

Data Management

Getting started

Planning is a critical step in Data Management and this should commence before the Protocol has been finalised. See the steps below for getting started with Data Management, the link in the sidebar will take you to more detailed sections on the topics covered.

Requirements and Guidelines

Make sure you are familiar with the regulations associated with your data and any funder and/or institutional requirements with regard to data management, they may have policies or guidance stipulating how data should be monitored, shared, archived and timeframes for these activities.

Protocol Development

'The Clinical Data Management (CDM) process, like a clinical trial, begins with the end in mind. This means that the whole process is designed keeping the deliverable in view. As a clinical trial is designed to answer the research question, the CDM process is designed to deliver an error-free, valid, and statistically sound database. To meet this objective, the CDM process starts early, even before the finalisation of the study protocol.' [Data management in clinical research: An overview](#)

Good data management requires proper planning and should begin in parallel with protocol development to ensure that all of the protocol-specified data is accurately captured. Plans for how assessments will be performed, what and how data will be collected, entered, coded, stored, protected, analysed and quality controlled all need to be captured in the protocol or separately with a reference to where this information can be located. The **SPiRiT Checklist** provides a list of recommended Items to address in a clinical trial protocol several of which are focused on data management.

For further information on developing your protocol see node 4 on the [Process Map](#) and the [Protocol Development Toolkit](#).

Data Management Plan

The **Data Management Plan (DMP)** is a very important piece of study documentation and should be included as annex to the protocol. Depending on the study the plan may be made up of several documents. It should give a complete picture of how the data will be handled throughout the study by outlining all of the information relating to the study's data management procedures.

The elements you need to consider when writing a DMP are covered in the [Data Management Plan](#) tile.

Costs

The plans you make with regard to data management are likely to affect the cost of your research proposal. You will need to consider the long-term requirements for physical storage of the data in addition to the cost of data collection.

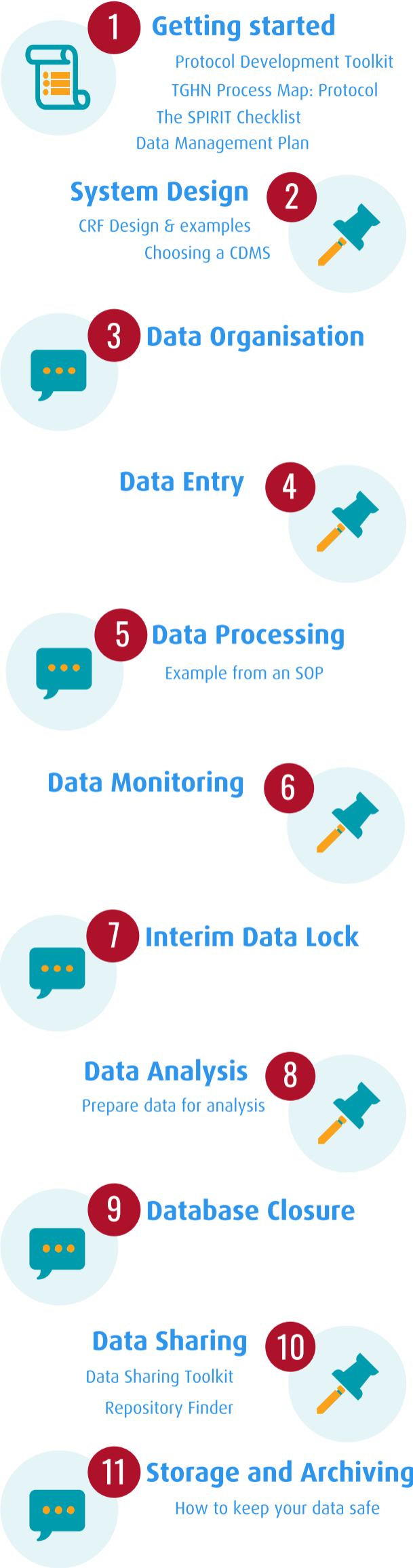
Prepare to share

It is now a standard requirement by publishers, funders, research institutions and regulatory agencies to share data.

You will find 'Prepare to share' links throughout the Data Management Portal with tips on how to prepare your data so it will be findable, accessible, interoperable and reusable. The [Data Sharing Tile](#) will give you further information and the [Data Sharing Steps](#) will take you step by step through how to share your data.

Training and Resources

- TGHN Global Health Data Management
- TGHN Data Management Training
- TGHN Data Management Tools and Templates



1

Getting started: Data Management Plan

Data Management

Data Management Plan (DMP)

Why plan?

A DMP will help you to work through how to manage your data. Which data will you collect? How will you process and analyse the data? You will need to describe methods that you plan to use for these activities and document them using a DMP. In addition to the benefits to your study many funders now ask grant applicants to submit a DMP document.

‘Planning for the effective creation, management and sharing of your data enables you to get the most out of your research.’ [Digital Curation Centre: How to Develop a Data Management and Sharing Plan.](#)

You should plan to revisit the Data Management Plan throughout the project to make any necessary adjustments. Think of this document as something that is evolving and flexible, rather than fixed.

Common headings in Data Management Plan templates:

Click a heading to see associated questions to consider for your Data Management Plan

- Background and methodology
- Documentation and Metadata
- Ethics and Legal Compliance
- Data Storage and Back-up
- Long-term Preservation of Data
- Data Sharing
- Responsibilities and Resources

Getting started

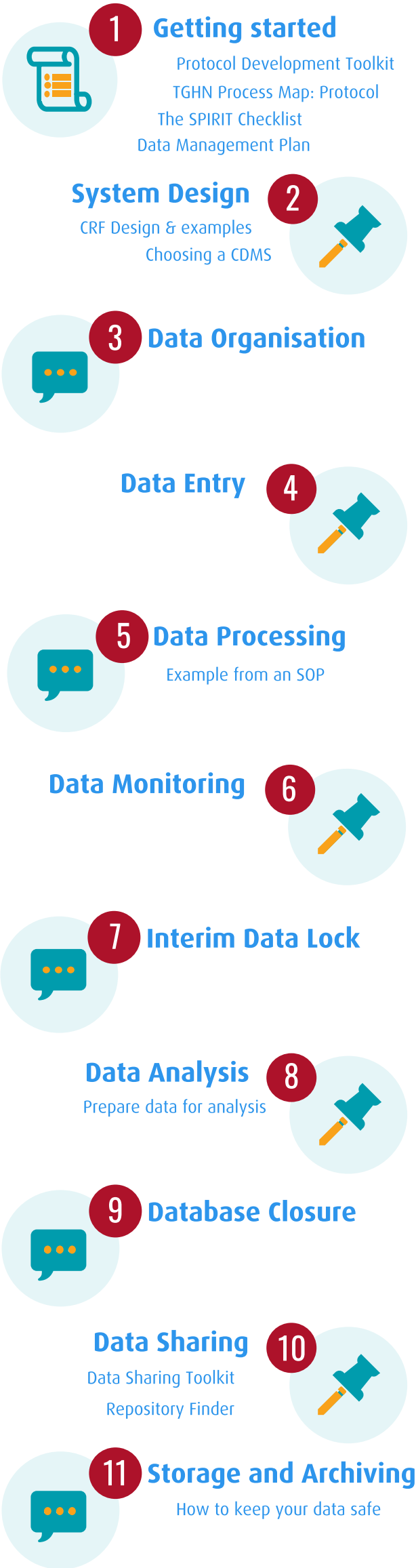
The best way to get started writing a data management plan (DMP) is to begin with a suitable template. Does your funder require a DMP? If so is a template or guidance on the content required provided?

