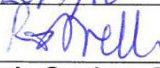
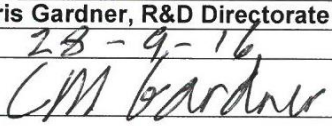
STANDARD OPERATING PROCEDURE

DATA CAPTURE AND CLEANSING SOP
S64

Version:	**1.1**
Effective Date:	**22 September 2016**
Review Date:	**22 September 2018**
Author & Position:	**Dr Roy Powell, Statistical Advisor**
Date:	23/9/16
Signature:	
Approver & Position:	**Chris Gardner, R&D Directorate Manager**
Date:	23-9-16
Signature:	

Controlled document

This document has been created following the Royal Devon and Exeter NHS Foundation Trust Policies, Procedures, Protocols, Guidelines and Standards Policy. It should not be altered in any way without the express permission of the author or their representative.

IT IS THE RESPONSIBILITY OF ALL USERS OF THIS SOP TO ENSURE THAT THE CORRECT VERSION IS BEING USED. If you are reading this in a paper format please go on-line to confirm you have the latest version.
https://hub.exe.nhs.uk/a-z/research-and-development/research-and-development-documents/?opentab=2

DISCLAIMER

This generic R&D Standard Operating Procedure (SOP) must be followed unless;

- A study specific SOP exists
- A departmental SOP dictates a different working practice

Once printed this is an uncontrolled document.

VERSION HISTORY LOG

This area should detail the version history for this document. It should detail the key elements of the changes to the versions.

VERSION	Reviewer	Date Implemented	Details of significant changes
1.0	Rohan Chauhan, NHS Research Advisor	14 November 2014	
1.1	Lisella Wilkinson Quality Assurance Coordinator, R&D	22 September 2016	Removal of Copyright symbol Updated link to online SOPs on the new Hub Intranet, live from August 2016

CONTENTS

1 BACKGROUND

Data collected in clinical trials and other well-designed research studies needs to reflect what truly occurred in the study in terms of the outcomes recorded. Quantitative (numerical and coded) data and some text are often initially recorded in the field by research nurses or other research assistants and technicians using either paper forms and questionnaires or an electronic device such as a laptop computer or smart phone. Other types of data that might be recorded could involve audio or video recording equipment (such as a digital Dictaphone, still or video camera). The integrity of that type of data is beyond the scope of this document. In the collection of data, there is plenty of scope for errors to be made – especially in the transference or transcription of data from the original data collection form or file to some other type of document – often a spreadsheet file on a computer.

2 PURPOSE

In order to ensure that data collected during research projects is accurate and valid.

3 SCOPE

To be implemented in all non-commercial research studies.

4 RESPONSIBILITIES

It is the responsibility of the Chief Investigator to ensure that data has been captured accurately and cleaned before any analysis takes place, either by the CI themselves or by an independent statistician.

5 PROCEDURES

The procedures are as follows.

Data capture

There must be detailed discussion between the CI and the statistician about the data to be captured in a trial. What is being measured and how is it being recorded? What are the units of measure in each case? The data should be classified according to its type i.e. nominal, ordinal or scale measurement. If these are defined at this stage, this makes the job of preparation of the data for analysis much quicker and more efficient.

Bespoke database

A database or spreadsheet must be designed to hold the eventual data. In some cases, especially if a Clinical Trial Unit is employed, a bespoke database to capture the data can be designed by the specialist IT staff of the CTU. This may be a Microsoft Access database, for example. If the clinical trial involves research nurses working in the community, they can be supplied with smart phones, tablets or laptops which can link with the database and data can be entered on the device for each recruit in the trial and transmitted to the database. Other facilities such as web-based randomisation can also be available to the research nurses this way. Each time a participant is recruited, the database can be set up to alert the CI of any new recruits and also any withdrawals, outcomes or adverse events.

Paper CRFs

Some feasibility and pilot trials do not have the funding to afford the services of the local CTU. In this case, the data from the clinical trial should still be entered into a dedicated database if the skills exist within the trial team to build one e.g. using Microsoft Access. This can then be used on a dedicated research computer if the trial is taking place in a static clinic or on an encrypted laptop or tablet if the research nurse is working away from the base. If these facilities are not available then the researchers will resort to using data collection of paper CRFs. The format of the form will need to be agreed between the team and the statistician in advance. This is to ensure that the various data items are recorded in a way that can later be understood and interpreted by people using the CRFs. This will include transcribers and the statistician. If data is captured using an Access database, it will need to be piloted to ensure that it stores the data as expected and can export the data as a flat-file spreadsheet for analysis in the expected way.

