Review Article

How to set up a database?—a five-step process

Alice Brembilla[1,2], Bérenger Martin[1], Anne-Laure Parmentier[1,2], Maxime Desmarets[1,3], Pierre-Emmanuel Falcoz[4,5,6], Marc Puyraveau[1,2], Frédéric Mauny[1,2]

[1]Unité de Méthodologie en Recherche Clinique, épidémiologie et de Santé Publique, Inserm CIC 1431, CHU de Besançon, France; [2]UMR CNRS 6249 Chrono-Environnement, [3]UMR1098 Inserm, Etablissement Français du Sang, Université Bourgogne Franche Comté, Besançon, France; [4]INSERM (French National Institute of Health and Medical Research), UMR 1260, Regenerative Nanomedicine (RNM), FMTS, Strasbourg, France; [5]Faculté de Médecine et Pharmacie, Université de Strasbourg, Strasbourg, France; [6]Hôpitaux Universitaires de Strasbourg, Service de Chirurgie Thoracique - Nouvel Hôpital Civil, Strasbourg, France

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: None; (V) Data analysis and interpretation: None; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Dr. Alice Brembilla. Unité de méthodologie en Recherche Clinique, épidémiologie et de Santé Publique, Inserm CIC 1431, 25000 Besançon, France. Email: abrembilla@chu-besancon.fr.

Abstract: Database set-up directly impacts the quality and viability of research data, and therefore is a crucial part of the quality of clinical research. Setting up a quality database implies following a strict data-management process. Too much collected information threatens the quality of the information required to achieve the objectives of the study. Therefore, the data that will be collected and managed have to be cautiously discussed and selected. Case report forms (CRF) are the tools the most frequently used to collect the data specified by the protocol. An informative and well-structured document simplifies database design and data validation. Key elements are about choice of sequential or thematic structuring, information and type of information that should be entered and the importance of data standards and coding guide. Final database must be structured with unique ID patient, with one record per subject or per measure. Specific information must be provided for each variable according to the database specifications. The quality of the results is directly related to the quality of the collected data. The CRF should then be completed as fully and accurately as possible. Data validation relies on three key points: the CRF completion guidelines, the Edit Checks process and the Data clarification process. Various open source or business software applications provide all functionalities to set up a clinical data base and CRF. The General Data Protection Regulation (GDPR) standardizes and strengthens the protection of personal data across the EU and for other country's data being "processed" within the EU. The General principles include lawfulness, fairness and transparency, restricted use of data, data minimization, accuracy, limited storage, confidentiality and probity, and accountability.

Keywords: Database; software; General Data Protection Regulation (GDPR)

Submitted Sep 14, 2018. Accepted for publication Sep 27, 2018.
doi: 10.21037/jtd.2018.09.138
View this article at: http://dx.doi.org/10.21037/jtd.2018.09.138

Introduction

Database set-up directly impacts the quality and viability of research data, and therefore is a crucial part of the quality of clinical research. Setting up a quality database implies following a strict data-management process (1). This process should be described in a data-management plan, "a to-do list that details how one plans to collect, clean, store and share the products of their research" (2,3). Data-management had been the subject of continued interest and development for over three decades (*Figure 1*), largely before the emergence of the concept of Big Data and the

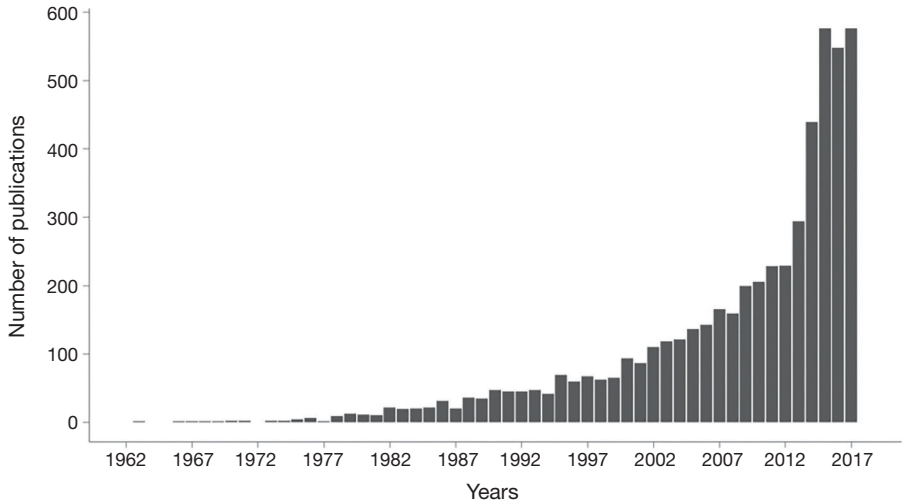

Figure 1 Annual number of publications indexed in Medline® and associated with the "Data Management".

data management issues associated with it (4). Whatever the size of the database, data-management is based on the same main rules and concepts. This paper aims to provide the basics of database set-up, with a particular focus on the case report form (CRF). We will discuss data collection, database conception, building, computerization and data validation, as well as regulatory aspects.