Data Management Plan templates from 12 major institutions and funders were compared to produce a list of questions that should be considered when creating a Data Management Plan. Select the button below to view the collated categories and questions.

Data Management Plan
Template Questions

Training and Resources

[Ten Simple Rules for Creating a Good Data Management Plan](#)



Back:
Overview



Next:
System Design

Data Management

Data Capture and Data Management Systems

Research data can be collected using different tools but the most frequently used is the Case Report Form (CRF). CRFs are used to capture information from source data (e.g. medical notes, X-rays, lab tests, interviews etc.) in order to address the protocol endpoints (primary and secondary outcomes) and report regulatory requirements (safety data). A CRF turns the protocol into the data capture system.

Clinical Data Management Systems (CDMS) are software tools available to assist with data management. In a study with multiple sites all collection data a CDMS has become essential to handle the huge amount of data.

Data Capture Systems

There are two types of CRFs:

Paper Case Report Forms (pCRFs)

pCRFs are collected by the research team and the originals are returned to the Principle Investigator, or Project Management Office/Coordinating Team to be entered on to the clinical database. This method may suit smaller studies and those studies where resources are limited or rural research where connectivity is an issue.

Electronic Case Report Forms (eCRFs)

eCRFs are increasing in use—the process used to enter the data is referred to as Electronic Data Capture (EDC). eCRFs offer immediate data entry at the research site, with basic quality checks to prevent the ‘data entry users’ entering ambiguous data resulting in a reduction in errors.

In order to set up eCRFs, a clinical management application must be set-up and all sites must have access to the appropriate equipment which is fit for purpose, e.g. secure log-in, robust internet access and anti-virus software etc.

CRF Design

The CRF design should run parallel to protocol development guidance on CRF design and some examples can be found in the sidebar.

Data Management Systems

Administrative Database

The system for entering and managing ‘personal’ data is often referred to the Administrative Database.

The Administrative Database can be used to manage automated mail-outs of appointment letters and up to date contact details. It can also help you manage your study – show whether you are on target with your recruitment, and hence produce reports to Funders etc. This can save a lot of management and clerical time.

It is good practice to have personal identifiers separate to the clinical data due to security monitoring and data protection.

Clinical Data Management System (CDMS)

The system for entering and managing the ‘clinical’ data is often referred to the study database and CDMS. Participants are ‘assigned’ a unique identifier and so no personal identifiers can be linked e.g. Date of birth, Name, Address etc. This is important in order to ensure patient confidentiality.

MS Excel does not offer any of the features discussed, and is therefore not considered to be a robust tool – users can easily ‘overwrite’/‘amend’ data with no ‘audit/explanation for the change’. This could even happen by mistake, without anyone noticing - yet, this does not stop people from using Excel to record their data.

Choosing a CDMS

There are a large number of data management systems to choose from with wide ranging costs and functionality. Guidance on how to choose a CDMS can be found in the sidebar.



1

Getting started


Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

System Design

CRF Design & examples
Choosing a CDMS




2



3

Data Organisation

Data Entry



4




5


Data Processing

Example from an SOP

Data Monitoring



6



7


Interim Data Lock

Data Analysis

Prepare data for analysis



8



9


Database Closure

Data Sharing

Data Sharing Toolkit
Repository Finder



10



11

Storage and Archiving

How to keep your data safe

Data Management

CRF Design

CRF Design is a very important step in research data management. Time and care must be taken to ensure that the CRFs generated accurately capture the data specified in the protocol. Because of this close link CRF design should be performed alongside Protocol development.

CDASH establishes a standard way to collect data consistently across studies and sponsors so that data collection formats and structures provide clear traceability of submission data into the Study Data Tabulation Model (SDTM), delivering more transparency to regulators and others who conduct data review.

The Global Health Network [Introduction to Data Management For Clinical Research Studies](#) bullet points the main points from the CDASH Best Practices for Creating Data Collection Instruments:

- Necessary data only: collect only data that will be used for analysis and avoid collecting redundant data. The protocol team should draft a statistical analysis plan to define what data is essential.
- CRFs should record sufficient identifiers to ensure that data can unambiguously be assigned to the correct participant but this needs to be balanced against data protection and anonymity requirements.
- Control and document the process of CRF design, printing and distribution. Create standard operating procedures for CRF design, development, quality assurance, approvals, version control and site training.
- Ensure that all members of the study team have adequately reviewed the CRFs before they are finalised.
- Pilot the CRFs if at all possible.
- Keep the end-user in mind and consider the workflow at the study site so that CRF is quick and easy for site personnel to complete. Also, consider the source data: will there be reliable medical charts at the site or will site staff require study-specific worksheets in which to record study observations or measurements; what will the lab reports look like; is it possible to use a central laboratory to ensure consistent results?
- Employ standards for data collection and use CDASH standards wherever possible.
- Use standardised and validated tools for the collection of qualitative data wherever possible (e.g. the Euroqol group’s EQ-5D which is a standardised instrument for use as a measure of health outcome).
- Keep the CRF questions clear and unambiguous and ensure that they are not ‘leading’.
- Wherever possible, avoid collecting ‘free text’ as it requires coding before it can be analysed. It is preferable to use ‘yes/no’ checkboxes or to provide a pre-defined list of possible responses.
- Ensure that translated CRFs are prepared using the same development process as the originals and are reviewed to ensure that the questions have a consistent meaning in all languages.
- Prepare CRF completion guidelines to assist site personnel in completing the forms.

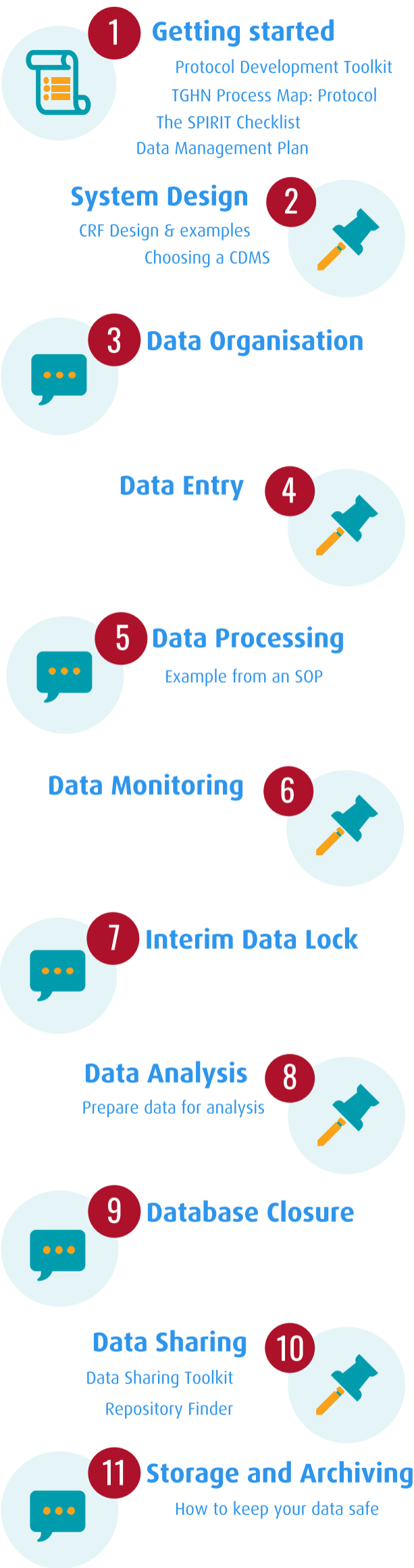
Example Case Report Form (CRFs)

