Identifiable Health Information (45 CFR, Parts 160 and 164, Subparts A and E) (“the Privacy Rule”), the Privacy Rule definition will govern. |
| **Primary Data Analysis** | Analysis of a research data set created specifically for the given research project. The primary analysis supports the original, stated aims of the research project, in support of developing the main outcome paper(s) for the project. |
| **Protected Health Information (PHI)** | Any information about health status, provision of health care, or payment for health care that can be linked to a specific individual. |
| **Research Data** | Information that is generated, collected, and/or reported for the purpose of research according to standard research rules.  Research data may be identifiable, anonymized, de-identified or aggregate format, provided by or on behalf of the covered entity from its patients’ or members’ medical records and other information, including but not limited to health care delivery practices, utilization data, pharmacy data, membership or other clinical or health plan information. |
| **Secondary Data Analysis / Secondary Data Use** | A reanalysis of data using the same or other appropriate procedures to verify the accuracy of the results of the initial analysis, or for answering different questions in a subsequent and separate study. |
| **Study Period** | The timeframe in which activities relating to a discrete research project occur. The study period ends when the main outcomes paper is accepted for publication. |

**Potential Policy Content Areas**

Most existing data sharing policy and guidance documents for HCSRN studies contain the elements below, though the degree of policy detail varies based on the nature, complexity and needs of the project.

| **Element** | **Considerations & Contextual Details** |
| --- | --- |
| Data consolidation procedures for analyses | * In multi-site studies, for either decentralized or pooled data sets, it is important to consider issues of ownership, duration of possession, adequate funding support for the model chosen, and freedom of sites to contribute or not to primary or secondary analyses. |
| Guidance and considerations relating to site participation | * Policies addressed whether or not PIs and sites who would contribute data could be from outside the consortium (or HCSRN). * Practical considerations were often included as general guidance (e.g., budget caps, staff availability, etc.) * Though final decision making and analytic authority resides with the project PI, consortium leaders may be available to advise on the most appropriate sites to include on a specific proposal, and the appropriate data sharing approach. |
| Longevity of study data sets | * Issues to be addressed include use of data set after funding has ended for the study, prime site IRB approval of novel use of data set at sub-sites with ceding, use of project-specific data sets in subsequent studies on a site-level basis, and data destruction or return of data. |

**Identified Best Practices**

* 1. Sharing data for primary analysis (i.e., within the originating HCSRN study team):
     + The HIPAA Privacy Rule for Research (HIPAA) does not require a Data Use Agreement (DUA) to share de-identified data or if there is no protected health information (PHI) in the data set. Sites may still require a Data Sharing Agreement (DSA) to restrict use.
     + Project teams are encouraged to develop data-sharing plans consistent with the expectations among the members of the multi-site project team. Specifically, project teams need to agree what data will leave the providing sites, when, how and to whom it will be sent, and any limitations on its use by the receiving site, including sharing with secondary parties.
     + During project development, it may be useful to create a flow diagram depicting all data sources and which sites/parties will give or receive data.
  2. Project teams are encouraged to discuss data-sharing issues and approaches during proposal development and plan proactively.
  3. All participating sites in a multi-site project should share data in accord with the research proposal and the agreed upon data-sharing plan.
  4. Discuss and decide among all sites whether the storage of the master data set will be centralized (one site) or a copy at each site.
  5. Discuss and decide whether final analyses on the merged data set will be centralized (one site) or across sites, with a subset of the final data set or with the full final data set (related to #3 above).
  6. Prime sites on a project should include details about data sharing in the scope of work of a subcontract to forestall any disagreements once the project is funded and operational.
  7. Because there is always a risk of re-identification, de-identified research data should be handled by the data recipient in the same manner as identifiable research data. Therefore, some HCSRN sites may require a Data Sharing Agreement (DSA) restricting use of a de-identified data set in a similar manner to a Data Use Agreement (DUA):
     + Specific analyses and secondary sharing of the de-identified data by the data recipient remain consistent with the purpose for which the data was shared.
     + Those to receive and analyze the de-identified data set (including third parties) are named in the agreement and all are held to the same standards and expectations.
     + The data recipient has appropriate safeguards in place to prevent re-identification or linking of the data (e.g., password protection, secured network, access limited to research team, etc.)
     + The data recipient will make no attempt to re-identify the data, link it to identified data sources, etc.
  8. Both limited data sets and de-identified data sets may include a study number assigned for the project as long as it is not a combination of numbers that would allow identification of the individual (such as a scrambled birth date and Social Security Number).

