Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-41

From Data Extraction to Analysis: Proposal of a Methodology to Optimize Hospital Data Reuse Process

Antoine LAMER[a,b,1], Grégoire FICHEUR[a], Louis ROUSSELET[a], Marine VAN BERLEERE[a], Emmanuel CHAZARD[a], and Alexandre CARON[a]
[a] *EA 2694,Univ. Lille, Department of Public Health, CHU Lille, F-59000 Lille, France*
[b] *Anesthesia Department, CHU Lille, F-59000 Lille, France*

Abstract. In the Lille University Hospital (North of France), data from the Anesthesia Information Management System (Diane®) are linked to the Hospital Information System and stored in a dedicated data warehouse since 2010. These electronic medical records need to be reused and analyzed for observational studies. The aim of this paper is to describe the framework developed to structure the operation of that anesthesia data warehouse for research purposes. The presented framework is structured around three meetings between clinicians, computer scientists, and statisticians. The data scientist acts as a coordinator, leads meetings, and checks each milestone. Reuse of anesthesia-related electronic medical record for research purposes is only allowed through this framework. The aim of the first meeting is to decide the primary and secondary objectives of the study. The aim of the second meeting is to validate the statistical protocol. The data are extracted and the statistical analyses are performed. Finally, the results are presented, explained and discussed during the third meeting. During a 6 months period, 27 projects were included in the framework leading to 5 scientific communications. As a result, case studies with extraction and/or analysis situations are presented. This collaboration led to an empowerment process between all three actors, which increased efficiency of the workflow. Implementation of this framework will keep encouraging collaborative publication in order to provide reproducible research evidence.

K**eywords.** Data Science, Healthcare Data Reuse, Statistical Analysis, Reproducible Research, Electronic Medical Records.

Introduction

In most hospitals of developed countries, a large amount of clinical data is routinely generated through the healthcare process. Some hospitals developed Clinical Data Warehouses to collect and store these structured medical records. Gathered data can thus be reused for hospital management, medical decision making and research [1]. In the Lille University Hospital (North of France), data from the Anesthesia Information Management System (Diane®) are linked to the Hospital Information System and stored in a dedicated data warehouse since 2010. These electronic medical records need to be reused by the department of anesthesiology for observational research studies.

[1] Corresponding Author: Antoine Lamer, CHRU de Lille, 2 Avenue Oscar Lambret, 59000 Lille, France; E-mail: antoine.lamer@chru-lille.fr.

Reproducible research is considered as an attainable minimum standard for assessing the value of scientific claims [2]. In the last decades, Knowledge Discovery in Databases, CRISP-DM and SEMMA processes were set up to structure the data reuse through a roadmap consisting of five to six phases [3, 4]. However, reproducibility is challenging when operating databases. Indeed, a detailed description of the data extraction, management and statistical analysis is required [5]. In a previous work, we published the algorithm enabling reproducibility of the detection of abnormal parameters in anesthesia time-series data [6].

When the operation of the data warehouse started, the process was performed sequentially with minimal interaction between the actors (clinicians, computer scientists, and statisticians). Numerous queries were then received by the computer scientists in charge of data extraction without medical background justification, and without consistency with the subsequent statistical analysis. In a second step, the clinician used to send data to the statistician in order to get the statistical analysis.

However, the design of the study, the data extraction and management, and the statistical analysis are not sequential and independent steps. Indeed, numerous back loops were observed after delivery of the statistical analysis, leading to a substantial waste of time, and preventing reproducible research. Thus, a tight collaboration between clinicians, computer scientists, and statisticians was required at all stages of the research process.

The aim of this paper is to describe the framework we developed to structure the operation of the anesthesia data warehouse for observational clinical research purposes.

1. Material and Methods

1.1. Anesthesiologic data warehouse

During the routine care process, patients' demographics, clinical characteristics, and postoperative data are collected into the Electronic Medical Records (EMR) system. At the same time, preoperative data, intra-operative surveillance, and postoperative follow-up are routinely collected into the Anesthesia Information Management System (Diane, Bow medical®) [6-8]. All those data are then loaded into the anesthesia data warehouse through an ETL (extract, transform and load) process. ETL process notably includes data cleansing, terminological alignment, and domain-specific transformations and computations.

1.2. Framework for the operation of the data warehouse

The proposed framework is structured around three meetings between clinicians, computer scientists, and statisticians. The data scientist acts as a coordinator, leads meetings and checks each milestone. Regarding anesthesia data, the reuse of EMR for observational research purposes is only allowed through this framework (Figure 1).

For the first meeting, clinicians have to provide a detailed background, and list all relevant variables (end-points, exposures, confounding factors, etc.) through a literature review. Furthermore, an overview of the methodology of each study (including sample size) is also required. A Strobe-based template is used to structure the review [9].

The aim of the first meeting is to decide the primary and secondary objectives of the study. The feasibility is assessed by the data scientist who compares the objectives

with the data available in the data warehouse. The opportunity to compute new variables from existing data may be considered. When the data are only available on paper, the feasibility of manual collection and further merging is discussed between the clinician and the data scientist. If necessary, a sample size computation is made by the statistician.

As a result of this first meeting, a list of variables to extract and/or collect is determined. De-identification process is carried out during the extraction step and data are aggregated if necessary (indirectly nominative). Once available, data are merged by the data scientist. A comprehensive descriptive statistical analysis (including missing data) is performed, and then enables the clinician to control the quality of the data.

Before the second meeting, clinicians are requested to provide "dummy results", i.e. empty tables, text and figures, that show which kind of results they would like to obtain. The aim of the second meeting is to validate the statistical protocol proposed by the statistician according to the "dummy results" and the quality of available data.