Forms designed to CDASH standards:

- Clinical Data Acquisition Standards Harmonisation (CDASH) [Library of example CRFs](#)
- The Infectious Diseases Data Observatory (IDDO) [Malaria Toolkit](#)
- The International Severe Acute Respiratory and emerging Infection Consortium (ISARIC) [Ebola Data Tools](#)
- ISARIC, PREPARE Europe, and partners developed a number of [Zika Research Tools](#), including a set of maternal and neonatal CRFs that aim to capture core data related to ZIKV infection and a potential link to Microcephaly

Other CRFs on the platform:

- Viral Haemorrhagic Fever Data Tools: [The ISARIC-WHO Data Tools for Viral Haemorrhagic Fever Infections](#)
- [Severe Acute Respiratory Infection Data Tools](#)



Data Management

Choosing a Clinical Data Management System

If designing an in-house system, you need a specification document describing the objective of the system and its functionality - this guide should help with the implementation and testing of the system.

There should be a ‘live’ system and ‘test’ incidence, which can be used to test functionality during development and to train new users. Testing of your new system is very important - take your time and don't skip this step. It is much better to make any amendments up front rather than discover a range of issues halfway through data entry.

There are some key features to good design of a data entry system.

1. Is CFR part 11 compliant

2. Allows users to build trial databases

3. Allows users to develop case report forms with validation checks

4. Has in-built version control functionality to control the release process

5. Includes, or can integrate with, a randomisation service

6. Includes, or can integrate with, a Clinical Trial Management System and treatment supply system

7. Can provide validated outputs in CDISC-compliant format eg, SDTM

8. functionality to allow for a high degree of flexibility with the form layout plus database field and entry field naming.

9. means to interrogate the database directly plus a very flexible report building system (if offered as part of the package)

10. Flexible user management allowing multiple roles with different access rights

11. Ability to import (and validate) data from external files

12. Ability to implement Source Data Verification

13. Provides an audit trail for data entry and system activity

System Selection and Acceptance

It is important to use a ‘validated’ system to demonstrate that the system is reliable and is working in accordance to its remit. Validation is good practice, and sometimes is a requirement of the regulatory authorities.

A key decision is whether to (A) build an in-house system, or (B) purchase an existing commercial system, or (C) use the community version of a commercial system, which is usually at cost zero, but may not be validated - you may need to validate the community edition yourself to make it compliant.

Examples of the commercial products available are:

[InForm Electronic Data Capture](#) (Exeter, New Hampshire,USA)

[MACRO Electronic Data Capture](#) (Elsevier, London,UK)

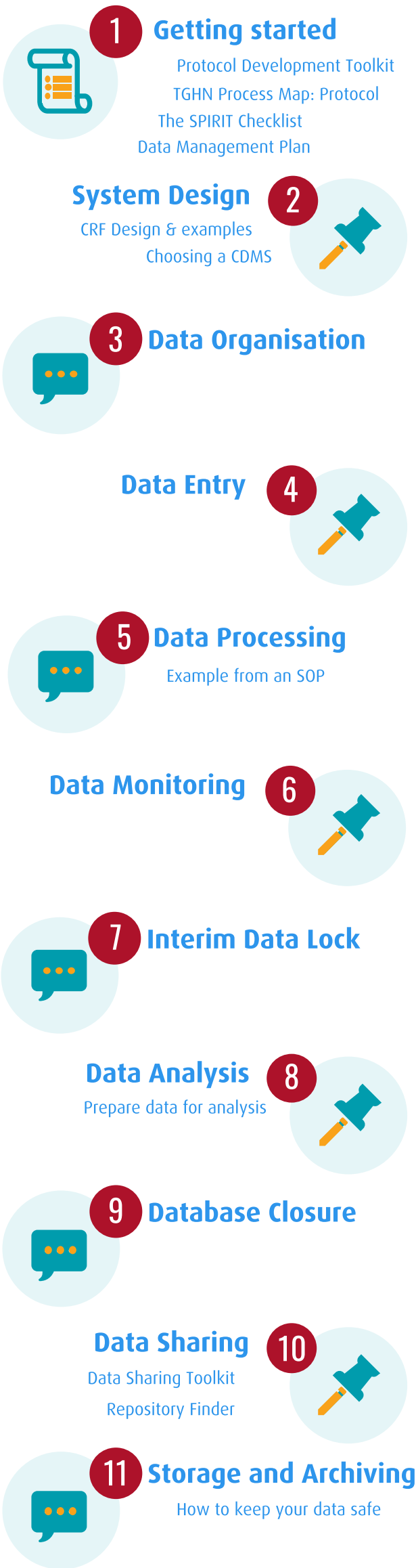
[Rave](#) (Medidata, New York,USA)

[OpenClinica](#) (Waltham, Massachusetts, USA) - FDA, EMA, GDPR and HIPAAcompliant

[REDCap](#) (developed by Vanderbilt, partially supported by the National Center for Research Resources and National Institutes of Health)

For some general guidance on choosing a suitable CDMS for your data see the pilot CDMS Finder tool.

CDMS Finder



Back:
Getting
Started

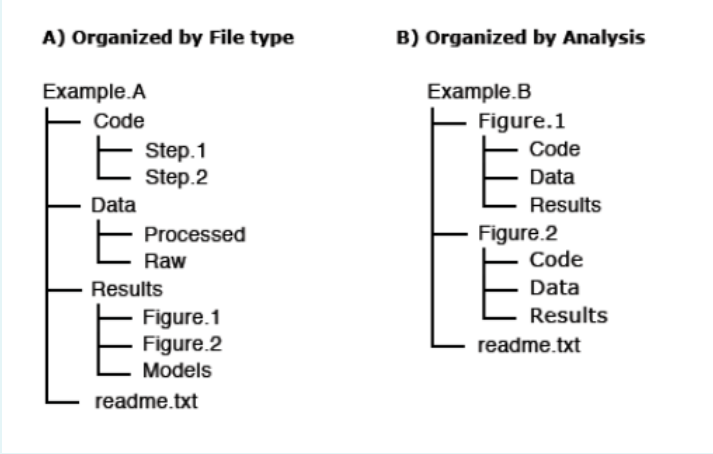


Next:
Data
Organisation

Data Management

Data Organisation

The organisation of you data has a huge impact on how easy it is to work with your data both during and after your research is completed. By structuring and naming your files in a logical way you will help yourself and others.



Example from [Dryad](#)

Describe your data

By describing you data you will help others (and you) to make sense of your data in the future. Things which are perfectly clear when you are doing them my become hazy with time and people will need to understand the processes you followed to collect, process, and analyse your data.

Preparing to share

The file formats you use affect your ability to open those files at a later date. They will also affect the ability of other people to access those data. It is therefore important that you save your data in non-proprietary (open) formats whenever possible.

Ideally, the formats you use should be:

- non-proprietary
- [uncompressed](#)
- commonly used within your research community
- [interoperable across variety of software and platforms](#)

Further information on Data Organisation can be found in the Data Sharing Toolkit.

