Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230103

Why Are Data Missing in Clinical Data Warehouses? A Simulation Study of How Data Are Processed (and Can Be Lost)

Sonia PRIOU[a,1], Guillaume LAME[a], Marija JANKOVIC[a], Gilles CHATELLIER[b], Romain BEY[c], Christophe TOURNIGAND[d], Christel DANIEL[c,e], and Emmanuelle KEMPF[d,e]

[a] *Université Paris-Saclay, CentraleSupélec, Laboratoire Génie Industriel, Fr*
[b] *Université de Paris, APHP-CUP, Department of Medical Informatics, Paris, Fr*
[c] *AP-HP, Innovation and Data, IT Department, Paris, Fr*
[d] *Université Paris Est Créteil, AP–HP, Department of medical Oncology, Henri Mondor and Albert Chenevier Teaching Hospital, Créteil*
[e] *Sorbonne Université, LIMICS, Paris, Fr*

Abstract. In recent years, the development of clinical data warehouses (CDW) has put Electronic Health Records (EHR) data in the spotlight. More and more innovative technologies for healthcare are based on these EHR data. However, quality assessments on EHR data are fundamental to gain confidence in the performances of new technologies. The infrastructure developed to access EHR data - CDW - can affect EHR data quality but its impact is difficult to measure. We conducted a simulation on the Assistance Publique – Hôpitaux de Paris (AP-HP) infrastructure to assess how a study on breast cancer care pathways could be affected by the complexity of the data flows between the AP-HP Hospital Information System, the CDW, and the analysis platform. A model of the data flows was developed. We retraced the flows of specific data elements for a simulated cohort of 1,000 patients. We estimated that 756 [743;770] and 423 [367;483] patients had all the data elements necessary to reconstruct the care pathway in the analysis platform in the "best case" scenarios (losses affect the same patients) and in a random distribution scenario (losses affect patients at random), respectively.

Keywords. Clinical data warehouse, EHR data, data quality, simulation

1. Introduction

Real-world data, and especially Electronic Health Records (EHR) data, are increasingly used to develop innovative digital technologies and new services supporting various activities of health professionals, in research, care and training. To integrate data from multiple sources and provide associated services for secondary use of this data, hospitals have recently started to develop Clinical Data Warehouses (CDW) [1]. However, EHR data are not without flaws, and caveats and recommendations have been raised when re-using them for research [2,3].

[1] Corresponding Author: Sonia PRIOU, Université Paris-Saclay, CentraleSupelec, Laboratoire Génie Industriel, Gif-sur-Yvette, France; Email: sonia.priou@centralesupelec.fr

A major aspect of re-using real-world data is data quality assessment. For EHR data, it doesn't just concern the data entered in the Hospital Information System (HIS). Issues can occur and data can be lost or somehow modified during the extraction, the transformation, or the loading phases, ultimately affecting the dataset provided to researchers [4,5]. This could lead to a loss of power or to systematic bias, by shifting or reversing the results of analyses. Since CDWs combine data from multiple pieces of software, many of them proprietary, they ultimately resemble black boxes, and tracing back how data were processed is hard. One can manually compare the data accessible in the CDW with the data in the HIS [6], but this procedure is time-consuming, and impossible when CDW data are de-identified.

The objective of this paper is to model and simulate the effect of the complex structure of a CDW on the data extracted from the HIS. To illustrate this simulation, we will place ourselves in the position of researchers trying to reconstruct breast cancer care pathways using data available in the Assistance Publique – Hôpitaux de Paris (AP-HP) CDW. By simulating a cohort of breast cancer patients treated at the AP-HP whose data is entered in the HIS, we estimate the proportion of the care pathways that can be fully reconstructed using data available in the analysis platform to researchers via the CDW.

2. Methods

2.1. The Hospital Information System

The AP-HP HIS is composed on a main EHR software and numerous specific software dedicated to specialised medical fields (e.g., imaging, pathology, laboratory tests results…). The main EHR software was not installed in all hospitals at the same time, leading to different EHR software being used across the AP-HP at first. As well, specific software varies from hospital to hospital and are prone to new versions. We conducted an inventory of which software was used each year for each hospital of the AP-HP. We hypothesized that, for a given hospital, the software used was consistent during the entire year and across all departments of the hospital.

The data entered in the HIS is composed of different elements: reports (consultation, hospitalisation, meetings…), laboratory results, procedure and diagnostic codes… Each of these data elements is entered in the HIS via either the main EHR software or one (or more) specific software, depending on both hospital and year.

2.2. Modeling care pathway reconstruction

We considered the following standard care pathway, common to all hospitals included in the study: first consultation of a hospital specialist, cancer diagnosis, therapeutic strategy choice, and administration of treatment. At each step of this care pathway, distinct data elements are recorded in the HIS by clinical staff.