Data security

Clearly data should be stored securely so that only the members of the research team have access to it and nobody outside that team can see it, copy it, edit it or tamper with it in any way.

Bespoke database

In the case of a dedicated database, this should be password protected and stored encrypted on a hard drive of a computer and backed up on the host organisation's secure server. The computer will be password protected and stored in a lockable office to be used only by the research team. Data should be stored as per National Data Protection Guidelines. The data should be stored in such a way that it is pseudo-anonymised i.e. that each participant is given a trial number and no personal identifiers are stored in the data file. A separate list of trial numbers can link the data to participants. This must be stored in a separate secure place but must be readily available so that un-blinding can take place rapidly if this is required during the course of the trial.

Paper CRFs

If data is captured first on paper CRFs, these must be filed securely in a locked filing cabinet which is situated in a lockable office and be only accessed by the research team.

Data transfer into a spreadsheet for analysis

Entering data in an Excel spreadsheet in the correct way can significantly reduce the amount of preparation time and cleaning required before undertaking statistical analyses. This can save time for both the investigator and the statistician if data can be transferred directly from Excel to a standard statistics software package such as SPSS, STATA, R, StatsDirect or MINITAB etc. If data are entered into Excel as suggested below, the data should be immediately ready for analysis using most statistical software including the basic options available in Excel.

All personal identifying data about participants must be kept separately from data representing the results of studies. An independent Excel file or worksheet for personal data should be set up which is kept in a secure system. In this, each individual participant is given a unique identifying code which can be used in the Excel results and statistics spreadsheets.

All data required for analysis should be entered into a single worksheet – not different sheets within a workbook – as many statistics packages can only import one sheet (a flat file).

1. Codes or identifiers of no more than eight letters representing each study variable should be entered across the first row of the spreadsheet starting with a "Unique ID" column. Mathematical characters such as + - % ! < > etc. should not be used as these may not be accepted by the statistics software; underscores (_) are usually acceptable in variable names for most statistical packages.

2. Codes representing individuals should be entered in the first column starting at the second row.

3. The code or name given to each individual or study variable should be unique.

4. The data should start in the second row.

5. A separate row should be given for each patient or participant and each person should be given only one row in the worksheet. If you have a before and after measurement for an individual, then these should be entered in separate columns identified under row 1 headings rather than separate rows.

6. There should be no blank rows in the worksheet.

7. Avoid putting blank rows between patients receiving different treatments (or even putting patients receiving different treatments in different worksheets). Instead create a separate column to flag which treatment the patient is receiving, coded perhaps 1, 2 etc.).

8. All numeric data should be numeric! Characters such as '?' should not be entered when a value is not known, or '<4' for example, as statistics software cannot make sense of these characters. If you are still waiting for results to be entered, leave the space blank until you are able to fill in the data (see below for handling missing data).

9. The data in each column of an Excel spreadsheet should be formatted appropriately (by right-clicking on the column header) – otherwise it will get imported into a statistical package in the default alphanumeric text format ('string' format) – which the package can't handle. Columns of integers and scale data need to be formatted as 'number' data in Excel, with the desired number of decimal places. Columns containing dates need to be formatted as 'date' fields – otherwise they get imported as long numerical strings which cannot be interpreted. There are also 'time' and 'currency' formats, for example.

10. All text data should be coded into numeric data where possible. For example, for gender, code male as 1 and female as 2. A separate note of what the codes mean should be kept.

11. Dates should be entered as four figure dates (for example 1927, rather than 27) to avoid confusion when changing between packages and versions of packages. This is particularly important if you have date of birth information and you are dealing with older participants, who were born in the previous century.

12. Summary statistics (means, medians etc.) or any other notes should not be entered or calculated at the bottom of the worksheet. This causes confusion if the data needs to be put into a statistical package.

13. Missing data should be coded using numbers that don't normally occur in the range for a particular variable e.g. -1, 99 or 999. A zero (0) could be a valid data value, depending on the variable. In SPSS, a blank cell is assumed to be a missing value but other missing values can be defined e.g. -1, 99 or 999 so that they are not included in any analyses.

14. If in doubt about how to enter your data into Excel, then seek advice from R&D or CTU staff.

Examples of data entered into Excel correctly and incorrectly are shown in the tables below.