Data Sharing Toolkit

1

Getting started

Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

2

System Design

CRF Design & examples
Choosing a CDMS

3

Data Organisation

4

Data Entry

5

Data Processing

Example from an SOP

6

Data Monitoring

7

Interim Data Lock

8

Data Analysis

Prepare data for analysis

9

Database Closure

10

Data Sharing

Data Sharing Toolkit
Repository Finder

11

Storage and Archiving

How to keep your data safe

←

Back:
System Design

→

Next:
Data Entry

Data Management

Data Entry

Depending on the set-up of your study data capture may be paper-based, with data from paper CRF being entered onto a database, or Electronic Data Capture (EDC), where data are entered directly into an EDC system.

Paper CRFs

When paper CRFs are first received they should be logged as received, entered on a database, and subjected to query management and audits as per your Data Management SOP. The logging of receipt is an important step in tracking your paper CRFs, a date stamp can be a useful, simple tool to assist with this.

Double or Single Data Entry

In a double entry system two separate users enter the data from each paper CRF independently the system then compares the two records and will only accept them if they are identical. A system must be in place for resolving discrepancies with a view to ensuring that the data available in the database are a true reflection of the information on the forms. An alternative to double data entry is data validation, where another user confirms that what has been entered is consistent with what is recorded on the paper CRF. This might be limited to key data fields related to study endpoints.

Please see [Single vs. Double Data Entry](#) for further information on single versus double data entry.

Electronic Data Capture (EDC)

EDC allows for real-time data entry at “bedside”, close to the source of information. Using an EDC system assists in identifying ‘out of range, or invalid data’ at the point that the data is entered - so the entry can be checked immediately and corrected if it is a simple entry error, or flagged up for querying. Programmers can program the system to check for logic, comparing several answers from a series of questions, and ‘raise’ inconsistencies. E.g. if an infant is on Oxygen for 20 days, and discharged on day 19, then there is an error with one or both answers, and so the answers to both questions must be confirmed. Similarly, if the record is for a man, but also specifies that the patient was pregnant we clearly have an error on our hands!

Data Entry Conventions

To ensure data is entered consistently the DMP or referenced documents should define data entry and processing plans. Data handling guidelines provide details of general study rules, which may cover acceptable abbreviations, symbol conversions, incomplete dates, illegible text, allowed changes and self-evident corrections [Good Clinical Data Management Practices](#).

Medical Coding Dictionary

All free text should be coded to make sense of the data in the analysis stage. What do we mean by coding? Coding is the classification of similar terms using a validated dictionary, it can be done automatically using a program that matches up words to a specific code, or manually, where a specialist assigns a code for each reported term, or a mix of the two. Medical terms, signs, symptoms, diseases, diagnosis, surgical procedures, medical and family history-all can be coded to make data recording, processing and analysis easier.

Data can be just listed, or grouped manually with the input of a physician. However, in large multi centre studies, particularly where there are a number of countries and sites involved, consistency of reporting can be an issue. This impacts on the ability to provide a standardised data set to regulatory authorities.

Examples of medical coding dictionaries:

[Medical Dictionary for Regulatory Activities \(MedDRA\)](#)

[World Health Organization Drug Dictionary \(WHO Drug or WHODD\)](#)



1 Getting started

Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

System Design

CRF Design & examples
Choosing a CDMS



3 Data Organisation

Data Entry



5 Data Processing

Example from an SOP

Data Monitoring



7 Interim Data Lock

Data Analysis

Prepare data for analysis



9 Database Closure

Data Sharing

Data Sharing Toolkit
Repository Finder



11 Storage and Archiving

How to keep your data safe



Back:
Data
Organisation



Next:
Data
Processing

Data Management

Data Processing

The process for handling data discrepancies and query management should be documented in the DMP or referenced documents. An example of an SOP for making corrections to data is in the [sidebar](#).

Checking

Data (pCRF and eCRF) should be checked for quality as they are entered into the eCRF or database and verified for:

- **Completeness:** Data are recorded fully, in the appropriate sections of the CRF and database. The latest CRF version has been used and is completed by delegated persons.
- **Accuracy:** Data is clear and entered correctly into database, without any errors.
- **Logic:** Data must be logical, e.g. the date the participant was entered into the study must be before the date the participant received the intervention (e.g. CTIMP); or the second hospitalization date must be after the first date; etc. All data limits or standards must be satisfied, e.g. head circumference measurement must be within the acceptable ranges or clearly explained if they are not.

Validation

Data validation and query management should be part of the day to day management of your data during or after they have been entered into the database.

Query Management

Sometimes data will validation check, for instance when some CRF boxes were left blank. Every box/field needs to be answered and we can never make assumptions about any of the answers. This is why it is important to set up CRFs correctly, allowing such responses as "Unknown" or "Not done". For example, if a lab test has not been done, we would like to see a "Not done" answer, not an empty box - we cannot assume that the test was not done based on an empty box alone, as it is possible that the data exists, but hasn't been input correctly. We therefore have to send a query out to the research site for confirmation to the question if any of the fields are left blank.

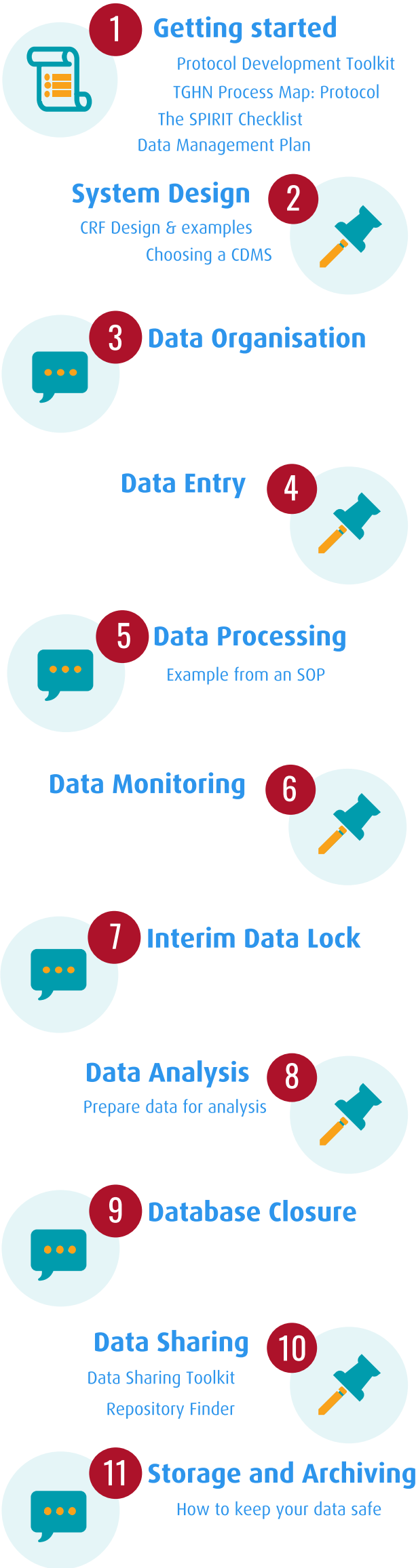
Some of the answers may conflict with answers to other questions, or may not be what is expected - e.g. clinical data outside of the 'normal' range. These too need to be sent back to the centre and queried.

A missing data/discrepancy form should be sent regularly to the research site, should they have any queries. When the form is returned with a resolution the data needs to be amended on the database, following data entry processes. Any changes should be documented for audit purposes.

