



**DRAFT**

**Paper 3 - Data Sharing– Advice for Reviewers**

**Prepared by the NCRI Informatics Coordination Unit  
March 2004**

## Why share data?

Member organisations of the National Cancer Research Institute in the UK have agreed to work together to ensure that they maximise the impact of the results of research they fund for the benefit of cancer patients. Their shared vision is to create an internationally compatible informatics platform in the UK that facilitates, access to and integrated analysis of, data generated from research funded by all NCRI Partner organisations, across the spectrum from genomic data to data generated from clinical trials and large-scale epidemiological data.

A key element of this long-term vision is to ensure that data generated through research is put to maximum use by the cancer research community through encouraging data-sharing. The NCRI members believe that all data should be considered for sharing and that data should be made as widely and freely available as possible whilst safeguarding the privacy of participants, and protecting confidential and proprietary data. ***To facilitate data sharing, investigators are required by the NCRI member organisations to include a data-sharing strategy in any cancer-related research grant application in the areas of functional genomics, tissue collection and clinical trials, or state why data sharing is not possible.***

As well as ensuring that they get the most out of their investment in cancer research for the benefits of patients, the NCRI Partners are also responding to a changing research environment in the post-genome age. There is a growing realisation within the scientific community that, if we are to realise the opportunities offered by today's research technologies, we will need to change our approach to data handling and work more collaboratively. There are an increasing number of examples of research groups that already make their data publicly available:

Illinois State Cancer Registry Public Datasets

<http://www.chas.uchicago.edu/healthdata/illinois/idphcancer>

Age-specific incidence of all cancers: by gender and age 1997: Social Trends

<http://www.statistics.gov.uk/statbase/ssdataset.asp?vlnk=5218&B3.x=48&B3.y=8>

The clinical cancer community are also advocating the reporting of clinical datasets linked to experimental results in order to maximise their potential for reuse in other analyses (Brenton & Caldas, 2003). Furthermore, data sharing is already either required or recommended by Funders in other disciplines, e.g. British Academy, Wellcome Trust's History of Medicine Programme, Economic and Social Research Council (ESRC).

Experience from genome research has also highlighted the benefits of data-sharing to

### **The wider community**

- Facilitate communication within the research community
- Encourage the development of new areas of research
- Facilitate further research and debate of current research themes
- Promotes data quality

### **Individual researchers**

- Your collection gains professional recognition within the research, teaching and learning communities
- You will receive recognition and citation for your datasets as they are incorporated into future research
- Preserve your data for future use

### **What sort of data does this apply to?**

The NCRI policy on data sharing applies to:

- The sharing of final research data for research purposes
- To the entire spectrum of translational research to encompass basic research, clinical studies, surveys, and other types of research supported by the NCRI Funders. It applies to research that involves human participants and laboratory research that does not involve human participants. It is especially important to share unique data that cannot be readily replicated.

### **How will this impact on me as a reviewer?**

As a reviewer you will assess the data sharing strategies submitted to your organisation by applicants. Initially, NCRI member organisations will be introducing this requirement in a flexible manner over the coming academic sessions. You should check with the individual funding organisation regarding their particular requirements at this time. Applicants will also be required to report on progress of data sharing at the end of the grant or its renewal.

### **Guidelines for assessing how validity of a data sharing strategy**

The NCRI data-sharing policy applies to all cancer-related research applications. Given the breadth and variety of science that the NCRI Funders support neither the precise content for the data documentation, nor the formatting, presentation, or transport mode for data should be stipulated. Instead there should be a degree of flexibility so that individual Funders can determine the most sensible approach for their organisation. It is sensible to adopt the approach that what is appropriate for one scientific discipline may not work for all others.

#### **1. Method of Data Sharing**

There are many ways to share data. The method for sharing that an investigator selects is likely to depend on several factors, including the sensitivity of the data, the size and complexity of the dataset, and the volume of requests anticipated. Investigators will need to determine which method of data sharing is best for their

particular dataset. The following is an introduction to the different ways in which data can be shared:

- **Under the auspices of the Principal Investigator**

Investigators sharing under their own auspices may simply mail a CD with the data to the requestor, or post the data on their institutional or personal Website. Although not a condition for data access, some investigators sharing under their own auspices may form collaborations with other investigators seeking their data in order to pursue research of mutual interest. Investigators sharing under their own auspices should consider using a **data-sharing agreement** to impose appropriate limitations on users. Such an agreement usually indicates the criteria for data access, whether or not there are any conditions for research use, and can incorporate privacy and confidentiality standards to ensure data security at the recipient site and prohibit manipulation of data for the purposes of identifying subjects. An example of a data sharing agreement can be found in Appendix 1, page 7.