The query was conducted using PubMed® (builder = "data management" in "All fields", 09 August 2018).

Data collection

Which data should be collected? (5-7)

This question of great importance is about the relevance of the collected data with regard to the objectives of the study. Too much (collected) information threatens the quality of the information required to achieve the objectives of the study. Therefore, the data that will be collected and managed have to be cautiously discussed and selected (see also "Evolution of the legal consideration in Europe" below). The golden rule "one scientific question, one objective, one study" also applies when the data are selected for the study, keeping in mind that the usefulness of all the collected pieces of data is unequal. The retained data will provide different information related to specific roles and functions. The function of the data should be clarified and be in line with the objective of the research: data which will be used to assess the respect of the protocol, those for safety and quality purposes, those to give the main general characteristics of the study sample, those to analyze to address the objectives. Indirect identification by the data management should be maintained for validation, correction, for patient and medical records queries (even the recorded data must be anonymized). The actual collected data will be used to define:

(I) The outcomes (primary and secondary);
(II) The eligibility criteria;
(III) The matching/stratification criteria;
(IV) The relevant groups of patients (allocated treatment, case *vs.* controls, exposed *vs.* unexposed…);
(V) The known and potential confounding factors;
(VI) The factor potentially involved in effect modification or mediation;
(VII) The general characteristics of the study sample.

How should the data be collected? The case report form

Case report forms are the tools the most frequently used to collect the data specified by the protocol. An informative and well-structured document simplifies database design and data validation. The CRF should be designed in accordance with the following rules:

(I) Use a multidisciplinary approach to designing and developing the CRF: the project leader, the methodologist(s), the clinical and safety personnel, and the data manager(s) provide valuable perspectives to help optimize the CRF;
(II) Design the CRF with safety and efficacy endpoints in mind. In order to do this, consult the study, and review the statistical analysis plan (if available) to

ensure that all key endpoints are collected;
(III) Design the CRF to follow the data flow from the perspective of the person in charge of the completion, and take into account the flow of study procedures. Data that are logically related should be grouped together in CRF sections. Use the time schedule of enrolment, interventions and visits to structure the document;
(IV) Keep the CRF questions, prompts, and instructions as clear and concise as possible.

Database conception: structure and organization of the data

Sequential or thematic structuring?

Two major organizations/strategies can be implemented. The sequential organization structures data in chronological blocks (i.e., regularly/periodically collected blocks of data), such as daily body temperature or dosage antibodies. The thematic organization is adapted when one data statement is necessary like medical history. The type of organization must be linked with the variable definition.

Which information should be entered?

To avoid the loss of information, crude data will be relevant. For example, if we want to collect body mass index, we must ensure that crude data such as size and weight are entered. The body mass index will be calculated automatically in a post-processing phase. If necessary, information about weight will be available and can be used for another description or calculation.

Which type of information?

As much as possible, the data collection should be based on closed-ended questions. In this case, the doctor/Data Entry Clerk chooses among a fixed set of choices. Some data may also be collected with open-ended questions and free-form text. For example, it can be allowed to specify additional information when a category "other" is present. This strategy prevents the loss of data without building an overly long list of choices, some of which will occur only rarely over the course of the study. The heterogeneity of such written information requires specific processing to homogenize the information. It is therefore recommended to use open-ended questions sparingly.

Importance of data standards

The unit of measure and the rounding value of each variable must be specified. The international system of units should be preferred. Thus, the data survey will be homogeneous making analysis possible.

Coding guide

The organization/arrangement of data is a precondition/prerequisite to computerize the data base. It's recommended to synthesize in a document the name and the definition of the collected data, with its unit of measure and rounding value. This document allows the data-manager to faithfully build the data base.

Database building: data computerization

How should the database be structured?

