UKCTAS Data Management and Sharing Guidelines

Introduction

This document provides guidelines for the management and sharing of data for studies funded, completely or partially, by the UK Centre for Tobacco and Alcohol Studies. These are written in accordance with the agreed UKCTAS Centre Data Management Plan, which will be able to be accessed via the Centre resources library. The aim of these guidelines is to share good practice in data management and to ensure that MRC policy for data management and sharing [1-2] are maintained. The overarching aim of the MRC policy is for data-sharing *to "maximise the life-time value of research data assets for human health* and to do so *timely, responsibly,* with as few restrictions as possible, in a way *consistent with the law, regulation and recognised good practice".*

The UKCTAS guidelines cover the aspects of data management and sharing that should be considered at the design and planning stage of a project taking place within UKCTAS including data documentation, formatting, data storage and back-up, plans for sharing, ethics and consent issues in relation to sharing, copyright and ownership.

Individual UKCTAS researchers (Principal Investigators) are responsible for managing their own data, and for preparing them for data sharing if appropriate, in accordance with these guidelines, and should ensure that a plan for data management and data sharing is prepared for each study, using the form provided in the Centre resources library. **Study protocols and data management plan forms should be submitted to the UKCTAS Data Manager for approval from the Data Management Committee (DMC) prior to commencement of the study.**

Guidelines for data management and sharing

Plans for data management and sharing should be considered during the design and planning stage and detailed in the study data management plan. This should include a description of:

- the data that will be generated during the research, either by collecting new data or through creating new linkages or resources from existing data.
- how the data will be documented & formatted
- the quality assurance measures to be undertaken
- data storage and back-up measures
- data security and anonymisation
- plans for sharing data
- consent and ethical issues in relation to data sharing
- copyright and intellectual property rights of data
- data management roles and responsibilities

Data documentation

Data documentation explains how data were created or digitised, what data mean, what their content and structure are and any data manipulations that may have taken place. Documentation allows researchers within a project or as re-users to make the best use of the data. Projects undertaken within the Centre should aim to create and sustain comprehensive documentation during the research process, explaining how data were created, its content, structure, validation, and any manipulations undertaken. Any additional value added through variable derivation should be recorded and syntax kept as documentation. The study data management plan should provide a list of the documentation that will be created during the project. For example:

Document	**Format**	**Description**
Questionnaires	Pdf	Questionnaires and self-completion questionnaire for pregnant smokers
Project instructions	Pdf	Information provided to the interviewer / participant / coding instructions etc
Dataset information	Pdf	Dataset documentation eg variable lists / derived variables etc
Stata code	Stata do file (converted to text file)	Stata code for deriving variables
Technical information	Pdf	Technical information for example conversions between formats
Study information	Pdf	Background information needed to put the data into context

For some studies, and in particular when data sharing will be through the UK data archive, it may be appropriate to provide **metadata,** that is a subset of core data documentation, which provides standardised structured information explaining the purpose, origin, time references, geographic location, creator, access conditions and terms of use of a data collection [3].

Studies must ensure that the documentation that is produced can support the reasonable understanding and use of study datasets by new and external researchers. Completed documentation should be provided to the Centre Data Manager for inclusion in the Data Inventory within three months of completion of the study.

Data Formatting

Using standard and interchangeable or open lossless data formats ensures long-term usability of data. The following is the file formats currently recommended by the UK Data Archive for long-term preservation of research data [4].

Type of Data	**Recommended file formats for sharing, reuse and preservation**
Quantitative tabular data with extensive metadata, e.g. a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file [5]
Quantitative tabular data with minimal metadata, e.g. a matrix of data with or without column headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
Geospatial data vector and raster data	ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data
Qualitative data textual	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)
Digital image data	TIFF version 6 uncompressed (.tif)
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentation	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

Individual researchers can use the most suitable data formats and software for the planned analyses, but they need to consider converting their research data to standard, interchangeable and longer-lasting formats for back-ups and archive of data once data analysis is completed and data are prepared for storing. The conversion of the data should be documented, including the original data format and the software applied. In addition, when data are converted from one format to another, certain changes may occur to the data. It is the responsibility of the data creator(s)/owner(s) to check for errors or changes of the data after conversions. Details of the formats to be applied should be provided in the study data management plan.

Data quality

The key details of quality assurance will inevitably depend on the institution, design of the study and nature of data collected. Plans for quality assurance should be described in the study data management plan. These will be expected to encompass data collection (for example, standardised methods and protocols will be expected for interviews and questionnaires), data entry or digitisation (for example an appropriate database should be designed for numerical data entry with double entry of data, quality of data transcription [6]), and data checking (for example, quantitative data should be checked for out of range values and data completeness, and transcripts compared with the original recordings).

Data storage and backup

Looking after research data for the longer-term and protecting them from unwanted loss requires having good strategies in place for securely storing and backing-up data. Data should be backed-up either to an institutional back-up server (local/university server) or on a separate data storage device (CD/DVD, external hard drive, etc.) that is kept in a secure and fireproof location, separate from the main data point. For some data it may be appropriate to make multiple back-ups.

Plans for data storage should consider the use of the data in the longer term and that accessibility of any data depends on the storage medium and availability of the relevant data-reading equipment (software) for that particular medium.