- **Data archive**

You may simply share the data by transferring them to a data archive facility to distribute more widely to interested users, to maintain associated documentation, and to meet reporting requirements. Data archives can be particularly attractive for investigators concerned about a large volume of requests, vetting frivolous or inappropriate requests, or providing technical assistance for users seeking help with analyses. An example of a data archive is the UK Data Archive ([www.data-archive.ac.uk](http://www.data-archive.ac.uk)).

- **Data enclave**

Datasets that cannot be distributed to the general public, for example, because of participant confidentiality concerns, third-party licensing or use agreements that prohibit redistribution, or national security considerations, can be accessed through a data enclave. A data enclave provides a controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.

- **Mixed mode sharing**

Investigators may wish to develop a "mixed mode" for data sharing that allows for more than one version of the dataset and provides different levels of access depending on the version. For example, a redacted dataset could be made available for general use, but stricter controls through a data enclave would be applied if access to more sensitive data were required.

## **2. Examples of Data-Sharing Strategies**

The precise content and level of detail to be included in a data-sharing strategy depends on several factors, such as whether or not the investigator is planning to share data, the size and complexity of the dataset, and the like. Below are several examples of data-sharing strategies:

### **Example 1**

The proposed research will involve a small sample (less than 20 subjects) recruited from clinical facilities in the Manchester area with Williams syndrome. This rare craniofacial disorder is associated with distinguishing facial features, as well as mental retardation. Even with the removal of all identifiers, we believe that it would be difficult if not impossible to protect the identities of subjects given the physical characteristics of subjects, the type of clinical data (including imaging) that we will be collecting, and the relatively restricted area from which we are recruiting subjects. Therefore, we are not planning to share the data.

### **Example 2**

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioural data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

### **Example 3**

This application requests support to collect public-use data from a survey of more than 22,000 UK Citizens over the age of 50 every 2 years. Data products from this study will be made available without cost to researchers and analysts on a website. User registration is required in order to access or download files. As part of the registration process, users must agree to the conditions of use governing access to the public release data, including restrictions against attempting to identify study participants, destruction of the data after analyses are completed, reporting responsibilities, restrictions on redistribution of the data to third parties, and proper acknowledgement of the data resource. Registered users will receive user support, as well as information related to errors in the data, future releases, workshops, and publication lists. The information provided to users will *not* be used for commercial purposes, and will *not* be redistributed to third parties.

Thus, it is important to remember that there may be some data that is too sensitive to be shared (e.g. Example 2). It will be up to you as a reviewer to assess whether or not this is the case in the applications you review or whether, with appropriate measures, the data could be appropriately shared (e.g. Example 3).

### **The application I am reviewing is part of a proposed large multi-centre trial involving international partners – how does this affect data sharing?**

Policies with respect to data sharing vary across countries. Investigators submitting research applications which include collaborators from foreign institutions should familiarise themselves with the policies governing data sharing in the countries that will be involved in the work and address any specific limitations arising from this in the data sharing plan. The involvement of International partners is not a valid reason in itself to justify not sharing the data.

**The application I am reviewing wants to link together existing datasets and will not collect any new data – does this application need to include a data sharing strategy?**

If support is sought from an NCRI member organisation to transform or link existing datasets (as opposed to producing a new set of data), the investigator should still include a data-sharing strategy in the application. If there are limitations associated with a data-sharing agreement for the original dataset that preclude subsequent sharing, then the applicant should explain this in the application.