The final structure of the database should be constituted as follows:

(I) If there is only one measure of the main outcome per subject one record per subject in the database is appropriate;
(II) If there are repeated measures, then the database should contain one record per measure for the subject.

Each subject needs to be identified with a unique ID. If there are several records per subject, they should be identified with a unique record ID.

Each variable must be presented in a distinct column with unique and synthetic brief but clear name.

How must variables be coded? (8)

Database specifications should at least provide the following information for each variable:

(I) Name and label (questions asked on the CRF);
(II) CRF section, forms, or other logical group to which the data belongs;
(III) Type (e.g., numeric, text, date, time);
(IV) Length (including number of characters before and after the decimal point, when applicable);
(V) Definitions for all coded values included in code lists;
(VI) Algorithms for derived or calculated variables;
(VII) Dependent relationship.

The CRF should be annotated to map the sections, visits, forms and collected items to their corresponding database tables and variable item names.

Data validation

How to increase data quality?

The quality of the results is directly related to the quality of the collected data. The CRF should then be completed as fully and accurately as possible. This section proposes a focus on three key points: the CRF completion guidelines, the Edit Checks process and the Data clarification process.

CRF completion guidelines (8)

CRF completion guidelines will help to ensure that all required fields will be completed, and that the data provided within these forms are in compliance with the study protocol. The guidelines should not provide guidance or suggestions that could be considered leading the user. In order to standardize the data collection, the CRF guidelines should be developed in accordance with the following rules:

(I) Provide any instructions on mandatory/optional fields, and clearly define acceptable notations if a data item is unavailable or unknown;

(II) Provide any special instructions for completing missing values, clearly reporting:
- any visit that a subject failed to make;
- any clinical biological test that was missed, examinations that were not performed;
- all withdrawals and dropouts of enrolled subjects from the trial.

(III) Provide instructions for recording Adverse Events and Serious Adverse Events (e.g., record diagnosis instead of symptoms whenever possible).

Edit checks process

The edit checks process allows to automate data reviewing in order to reduce potential data errors and inconsistencies in accordance with the study requirements. It improves the quality of the data and reduces the data reviewing and data-cleaning activities.

Edit checks generate warnings related to missing, out of range, unexpected, incompatible data. They may also point to discrepancies with other data or study parameters.

Edit checks are mainly focused on data related to eligibility criteria to avoid protocol violations, on data related to endpoints, and on safety data. This process is directly derived from the study protocol, the CRF and the database documentation. The edit checks process should be performed directly during the data entry process (to draw attention before saving the data) or should be performed in post-processing and then will generate queries.

Data clarification process

The process allows to describe how queries will be generated, transferred to the clinical site, tracked and reviewed.

Software application

Various open source or business software applications provide all functionalities to set up a clinical data base and CRF (web-based data entry, edit check programming, queries management and data clinical trials management system). The choice between open source or commercial solutions often depends on time, budgetary considerations, and human resources and skills. These solutions generally meet regulatory requirements on data hosting, data transfer and data control access to ensure data security and confidentiality. Simple data file or even "home-made" solution should not be used without keeping in mind that these requirements have to be respected to the letter.

Legal evolution in Europe: the General Data Protection Regulation (GDPR)

The aim of the GDPR (European regulation effective May 25, 2018) is to standardize and strengthen the protection of personal data across the EU and for other country's data being "processed" within the EU (9). Under the new regulation, each structure (industrial or institutional research) processing of the personal data of subjects residing in the Union must comply. Personal data are defined by information that allows or could allow identifying the person. It is important to distinguish identifiable data (even if it is key coded) from completely anonymous rendered data, as the GDPR applies only to the first. Personal data may contain any information relating to a person (private, professional or public life including health) (10). GDPR applies to treatment of data by a third party (a person or legal entity).

The GDPR aims: (I) to increase the rights of patients to be better informed about how their data I used and (III) to set out clearer responsibilities and obligations on healthcare professionals using such data. Transparency, security, and

Table 1 The general principles of data protection regulation (11)