Audit Trail

An audit trail documents the history of decisions made and it can be paper or electronically based. In the case of data, it shows any manipulations or changes to the data, where it came from, who worked on it, when changes were made and why etc. IT can consist of documents, computer files and other records. Original records have to be date-and time-stamped.

This way any steps taken and any decision made can be re-traced. It should also be clear who made the changes - regulatory agencies may request all this information to verify that data are secured and handled correctly. Verbal discussions on data change decisions should be followed up with written documentation for auditing.



Data Management

Data Processing

The process for handling data discrepancies and query management should be documented in the DMP or referenced documents an example of an SOP for making corrections to data from [CRFs](#), [Source Documents](#), [Record Keeping and Archiving](#) SOP is below:

The PI or appropriate delegate should:

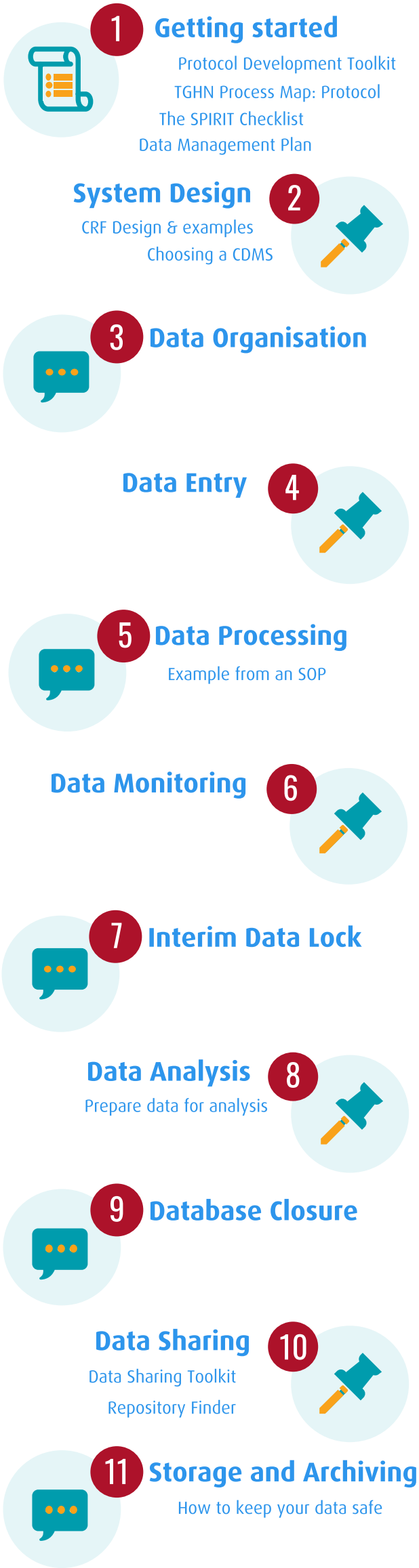
- Ensure that changes are made only by study team members authorised to do so.
- Ensure that the original data is not obscured.
- Changes/corrections are traceable i.e. auditable.
- Retain records of the changes and corrections.

For Paper Documentation:

- Ensure that any change or correction to a CRF, DCF or other source documents is dated, initialled, and explained (if necessary) and should not obscure the original entry (i.e. an audit trail should be maintained); this applies to both written and electronic changes or corrections.

For Electronic Records including ECM medical records

- Ensure that changes are made only by study team members/personnel authorised to do so.
- Ensure that the original data is not deleted and can be accessed if required.
- Changes/corrections are traceable i.e. auditable.
- For changes to information held in ECM:
 - Print the document to be amended, complete any changes or corrections according to good documentation practice including the signature of the person making the change and dated, the reasons for the change should be explained if required. Forward the amended document to HIS for scanning into the ECM with an explanation that this is an amended record on the scanning coversheet.



Back:
Data Entry



Next:
Data
Monitoring

Data Management

Data Monitoring

Monitoring is all about overseeing the trial and its progress - is everything working well, are the procedures all in place, is documentation kept to an adequate level and so on. For instance, you should review patient eligibility criteria before randomisation, confirm that all patients signed consent forms, review CRFs, monitor compliance.

Routine reports

During a study ad hoc or routine reports will need to be generated by the data management group for the ongoing management of the trial; in order to track progress, produce data validation queries, and also, where the protocol allows, for ongoing review of data for risk-based monitoring or safety data for review by oversight committees.

Protocol deviations

A protocol deviation is a departure from the approved protocol procedures, as outlined in the protocol. These should be monitored and recorded, with assessments as to why the deviation happened – are the processes described in the protocol too onerous/not practical within the clinical environment?

Statistical Central Monitoring

Statistical monitoring involves looking for data patterns - anything unusual should be flagged up for review and verified.

Site Visits

It is important to visit sites periodically to ensure that facilities and resources are satisfactory: you may want to have the storage and supplies reviewed, check the inventory, make sure samples are handled and stored appropriately, and that all the necessary documentation - such as signed consent forms - is in place. Review whether any further training is needed for the staff and carry out verification of the source data by comparing the paper CRFs with the information stored in the database.

All site visits should be documented: summarise what was done, what you found and what actions, if any, were performed.

Safety Management & Reporting

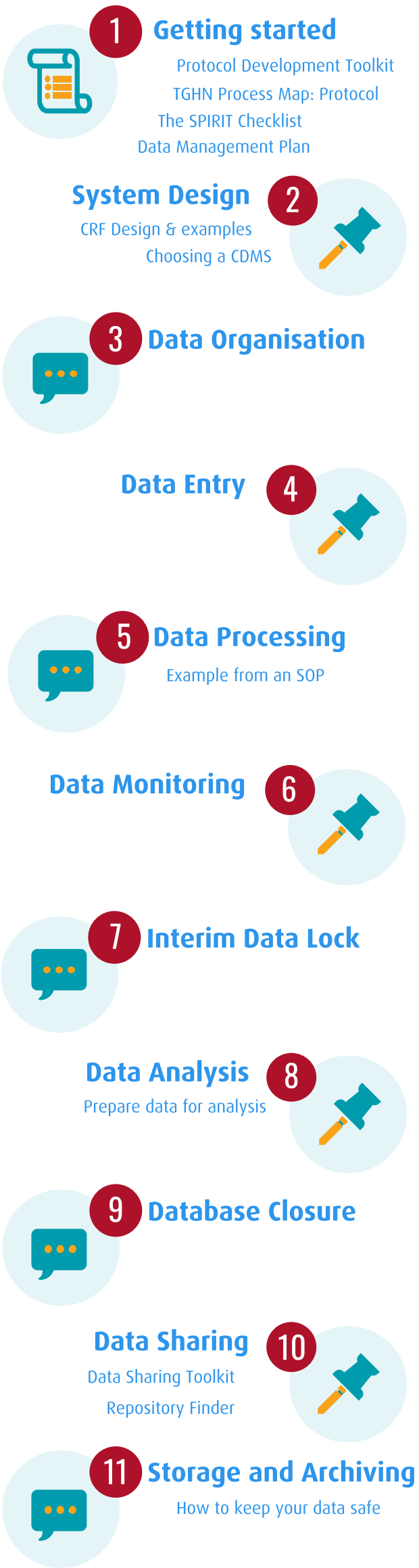
The safety of the participants in your research - patients, trial subjects - is the most important thing. Therefore, it is crucial to monitor the safety within your study throughout the life of the project and notify all concerned if you find the participants are at risk for any reason.