## Appendix 1 An example of a generic data sharing agreement

---

*This agreement must be signed by anyone seeking to use data in the [name of study](#) maintained by [the Sponsor /Host Institution/Funder](#) before access to such data can be granted. All data is confidential or proprietary except data specified for restricted access public release, or data authorised by [the Sponsor/Host Institution/Funder](#) and the original data source for re-release.*

---

### *Terms and Conditions*

- Any effort to determine the identity of any person contained in the databases (including but not limited to patients, clinicians, and research scientists) or to use the information for any purpose other than for research, analysis, and aggregate statistical reporting would violate the conditions of this data use agreement.
- No identifying information may be published or released in any way without the consent of the person who supplied the information or who can be identified by the information.
- The data provider has omitted from the data set all direct personal identifiers, as well as characteristics that might lead to identification of persons. It may be possible in rare instances, through complex analysis and with outside information, to ascertain from the data sets the identity of particular persons. Considerable harm could ensue if this were done. By virtue of this agreement, the undersigned agrees that such attempts will be prohibited and that information which could identify individuals directly or by inference will not be released or published.
- Users of the data must agree that they will not attempt to contact individuals for the purpose of verifying information supplied in the databases.
- This agreement also restricts the use of any information that allows the identification of establishments to the purpose for which the information was collected.
- Permission was obtained from the data sources (data organisations, hospital Trusts, and data consortia) to use the identification of hospitals (when such identification appears in the data sets) for the purpose of conducting research only. Such research purpose includes linking institutional information from outside data sets for analysis and aggregate statistical reporting. Such purpose does not include the use of information in the data sets concerning individual establishments for commercial or competitive purposes involving those individual establishments, or to determine the rights, benefits, or privileges of establishments.
- Users of the data must not identify establishments directly or by inference in disseminated material. In addition, users of the data must not contact establishments for the purpose of verifying information supplied in the databases.
- Any questions about the data must be referred to the original data source

The undersigned gives the following assurances with respect to the named data sets.

- I will not use nor permit others to use the data in these sets in any way except for research, analysis, and aggregate statistical reporting.
- I will require others in the organisation (specified below) who use the data to sign this agreement (specifically acknowledging their agreement to abide by its terms) and will submit those signed agreements to [the Sponsor/Host Institution/Funder](#)
- I will ensure that the data are kept in a secured environment and that only authorised users have access to the data.
- I will not release nor permit others to release any information that identifies persons, directly or indirectly.
- I will not release nor permit others to release the data sets or any part of them to any person who is not a member of the organisation (specified below), except with the approval of the original data source and [Sponsor/Host Institution/Funder](#).
- I will not attempt to link nor permit others to attempt to link the hospital records of persons in this data set with personally identifiable records from any other source.
- I will not attempt to use nor permit others to use the datasets to learn the identity of any person included in any set.
- I will not use nor permit others to use the data concerning individual establishments (1) for commercial or competitive purposes involving those individual establishments, (2) to determine the rights, benefits, or privileges of individual establishments nor (3) to report, through any medium, data that could identify, directly or by inference, individual establishments.
- When the identities of establishments are not provided on the data sets, I will not attempt to use nor permit others to use the data sets to learn the identity of any establishment in the data sets.
- I will not contact nor permit others to contact establishments or persons in the data sets to question, verify, or discuss data in the databases.
- I will indemnify, defend, and hold harmless the data sources and [Sponsor/Host Institution/Funder](#) from any or all claims and losses accruing to any person, organisation, or other legal entity as a result of violation of this agreement.
- I will make no statement nor permit others to make statements indicating or suggesting that interpretations drawn are those of data sources or the Sponsor.
- I will acknowledge in all reports based on these data that the source of the data is the [insert name of original data source and Sponsor/Host Institution/Funder](#)

I understand that these assurances are collected for the [insert name of Sponsor/Host Institution/Funder](#) to require compliance with its confidentiality requirement. My signature indicates my agreement to comply with the above-stated requirements with the knowledge that this agreement is governed by English Law.



## DEFINITIONS

**Data** - see Final Research Data

**Data Archive** - A place where machine-readable data are acquired, managed, documented, and finally distributed to the scientific community for further analysis.

**Data Enclave** - A controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.

**Final Research Data** - Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based. For the purposes of this policy, final research data do not include laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.

**Restricted Data** - datasets that cannot be distributed to the general public, because of, for example, participant confidentiality concerns, third-party licensing or use agreements, or national security considerations.

**Timeliness** - In general, the NCRI considers the timely release and sharing of data to be as soon as possible after publication of the main findings from the final dataset. However, the actual time will be influenced by the nature of the data collected.

**Unique Data** - Data that cannot be readily replicated. Examples of studies producing unique data include: large surveys that are too expensive to replicate; studies of unique populations, such as centenarians; studies conducted at unique times, such as a natural disaster; studies of rare phenomena, such as rare metabolic diseases.

## Reference

Brenton J, Caldas C. Predictive cancer genomics – what do we need? *The Lancet* 2003 **362** 340-341