Lawfulness, fairness and transparency: respect for the European and state laws must be applied scrupulously to data processing. Data controllers have to ensure that patients know the future of their personal data
Restricted use of data: the data is collected under a single, well-defined objective and should not be used for other purposes. However, an exception has been provided by the legislator as part of the scientific research on a possible re-use of data
Data minimization: only relevant data related to the research protocol can be retrieved. The data controller is responsible for compliance with this rule
Accuracy: data controllers are responsible for the consistency between the source data and the data collected. In case of discrepancies, data controllers must take the necessary measures to correct the data and avoid future inconsistencies
Limited storage: data can only be stored for a limited period. Notables exceptions are related to archiving and research
Confidentiality and probity: the heart of a data management process must be the confidentiality and integrity of the data
Accountability: European law empowers data controllers. They are responsible for the respect of the law and must be able to prove that they gave themselves all the necessary means to make it respected by all the persons involved in a research project

accountability of *Data Controllers* are determinant. In the GDPR, Data Controllers are defined as any person or entity which collect or process personal data. Structures involved in research must identify: (I) the data that is being processed, (II) where it is transferred to, (III) who processes the data, (IV) what it used for, (V) any risks and processes, and (VI) must ensure all people are trained.

The conditions for consent have been also consolidated: consent must be given in a clear, intelligible, and easily accessible form, using clear and plain language with the purpose of the research attached to that consent. Patients should be informed on how to withdraw consent. And finally a Data Controller has to be able to demonstrate that a person has given consent.

This GDPR does not only cover patients participating in clinical trials, but also any individual involved in medical and health research. The sponsor of the research has obligation to make sure that rules are in place, known and followed by all. Data impact assessments will be crucial, for both electronic and hard copy data. It should cover what the data will be used for, how it will be managed, and which actions will be needed. The *Data Protection Officer* has been also defined by the GDPR: he is responsible for overseeing data protection strategy and implementation to ensure compliance with GDPR requirements.

Patient's rights and the existing exemptions to these rights are also well-defined in the context of research (9). Patient's rights cover different main points such as:

(I) Access own personal data;
(II) Rectification or erasure of data;
(III) Being forgotten;
(IV) Rights in case of security breach;
(V) Lodging a complaint and right to compensation;
(VI) Be informed.

European law aware of the complexity of a scientific research project admits some exceptions to these principles. Derogation may be granted to the principle of information and deletion of data if this compromises the research, for example: when providing the information is a "disproportionate effort" (time or cost) or can be a problem in a blind trial.

The GDPR sets clear principles that apply to all use of personal health data and to all Data Controllers (article 5). Seven principles were extensively presented in a guide for patients and patient's organization (*Table 1*) (11).

Conclusions

The Data collection, the design of the case report form, the database conception, the structure and organization of the data and the data validation are essential steps of a quality database set-up.

At each all these stages, it is essential to follow rules that may appear constraining but are easy to follow and make common scientific sense. The focus should be not only on the data and their organization but also on data safety and confidentiality. Regarding this important issue, the General Data Protection Regulation has been implemented and deployed in the European Union to standardize and strengthen the protection of personal data. With this evolution of European community, it is expected that the legal framework will be prompted to evolve in Europe but also in a world-wide context.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Krishnankutty B, Bellary S, Kumar NB, et al. Data management in clinical research: An overview. Indian J Pharmacol 2012;44:168-72.
2. Everyone needs a data-management plan. Nature 2018 15;555:286.
3. Schiermeier Q. Data management made simple. Nature 2018;555:403-5.
4. Househ MS, Aldosari B, Alanazi A, et al. Big Data, Big Problems: A Healthcare Perspective. Stud Health Technol Inform 2017;238:36-9.
5. Kelsey JL, Whittemore AS, Evans AS, et al. Methods in Observational Epidemiology. Second Edition. Oxford, New York: Oxford University Press; 1996:448 p. (Monographs in Epidemiology and Biostatistics).
6. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.
7. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg 2014;12:1495-9.
8. Good Clinical Data Management Practices (GCDMP). Available online: https://www.scdm.org/publications/gcdmp/ Last access on 09 August 2018.
9. [GDPR] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng. Last access on 09 August 2018.
10. European commission - Press release. Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses. 2012. Available online: http://europa.eu/rapid/press-release_IP-12-46_en.pdf Last access on 09 August 2018.
11. European Patients' Forum. A guide on data protection for patients and patients' organizations. Available online: http://www.eu-patient.eu/globalassets/policy/data-protection/data-protection-guide-for-patients-organisations.pdf. Last access on 09 August 2018.

Cite this article as: Brembilla A, Martin B, Parmentier AL, Desmarets M, Falcoz PE, Puyraveau M, Mauny F. How to set up a database?—a five-step process. J Thorac Dis 2018;10(Suppl 29):S3533-S3538. doi: 10.21037/jtd.2018.09.138

