# TEMPLATE FOR A DATA MANAGEMENT PLAN

**0. Proposal name**

Pilot study of the utility of machine learning tools to accelerate systematic review and meta-analysis of findings of in vivo research

**1. Description of the data**

**1.1 Type of study**

This study will assess the utility of machine learning and text mining approaches to identifying and extracting information from published scientific works

**1.2 Types of data**

Background data comprise (1) text retrieved from literature search engines (usually .txt or .xml); (2) Full text of publications retrieved from PubMed Central, Institutional Repositories and publishers (usually .pdf or .html); and (3) categorical or numerical data extracted by human investigators from those full text publications. Foreground data comprise (1) computer code used to perform the tasks outlined in the main application (identification and retrieval of relevant publications, extraction of publication meta-data and extraction of publication outcome data); and (2) quantitative data generated from comparison of different approaches to the tasks outlines above.

**1.3 Format and scale of the data**

Foreground computer code will be developed in compliance with the Unstructured Information Management Architecture, an Apache Software Foundation open source project commonly used for the processing of language resources. All other data are in standard formats (.txt, .xml, .html, .pdf, .csv). We estimate around 50,000 unique publication records will contribute to this project, of which ~5,000 will be retrieved in full text, from which ~20 categorical variables and ~30 numerical variables each will be extracted. Human data extraction has been or will be performed in duplicate. All formats used enable sharing and long-term re-use of data?

**2. Data collection / generation**

**2.1 Methodologies for data collection / generation**

Current gold standard for data collection are given in Sena et al (Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically, Journal of Cerebral Blood Flow & Metabolism (2014) 34, 737–742; doi:10.1038/jcbfm.2014.28), and we will adhere to this standard.

**2.2 Data quality and standards**

Consistency and quality of data collection and generation will be controlled by having two investigators perform each task. The performance of the investigators will be confirmed one with the other using the kappa statistic, and the performance of developed machine learning/ text mining tools will be assessed as the sensitivity and specificity against the combined human "gold standard".

**3. Data management, documentation and curation**

**3.1 Managing, storing and curating data.**

Data will be stored on servers at the University of Edinburgh. These servers are backed up daily, with 30 days of backup held.

**3.2 Metadata standards and data documentation**

All data arising from the study, including computer code, will be made available on open access servers such at GitHub and FigShare. These depositions will include detailed annotation to allow reuse including documentation of the methods used to generate the data, analytical and procedural information, and detailed descriptions for variables.

### 3.3 Data preservation strategy and standards

Deposition in GitHub and Figshare will ensure long-term storage and preservation with no limit to the retention period.

## 4. Data security and confidentiality of potentially disclosive information

### 4.1 Formal information/data security standards

University of Edinburgh information and data security policy and standards are described at http://www.ed.ac.uk/polopoly_fs/1.162655!/fileManager/Information%20Security%20Policy_vsn_2.1.pdf.

### 4.2 Main risks to data security

No personal data will contribute to this study, and all data collected will be derived from published scientific works. Confidentiality is therefore not an issue. Access to the CAMARADES database is controlled through a password protected login, and an audit trial identifies those who have logged into the database, at what time and for how long. The risk of "information vandalism", for instance by animal rights activists, is we believe small (and has not occurred in the 10 years that the database has been in existence). However, to address this the next iteration of the database will include, for each field, an audit trial of entries and edits such that once identified any such vandalism can be easily reversed.

## 5. Data sharing and access

We use GitHub and Figshare to store, curate and share data from our studies.

### 5.1 Suitability for sharing

Our data is indeed suitable for sharing, as it can be used by others for analyses secondary to our purpose at the time of data collection. For instance, see Hirst et al (The Need for Randomization in Animal Trials: An Overview of Systematic Reviews, PLoS One DOI: 10.1371/journal.pone.0098856).

### 5.2 Discovery by potential users of the research data

New users can find our data

(a) through data descriptor publications (for instance the journal "Evidence Based Preclinical Medicine" (MRM is co-Editor in Chief) has data descriptor as a type of publication, and the Egan Alzheimers Disease database is under consideration at present, as the same time as the analysis paper is under consideration at a different journal.

(b) By direction from the CAMARADES website

(c) By keyword search of FigShare or Github using the keywords "systematic review" and "animal".

Our approach to data sharing is published on our study website at http://www.dcn.ed.ac.uk/camarades/default.htm#about.

### 5.3 Governance of access

Because all information is available without request, no decision on supply is required.

### 5.4 The study team's exclusive use of the data

All data are made available at completion of the study. Studies which for whatever reason are not accepted for publication would be made available in FigShare and on the CAMARADES website. The principle is that all data should be made available as soon as the dataset is complete.

### 5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

There are no restrictions to data sharing.

**5.6 Regulation of responsibilities of users**

We impose no responsibilities on external users.

**6. Responsibilities**

As the PI and Scientific Co-ordinator for the SLiM consortium, Macleod is responsible for study-wide data management, metadata creation, data security, and quality assurance of data.

**7. Relevant institutional, departmental or study policies on data sharing and data security**

*Please complete, where such policies are (i) relevant to your study, and (ii) are in the public domain, e.g. accessible through the internet.*

*Add any others that are relevant*

| **Policy** | **URL or Reference** |
|---|---|
| Data Management Policy & Procedures | This document is available online at www.camarades.info |
| Data Security Policy | This document is available online at www.camarades.info |
| Data Sharing Policy | This document is available online at www.camarades.info |
| Institutional Information Policy | http://www.ed.ac.uk/polopoly_fs/1.162655!/fileManager/Information%20Security%20Policy_vsn_2.1.pdf. |
| Other: | |
| Other | |

**8. Author of this Data Management Plan (Name)** and, if different to that of the Principal Investigator, their **telephone & email contact details**

Malcolm Macleod