Plans for data storage and back-up should be described in the study data management plan and will be expected to encompass:

- Where and in what medium will data be stored and backed-up
- What will be backed-up and how often
- How will the data be stored and in what format for longer-term software readability.

Data security and confidentiality of potentially disclosive personal information

Physical security, network security and security of computer systems and files all need to be considered to ensure security of data, and prevent unauthorised access, changes to data, disclosure or destruction of all project data. Data security arrangements will need to be proportionate to the nature of the data and the risks involved. Data that contain

personal information should be treated with higher levels of security than data which do not. Where personal data is involved, data security should be handled and stored in line with the Data Protection Act 1998 and with MRC policy. Plans for data security and participant confidentiality should be described in the study data management plan, and will be dependent on sensitivity of data collected but will be expected to encompass:

- Plans for physical data security (physical access to hard or electronic copies of data)
- Network security (e.g. not storing confidential data such as those containing personal information on servers or computers connected to an external network, firewall protection and security-related upgrades)
- Security of computer systems and files (may include locking computer systems with a password and installing a firewall system, implementing password protection to data files, encryption, transmission of only encrypted data)
- Security of personal information (anonymisation will be expected to be planned at an early stage and storage of personal information separately from data files, encryption for all personal information, and an anonymisation log created of replacements & removals).

Data sharing

The MRC policy on data sharing applies to all MRC funded research and hence all studies funded by the UKCTAS will need to be familiar with, and carefully consider, the MRC data sharing requirements [1-2]. These requirements are mandatory for MRC-funded cohort studies but the MRC states 'are likely to be applicable to MRC population and patient-based research more broadly' including much of the work of the Centre.
To meet the MRC requirements each study protocol will need:

- A statement of study policy on data-sharing which is consistent with MRC's overarching policy on data-sharing and preservation and sensitive to the interests of participants.
- Transparent and clearly justified priorities and criteria for sharing and access, and the various constraints on these.
- A clearly defined and justified statement on the type and extent of privileged use by the study team.

There may be specific issues with individual projects, for example care may be needed to ensure that those with completing interests – most notably the beverage alcohol and tobacco industries – did not get access.

Subject to these provisos, studies funded by the Centre should aim **to offer all data collected to the UK data archive, or otherwise make it available for sharing, within three months of final publications**, in accordance with the UKCTAS data management plan. We will need to report to the MRC on the performance and outputs of sharing achieved during the Centres funding period.

Data should be provided with high quality documentation for secondary users, which will be checked by the Centre Data Manager. A template for a data sharing agreement can

be accessed via the Centre resources library. The Centre data manager will support each research team to produce a data sharing agreement suitable for the individual study. Data sharing agreements should be signed before data are released and these will prohibit any attempt to identify study participants or otherwise breach confidentiality.

Ethics and Consent

Researchers are usually expected to obtain informed consent for people to participate in research and for use of the information collected. Where possible, consent and ethical approvals should also take into account any future uses of the data, such as the sharing, preservation and long-term use of the data. Further advice on how that might be achieved is provided in the UK data archive managing and sharing data guidelines and in the MRC policy and guidelines on sharing of research data [1,2,7].

Copyright and intellectual property rights of data

Intellectual Property Rights for all data collected by the Centre will rest with the organisation carrying out the research, i.e. the institution at which PI of the research is based, or based on the agreement between institutions, which needs to be detailed in the study protocol. The PI should ensure that IP relating to the value they create is suitably protected and managed in line with RCUK Knowledge Exchange Principles. Any delays or restrictions on sharing due to managing IP should be minimised as far as possible.

Where a study uses existing datasets, these remain the property of their respective owners, and the owner should be sufficiently acknowledged in any resulting publication. Where new data is derived or created from them, consideration should be given to whether permission could be sought form the owner to share these data. In the case of the commercial and licensed primary care data, whilst it may not be possible to share these data due to commercial sensitivities, it may be possible to obtain agreement to share the data in some form, for example, previously agreement was obtained to show (graphically) derived key aggregate monthly indicators of smoking behaviour from The Health Improvement Network (THIN) dataset, and NRT sales on the UKCTCS website.
A statement on copyright and intellectual property rights of data should be included in the study data management plan.

Data Management Roles and Responsibilities

The responsibility for data management for each study within the Centre lies with the principal investigator for the study, but responsibilities may be devolved to others involved with the project; a statement on data management roles and responsibilities should be provided in the study data management plan.

References

1. MRC policy on research data-sharing, http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm
2. MRC Policy and Guidance on Sharing of Research Data from Population and Patient Studies, Nov 2011, http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC008302
3. Documentation & Metadata Overview, http://data-archive.ac.uk/create-manage/document/overview
4. UK Data Archive, File formats table, http://data-archive.ac.uk/create-manage/format/formats-table
5. UK Data Archive, Catalogue Metadata, http://data-archive.ac.uk/create-manage/document/metadata
6. UK Data Archive, Transcription, http://data-archive.ac.uk/create-manage/format/transcription
7. UK Data Archive, Managing and Sharing Data- Best Practice for Researchers, May 2011, http://data-archive.ac.uk/media/2894/managingsharing.pdf