The analysis is then performed, and the results are presented and explained at the beginning of the third meeting. The results are then discussed. A turnkey paragraph of the statistical analysis is written by the statistician in order to be inserted in the future publication.

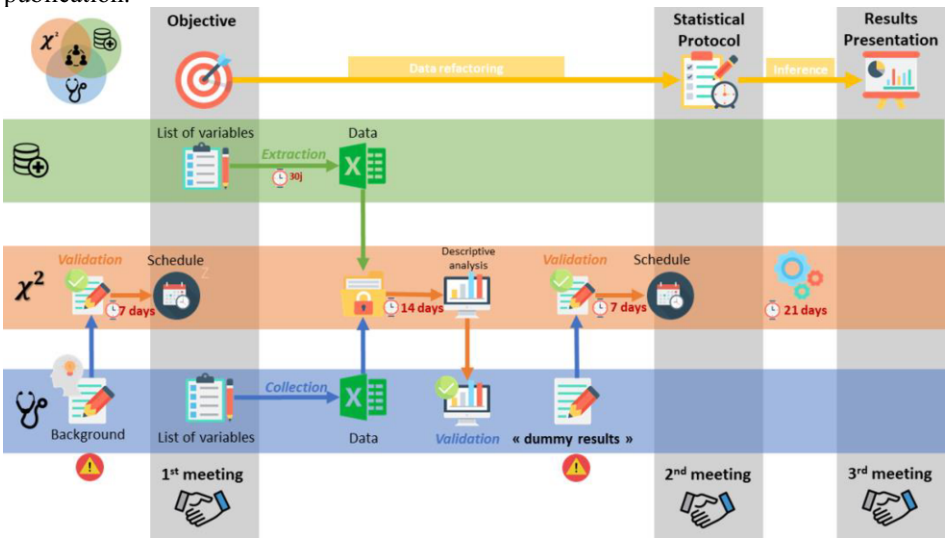

Figure 1. Timeline of the framework (from left to right). Tasks dedicated to the clinician, the statistician and the data scientist are represented in blue, red and green, respectively. Meeting are grayed.

2. Results

2.1. Projects currently underway

A test phase was carried out during 6 months from November 1, 2016 to April 31, 2017. The framework had been fully deployed since May 1, 2017. Table 1 details all the projects currently underway and their level of progress.

Table 1: Projects currently underway (October 31, 2017)

Progress	Number of projects
Background	4
Objective	6
Statistical Analysis	12
Communication/Publication	5

2.2. Study cases with extraction/analyses situations

As a result, we present some study cases with extraction or analysis situations, to illustrate the usefulness of the framework.

2.2.1. Emergency cesarean delivery and haemodynamic response under peridural anesthesia

During the first meeting, we defined the main objective of the study as the occurrence of hypotension after emergency cesarean section. This primary outcome was collected from the data warehouse following the reproducible methodology previously published by our multidisciplinary team [6]. On the one hand, operative data were extracted from the data warehouse, and on the other hand, medical history data were manually collected after delivery by the clinician. They were then merged by the computer scientist. The statistical analysis was then performed and the writing of the publication is now in progress.

2.2.2. Predictive factor of blood transfusion in liver resection

For this work, we intended to study predictors of blood transfusion in liver resection. During the third meeting, we decided to discard one of the secondary objectives. Indeed, after performing data extraction and descriptive analysis, the statistician argued that the occurrence of coelioscopy was too small to be used as a predictor, as initially hoped by the clinician.

3. Discussion

We described the framework developed to structure the operation of an anesthesia data warehouse for observational research purposes. After 6 months of full implementation, this framework enabled to increase data reuse efficiency by limiting the number of back loops. Despite stringency for the clinician, the acceptability was very good since delays were shortened, and quality of research was increased. That framework also enabled clinicians and statisticians to be aware of the complexity of the data extraction and management. Their participation in the process led to an empowerment process between all three actors, which increased efficiency of the workflow.

In a preliminary work, data extraction and management process was published by a multidisciplinary team of researchers. Implementation of this framework will keep encouraging collaborative publication in order to provide reproducible research evidence. Implementation of this framework resulted in the adoption of a unique shared folder between computer scientist and statistician. Collaborative documents increased the efficiency of the process. However, further work needs to be done since clinicians still don’t have access to the project management software and some documents still

are exchanged by mails. The set-up of a fully-shared workspace which avoids such exchanges is in progress. Full implementation of this framework will be possible when data from other information management systems (e.g. emergency, biology, etc.) will be integrated in the data warehouse.

References

[1] Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inf. 2017 Jun;102:21–8.
[2] Peng RD. Reproducible research in computational science. Science. 2011 Dec 2;334(6060):1226–7.
[3] Azevedo, Ana Isabel Rojão Lourenço and Santos, Manuel Filipe. KDD, SEMMA and CRISP-DM: a parallel overview. 2008.
[4] Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing. 2000.
[5] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007;14(1):1–9.
[6] Lamer A, Jeanne M, Marcilly R, Kipnis E, Schiro J, Logier R, et al. Methodology to automatically detect abnormal values of vital parameters in anesthesia time-series: Proposal for an adaptable algorithm. Comput Methods Programs Biomed. 2016 Jun;129:160–71.
[7] Lamer A, Jeanne M, Vallet B, Ditilyeu G, Delaby F, Tavernier B, et al. Development of an anesthesia data warehouse: Preliminary results. IRBM. 2013 Dec;34(6):376–8.
[8] Lamer A, Jeanne M, Ficheur G, Marcilly R. Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse. Stud Health Technol Inform. 2016;221:102–6.
[9] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet Lond Engl. 2007 Oct 20;370(9596):1453–7.