You should agree with your Sponsor about what are the appropriate timelines for the start and termination of safety reporting, and include these in your protocol, relevant SOPs and logs. Define the reporting procedures: how and when you will notify the Sponsor if anything does happen-do they need to be notified within 24 hours of the event being recorded? If so, is that for all types of events, or do they need to know about certain things immediately, but others could wait a couple of days?

Start and endpoints will depend on the type of study you are carrying out. Start points could be: when patients are consented into the trial, randomised or when you start their treatment. During each study visit, you would then ask patients about any Adverse Effects they are experiencing/have experienced. End point could be linked to the end of follow-up period.

Data Safety

In order to ensure that the study participants are not at risk you also need to take care of their data. Data collection protocols should be approved by an ethics committee and compliant with GDPR requirements. Access to both paper and electronic records should be restricted to authorised personnel only.



Data Management

Database Lock/Unlock

This is a controlled procedure to freeze data preventing further write/edit access to users of the system. A database lock is performed prior to an interim or full analysis, once all data entered has been cleaned, with no outstanding queries or discrepancies. Once completed, a statistician, programmer or a data manager will have to approve the lock and no further modification can be done to the data.

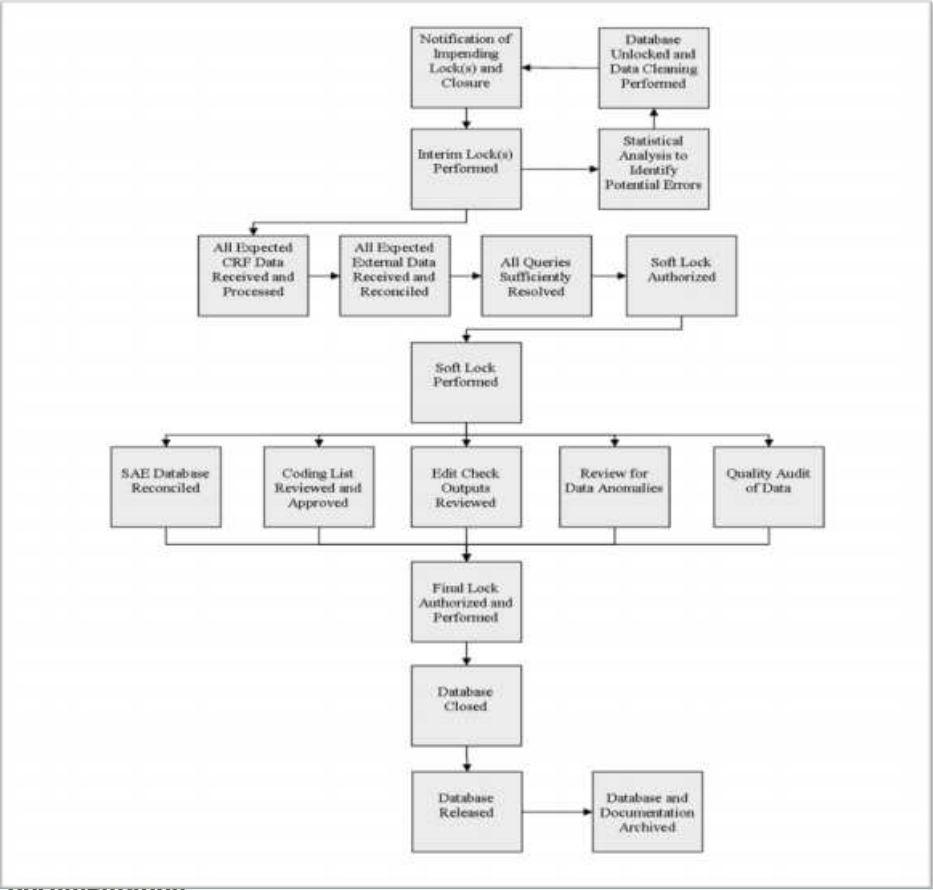
Interim-analysis

A temporary data lock for an interim analysis is an important milestone in a study and the timeline in the run up to the lock need to be well managed. Data Monitoring Committee (DMC) meetings are planned in advance to allow the preparation of data for the Committee to review. The DMC will make recommendations about the future of the study based on the interim analysis and in order to do this they will require the dataset to be as complete and query free as is possible.

In practical terms these interim analysis provide a useful focus for data management time. For example, an impending deadline can be an excellent motivator for sites to push on data entry or for clinicians with limited time to focus on responding to queries or coding data.

The process of performing an interim lock, statistical checks to find any errors, addressing these errors and cleaning the data will not always be in preparation for a DMC. This cyclic process can continue until the data is satisfactorily error free.

The Society for Clinical Data Management: [Good Clinical Data Management Practices](#) has auseful flow diagram showing an example of the database closure steps.



Documentation

A Database lock SOP should be in place which documents the procedures for locking and unlocking the database. This should include the process to resolve any queries or include further data.

If a database requires unlocking before the planned unlock the reason must be provided along with written approval and the effect on the statistical outcome. Unlocking the database or dataset should be limited to important corrections that have a significant impact on the reliability of the results and should therefore be done in consultation with the statistician. Only individuals required to complete the agreed changes should be granted access to the database. The database should be re-locked as soon as possible to prevent any other changes being made.



1 Getting started

Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

System Design

CRF Design & examples
Choosing a CDMS



3 Data Organisation

Data Entry



5 Data Processing

Example from an SOP

Data Monitoring



7 Interim Data Lock

Data Analysis

Prepare data for analysis



9 Database Closure

Data Sharing

Data Sharing Toolkit
Repository Finder



11 Storage and Archiving

How to keep your data safe



Back:
Data
Monitoring



Next:
Data Analysis

Data Management

Data management and data analysis

All clinical research generates data, produced to answer the research question and assess whether the intervention is effective. These data are meaningless unless subjected to statistical analysis. Conversely, the most sophisticated, cutting-edge statistical techniques will not be able to correct data generated by a poorly designed or implemented study.

In a clinical trial, the majority of the problems resulting in defective statistical analysis and interpretation have been attributed to:

- Poor study design
- Poor adherence to inclusion and exclusion criteria
- Inadequate subject recruitment
- Missing data
- Bias
- Wrong application of statistical methods
- Confounding factors

[Statistics from the Beginning](#)

Statistics

Statistical concepts can be difficult for non-statisticians to understand. However, because of this close link between data management quality and the quality of the resulting analysis, statistics are an integral part of clinical research spanning study design, **data monitoring**, analyses, and reporting so it is important that researchers involved with clinical research understand fundamental statistical issues.

CREDO – Statistics This module provides a background of statistical principles relevant to clinical research and trial design and highlights some of the statistical challenges that may occur with trial designs such as adaptive trial designs which may be used in clinical research during an outbreak.

Statistics from the Beginning – Article on the meaning of statistics, its importance and its application in clinical research including the terminologies commonly used in statistics in clinical research and an overview of the basic statistical issues.

Pre-specification

Given the influence of statistical decisions on research conclusions, well-documented and transparent statistical conduct is essential (Guidelines for the Content of Statistical Analysis Plans in Clinical Trials). A pre-specified Statistical Analysis Plan (SAP), will ensure study integrity while enabling the reproducibility of the final analysis and reducing the risks of false positive findings, or biased results due to “fishing” for the desired results (Guide to the statistical analysis plan). The plan should be reviewed and possibly updated as a result of the blind review of the data and should be finalised before breaking the blind (see below) ICH Topic E9 Statistical Principles for Clinical Trials.