1. Sending aggregate data without identifiers or dates qualifies as a de-identified data set and can be sent to a collaborator without executing a DUA provided the likelihood of an individual being re-identified is small. For example if the number in each cell is significant, the data can be shared with other researchers. Even with very small numbers in a cell, the data may be safe to send if the categories it represents are broad enough, e.g. ages in five- or ten-year groups.
2. Each site providing data should provide an expected timeframe for turnaround of data requests, either as planned per the protocol or for supplemental data requests.
3. Data that are shared for planned use on a specific project must not be used by the receiving site for any other purpose without the PRIOR APPROVAL of the providing site(s).
4. IRB applications should specify the type of data, source, frequency and how the data will be shared among the multi-site project team.

HIPAA regulations strongly encourage conducting research with “minimum data necessary,” though the adequacy of the data must satisfy the scientific needs of the project for high quality data analysis and maximum research productivity.

1. With regards to adverse event dates, de-identified data sets can contain only the year of the date, not the month or day. However, under HIPAA, there are special considerations for reporting adverse events. If the sponsor is the FDA, one may report adverse events without specific agreements. The minimum necessary standard applies. This would count as a disclosure and would need to be tracked.
2. IRB policies may not allow sharing of research data sets of any type after the IRB file for the original study has been closed. Check with your local IRB for guidance.
3. Consider keeping the pooled analytic data set for each study as a resource that, under the appropriate protections and oversight, will be usable by HCSRN investigators and others for new approved ancillary projects. These ancillary projects may go beyond the original study’s specific aims and could occur even after the end of the original project’s study period.
4. Research data in which no identifiable links remain are sometimes shared to support independent investigations by researchers who did not participate in the project creating the de-identified research data set (i.e., a new use of the data - not part of the original project). A Data Sharing Agreement (DSA) to restrict use of the de-identified data set may be required.
5. As a group, decide who will review and approve new data requests, most often for affiliate and ancillary projects (e.g., new proposals committee, steering committee). Agree upon and document composition and logistical issues at the onset.
6. For proposed new analyses or studies, the lead site will request the site PIs send an email affirmation that their data can be used for the proposed new study.
7. When releasing de-identified data sets (e.g., per NIH policy to support secondary analyses), to ensure integrity of the original research data, the primary investigative team [data holder] should provide:
   * + a brief description of the data source(s),
     + abstract describing the original intended use of the data and period of time the data was collected,
     + a data dictionary so that the data user understands and can interpret all data elements accurately,
     + how the original research data should be referenced in subsequent manuscripts written, and any authorship requirements.
8. De-identifying research data should not be a cost burden to the original study. If support for de-identifying data is not budgeted as part of the original project, secondary sharing of de-identified data is not obligatory.

**Other Data Sharing Resources**

The following materials are available to assist project teams developing data sharing policies.

* [HCSRN shared principles](http://www.hcsrn.org/en/Tools%20&%20Materials/GrantsContracting/SmallCells-final.pdf) for de-identified research dataset including small cell sizes
* [Checklist](http://www.hcsrn.org/en/Tools%20&%20Materials/GrantsContracting/Small%20Cells-Expert%20Assessment%20Checklist.doc) for use of the “expert determination method” for documentation of low re-identification risk

**Other HCSRN Project Policy Topics**

Best practices guidance, examples of any existing policies, and comparisons of similarities/differences between existing policies have been developed for these additional topics:

* Publications and Authorship
* New Project Proposals

Find them at [www.hcsrn.org](http://www.hcsrn.org) under Tools & Materials | Managing & Closing Out Projects

**Submitting Updates & Feedback**

This is a dynamic, living document. Additional best practices and resources can be added over time.

To provide an updated project policy, submit a new policy for the repository, share a best practice, ask questions or provide feedback, contact:

Ella Thompson, HCSRN Manager  
email: [thompson.e@ghc.org](mailto:thompson.e@ghc.org)  
phone: 206.442.5211