Note that:

1. The patient's gender has been coded into a numeric variable.

2. Patients are receiving treatments 1 and 2. A column and hence a separate variable (namely the variable ‘group’) identifies who is receiving treatment 1 and who is receiving treatment 2. Blank rows have not been left between patients receiving different treatments.

3. All dates have been entered with 4 digits for the year (although Excel only shows 2 until the cell is clicked on).

Table 1. Not recommended:-

Hospital no	group	sex	date-of-birth	date-F/Up	BP
RDE9876	**Rx**	F	26/08/37	07/06/13	134/78
RDE9753	Rx	M	23/07/12	09/08/13	132/60
RDE8642	Con	female	18/02/10	26/04/13	138/64
RDE6542	Rx	M	31st Jan 55	02/12/13	110/66
RDE1098	Treat	m	22/02/49	24/02/13	104/64
RDE7956	Ctrl	f	02/04/46	16/11/13	114/68
RDE2223	CTRL	Male	05/04/30	17/06/13	164/88

Table 2. Recommended spreadsheet setup:

id	group	sex	dob	date	systolic	diastol
1001	1	2	26/08/1937	07/06/2013	134	78
1002	1	1	23/07/1912	09/08/2013	132	60
1003	2	2	18/02/1910	26/04/2013	138	64
1004	1	1	31/01/1955	02/12/2013	110	66
1005	1	1	22/02/1949	24/02/2013	104	64
1006	2	2	02/04/1946	16/11/2013	114	68
1007	2	1	05/04/1930	17/06/2013	164	88

Transfer of data from a bespoke database

Databases written in software such as Microsoft Access will need to be able to export the data into a flat file spreadsheet in Excel format. This is the standard format that most statistical packages can use for future analysis. The nature of the data should be checked according to the guidelines above to ensure that the intended statistical package will be able to import and use the data.

Transfer of data from CRFs

There are really only two other choices for data entry: by manual (keyboard) entry or by the use of optical character recognition software such as FORMIC, Teleform or a bespoke program written for example by IT experts in a CTU for example.

Double manual data entry

Double data entry means that two operators manually enter the data into their own flat-file Excel spreadsheet from the same CRFs independently. At the end, there should be two (virtually) identical spreadsheets containing the same data. This has been the gold standard for a long time. The two spreadsheets should then be compared for any discrepant data and any found should be checked against the source data.

Dedicated form scanning: software such as FORMIC or Teleform

With software such as these, the CRF can be designed in a type of desk-top publishing system which designs a database to hold the data in the background. The resulting form will be scannable and will often contain codes (often bar codes) to identify for example the particular trial or even the participant number. If the forms do not have unique codes printed on them, they can be photocopied and the codes written on them by hand. The forms will contain grids and tick-boxes for various types of data and also spaces for free-hand text. Capital letters and numbers written in ink can be read by the software. Scales such as visual analogue scales can also be inserted in to forms. These are all read by the software when the completed forms are scanned by a dedicated double-sided scanner. The data are fed into the dedicated database. Free-hand text is not read by the software because of the huge variation in free hand writing. However, a picture of the text box is shown on the screen and the operator is able to type what is written within it and this text is entered into the results database.

Assessing reliability of the data

Double manually entered data

With double entered data, we end up with two spreadsheets that are hopefully identical. Human error will creep in though and in order to minimise this, the two spreadsheets should be compared using a specific Excel macro. This compares the two sheets cell by cell and highlights any cells that are discrepant. The operator can then look back at the original paper CRFs to read what was intended for that cell and enter it correctly.

Data from dedicated form scanning software

Data scanned by machine and processed by optical character recognition software such as FORMIC or Teleform is said by the software companies to be 99% accurate. There are tolerances built into the scanning so that if there is any doubt about a hand-written alphanumeric character that cannot be interpreted by the system, this is displayed to the operator for validation. The operator can either read the character and confirm it from the keyboard or go back to the original paper CRF to see what was written and enter it correctly manually.

6 FURTHER READING

Dawn-Marie Walker (Editor). *An introduction to Health Services Research*. 2014. Sage. 362pp.

Shields BM, Knight BA, Turner M, Wilkins-Wall B, McCammon K, Round A, Powell R, Hattersley AT. Improving the quality of maternity data – lessons from the Exeter Family Study of Childhood Health. *Midwives* 2004; 7(4):156-159.