Some journals, require the SAP to be submitted alongside the report of a clinical trial for use within the peer-review process.

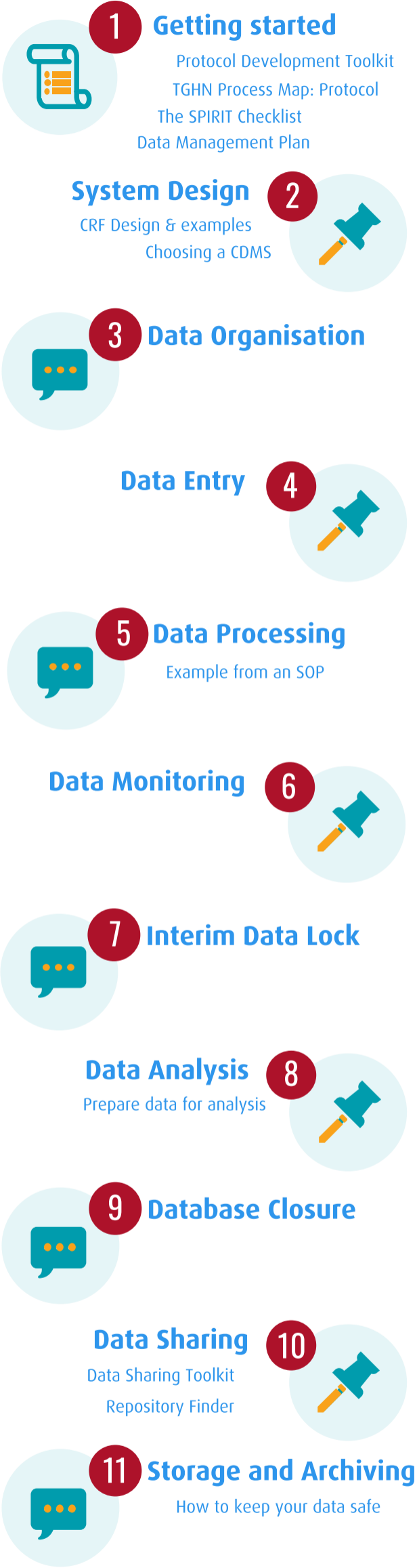
Protocol vs. SAP vs. DMP

The Protocol, SAP and DMP should not be not be developed completely independently. The principal features of the eventual statistical analysis of the data and data management should both be considered during the development of the Protocol.

Documents should be cross-checked to ensure that the statistical analysis and data management strategies are relevant and appropriate for the study design and research question outlined in the Protocol.

Data lock

Data must be **prepared** prior to analysis. Once as much data as possible has been received and data entered, any data which remains missing or data issues which remain unresolved can be ‘coded’ to indicate, for the purpose of analysis, that that data item was missing, invalid, etc. After all the data has been entered, the database will be **locked**. Once a database is locked no data can be added to, or edited within, the system (**Hackshaw A A. Concise Guide to Clinical Trials 2009. Wiley-Blackwell, Chichester**).



Data Management

Preparing data for analysis

In clinical research, errors occur in spite of careful study design, conduct, and implementation of error-prevention strategies. Data cleaning intends to identify and correct these errors or at least to minimize their impact on study results. [Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities](#).

Data Monitoring should occur throughout recruitment to ensure quality data collection. Once the study is closed, a final data cleaning exercise should be undertaken. This involves chasing any essential data missing from the CRFs (e.g. in the case of a birthing survey this could be whether the birth outcome was a live or still birth) and resolving out of range or logical issues and inconsistencies.

All changes should be justified and documented for audit purposes (see [Data Entry](#) and [Data Processing](#)).

Data cleaning: Process of detecting, diagnosing, and editing faulty data.

Data editing: Changing the value of data shown to be incorrect.

Data flow: Passage of recorded information through successive information carriers.

Inlier: Data value falling within the expected range.

Outlier: Data value falling outside the expected range.

Robust estimation: Estimation of statistical parameters, using methods that are less sensitive to the effect of outliers than more conventional methods.

[Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities](#)



1

Getting started

Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

System Design

CRF Design & examples
Choosing a CDMS



3

Data Organisation



Data Entry



5

Data Processing

Example from an SOP



Data Monitoring



7

Interim Data Lock



Data Analysis

Prepare data for analysis



9

Database Closure



Data Sharing

Data Sharing Toolkit
Repository Finder



11

Storage and Archiving

How to keep your data safe



Back:
Interim Data
Lock



Next:
Database
Closure

Data Management

Database closure

Research databases must be properly closed and edit access removed to ensure data integrity for the generation of results, and analysis.

Database lock versus database closure

The Society for Clinical Data Management: [Good Clinical Data Management Practices](#) has a useful flow diagram showing an example of the database closure steps.



The process of performing an interim lock, statistical checks to find any errors, addressing these errors and cleaning the data is cyclic and can continue until the data is satisfactorily error free.

Before you close your database the following tasks should be completed:

- Procedure for lock/unlock must be clearly defined
- All data must have been received prior to database lock
- All cleaning procedures must have been completed
- All queries should have been resolved
- External data (e.g. safety database, lab data) must have been reconciled
- Final consistency check of database (also with statistical methods)
- Conditions for unlock should be defined
- Coding should be reviewed
- A database audit might be useful (documenting error rate)

ECRIN-TWG: [GCP-compliant data management in multinational clinical trials](#)

Documenting database closure

Database closure must be documented as proof that access allowing the database to be edited was removed and when. This could include a checklist with the facility to sign-off tasks and add a completion date allowing the steps which go in to database closure to be recorded.

Once the database closure steps are complete, and the necessary approvals have been obtained, access to edit the database should be removed, and the date of the removal documented.

1

Getting started

Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

2

System Design

CRF Design & examples
Choosing a CDMS

3

Data Organisation

4

Data Entry

5

Data Processing

Example from an SOP

6

Data Monitoring

7

Interim Data Lock

8

Data Analysis

Prepare data for analysis

9

Database Closure

10

Data Sharing

Data Sharing Toolkit
Repository Finder

11

Storage and Archiving

How to keep your data safe



Back:
Data Analysis



Next:
Data Sharing

Data Management

Why share data?

It is now a standard requirement by publishers, funders, research institutions and regulatory agencies to share data. Data sharing achieves many important goals for the scientific community, such as:

- Reinforcing open scientific inquiry
- Encouraging diversity of analysis and opinion
- Promoting new research, testing of new or alternative hypotheses and methods of analysis
- Supporting studies on data collection methods and measurement
- Facilitating education of new researchers

[NIHR Information for authors](#)

How can you prepare to share?


Many elements of data management are crucial to data sharing if we want to share data in a meaningful way. How you will share your data should be considered at data management planning stage to ensure that the way you collect, organize, document and store your data will facilitate it being discoverable and reusable. Access the [Data Management Plan \(DMP\)](#) section for further information on the elements you need to consider when developing your DMP.

How to share your data

Despite funder and publisher requirements you may be unclear on how, where and when to share your data. The [Data Sharing Toolkit](#) contains extensive guidance on preparing your data for sharing, the [Repository Finder](#), an online tool to help you select an appropriate repository for your health research data and a searchable list of resources related to data sharing to help you comply with these data sharing requirements.

For excellent advice on how to prepare you data for archiving, from funder requirements and data ownership to organising your data and choosing a suitable repository for your data see the Data Sharing Steps in the Data Sharing Toolkit and the Repository Finder tool.

Data Sharing Toolkit




Repository Finder



Other initiatives supporting data sharing include platforms providing advice:

- [The Digital Curation Centre](#)
- The [Research Data Alliance](#)
- [Chatham House’s guide](#) to sharing health surveillance data
- Repositories where data can be archived (with [re3data.org](#) collating multiple repositories)
- Consortia working on standards supporting interoperability between different systems (e.g., the [Clinical Data Interchange Standards Consortium](#))
- Groups developing tools for specific diseases (e.g., Malaria Toolkit ([Infectious Diseases Data Observatory](#)))
- Ebola Data Tools ([ISARIC](#))
- Zika Research Tools ([ISARIC](#), [PREPARE Europe](#), and [partners](#))
- Trial registries facilitating discovery of the data sets such as [ClinicalTrials.gov](#)
- [ISRCTN](#)
- [EU Clinical Trials Register](#)

It is not enough that we require data to be shared; we have to make sharing easy, feasible and accessible too!



1


Getting started


Protocol Development Toolkit
TGHN Process Map: Protocol
The SPIRIT Checklist
Data Management Plan

System Design

2

CRF Design & examples
Choosing a CDMS







3

Data Organisation

Data Entry

4






5


Data Processing

Example from an SOP

Data Monitoring

6






7


Interim Data Lock

Data Analysis

8

Prepare data for analysis






9


Database Closure

Data Sharing

10

Data Sharing Toolkit
Repository Finder






11


Storage and Archiving

How to keep your data safe



Back:

Database Closure



Next:

Storage and Archiving

Data Management

Data Storage

There are two slightly different types of "storing" data:

- Short-term: Storing data that is being actively used in a current project (working data), including periodic back-ups of said data – the back-up storage serves as an insurance in case of data loss, as data can be restored/recovered from back up in such an event.
- Long-term: Storing data by submitting it to a long-term archive in case it is needed again, but where no more work and maintenance is being done on it.

Short-term: Data storage and back-ups

Safe storage of your working data and regular backups are essential during your research project. Backups are a key component of your research data management strategy protecting against the risk of damage or loss due to hardware failure, software faults, viruses or power failure etc. Your plans for back-up and storage of your data should be detailed in your [Data Monitoring Plan](#).

How to store your working data

The UK Data Service has a helpful section on the storage media and the correct physical conditions for storing data. Even for a short-term project, they recommend involving at least two different forms of storage, for example on hard drive and on DVD and checking data integrity periodically.

Version control should be part of data storage and maintenance – all changes to your data and code should be recorded. There are tools (version control systems) available that can automatically deal with this, creating a ‘history’ of all revisions, with time stamps.

Security

Security refers to keeping your data safe. This means both ensuring that data are not lost or corrupted and controlling access to your data to ensure that only authorised people can see your data. See '[How to keep your data safe](#)' for further information.

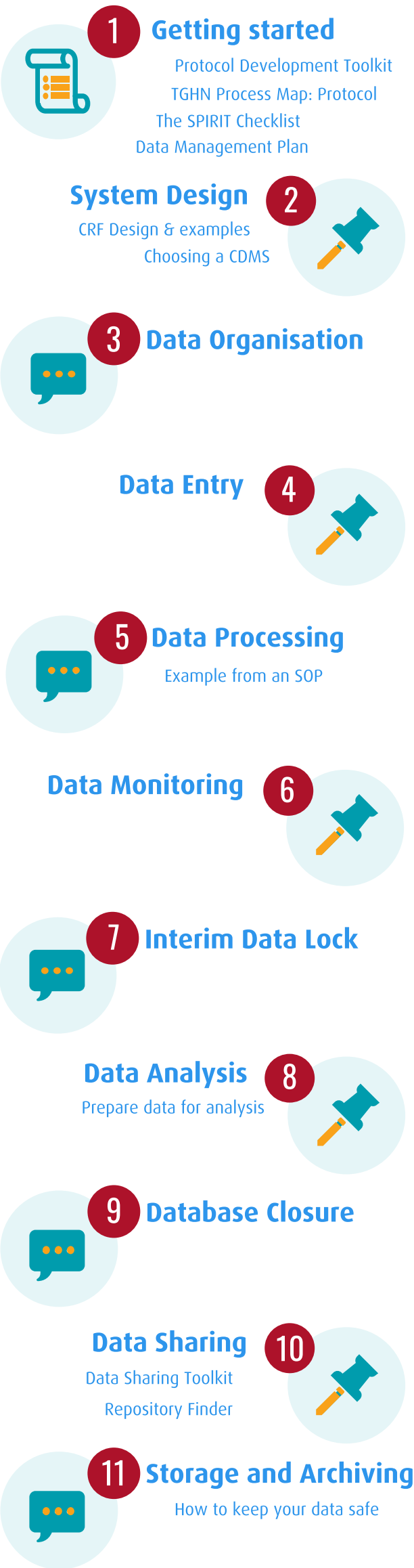
Long-term: Archiving

The purpose of archiving is preservation. Funders and regulators may state that the data must be archived for a certain period of time. Data archiving is vital to data sharing, discovery and dissemination (note: data should be archived whether or not they will be shared with others).

For excellent advice on how to prepare your data for archiving, from funder requirements and data ownership to organising your data and choosing a suitable repository for your data see the Data Sharing Steps in the Data Sharing Toolkit and the Repository Finder tool.

Data Sharing Toolkit

Repository Finder



Data Management

Security

Security refers to keeping your data safe. This means both ensuring that data are not lost or corrupted and controlling access to your data to ensure that only authorised people can see your data.

Password security

To protect your data files, you should use passwords to lock the computer systems used to access these data files. Besides choosing strong passwords, make sure to store and transmit them securely so they cannot be stolen:

Ways to keep your data secure:

- Do: store passwords in a sealed envelope in a secure place (e.g. a safe)
- Do: use secure password management tools. Remembering all of your passwords can be a challenge. Password management tools are one possibility of dealing with this problem. Examples are KeePassX (2017) and Lastpass (2017)
- Do not: write passwords down and leave them lying about openly (e.g. in your desk drawer)

Encryption

Encryption is the process of encoding digital information in such a way that only authorised parties can view it. It is especially useful when you are transmitting personal or confidential data. When you encrypt a file, the information it contains is “translated” into meaningless code. To translate this code back into meaningful information a key is required.

- Do: ensure that you do not lose the key to decrypt your files, e.g. by keeping it in a sealed envelope in a secure location such as a safe room
- Do: encrypt confidential data, especially before transmitting it online, uploading it to the cloud, or transporting it on portable devices. When working in a team, make sure that the key can be accessed by everyone who needs to access it (but only those people).

Physical, network and computer security

To prevent your data from being manipulated or stolen, sufficient security measures to block any unwanted access to rooms and buildings or computers and networks where they are held should be in place.

- Do: log and/or control access to physical sites where sensitive information is stored, e.g. with the help of key cards.
- Do: use strong passwords and encryption (see above).
- Do: use up-to-date virus scanners and firewalls.
- Do: ensure that systems used to access data are continually updated (e.g. security updates for the operating system).

From the [CESSEDA training module on Security](#)