We identified the key information needed to reconstruct the standard care pathway and selected the minimal dataset of data elements to find them (figure 1). To identify the first consultation related to the patient's cancer diagnosis, the date and the purpose of the visit are needed. Both are available in the consultation report. The cancer diagnosis and extension status are made by pathology examination and Positron Emission Tomography (PET) imaging, respectively. The reports give information about the diagnosis and the procedure date is linked to the procedure code. The choice of the therapeutic strategy is

made during the Multi-Disciplinary Meeting (MDM). The content and date of the MDM are available in the MDM report. Finally, the treatment administered is obtained through the following structured data: procedure and diagnostic codes, and laboratory results.

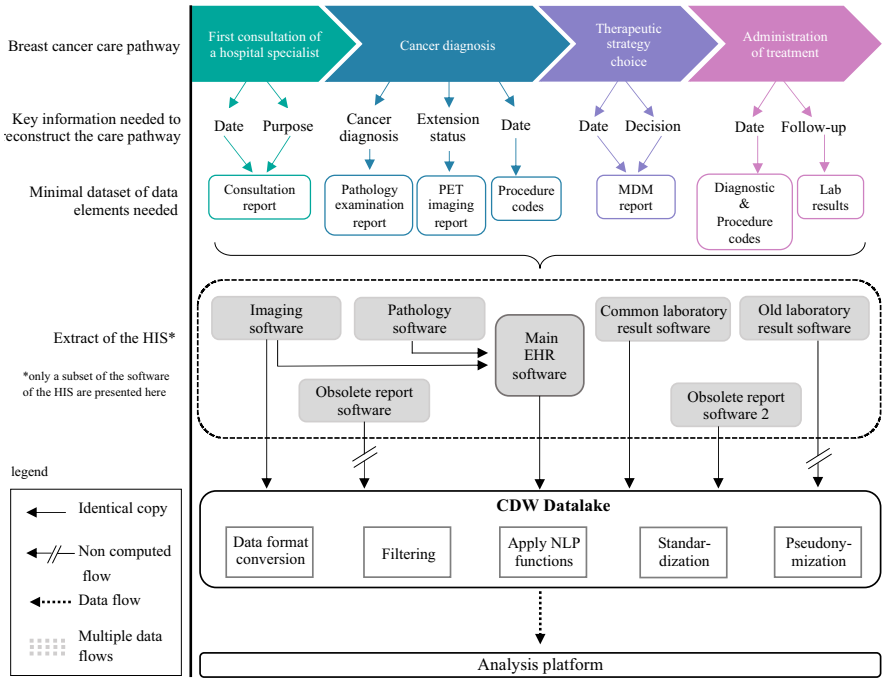

Figure 1. Simplified diagram of the data flow between the HIS and the analysis platform for the standard breast cancer care pathway

2.3. Modeling data processing

Thanks to the reading of internal technical documentation and the help from CDW professionals, we modelled the structure of the CDW. For each data element of the identified minimum dataset, we modelled its trajectory: entry in the HIS, transfer in the CDW and provision on the analysis platform. The trajectory followed by the data elements comprises several steps, between which data is transferred from one element to the next. Each data transfer was qualified according to the type of transformation applied: none, data format conversion, filtering, application of NLP functions, standardization and pseudonymization. Each of these steps was associated with a probability of correct data transfer: in case of impossible transfer, this probability was set to 0; in case of an identical copy, the probability was set to 1; for all other transformations, we modelled the probability of success using a uniform distribution between 0.95 and 1 (figure 1). A sensitivity analysis was performed on the lower bound of the uniform distribution.

2.4. Simulating a cohort of patients

We simulated a cohort of 1,000 patients treated at AP-HP for a breast cancer between 2018 and 2021. At AP-HP, breast cancer is mostly treated in 5 hospitals. One of these

hospitals uses a completely different EHR software than the other 4, so none of its data is integrated into the AP-HP CDW. For that reason, we decided to focus only on the other 4 hospitals, that we will name hospitals A, B, C, and D. We distributed our 1,000 patients across these 4 hospitals and per year following empirical distributions reproducing real proportions in the context of AP-HP. We hypothesised that patients stay in the same hospital during their care and that their entire care pathway occurs during the same year.

We used Monte-Carlo simulation on the success rate of data flows and simulated the success rate coefficients 10,000 times. Each time, we replayed the data trajectory to measure which data elements were available in the analysis platform. We measured the median and inter-quartile range of the percentage of patients for whom all the data elements necessary to rebuild the care pathway were available in two scenarios:

- the "best case" scenario, where all missing elements are concentrated on the same patients (i.e., the same subset of patients are missing their laboratory results, consultation reports, imaging reports…),
- the "random distribution" scenario, where data losses per data elements are randomly assigned to patients (i.e., missing laboratory results are attributed to a random subset of patients; missing consultation reports are attributed to another random subset on patients, possibly intersecting with the previous one).

3. Results

For all hospitals, the median success rate for transmitting the original data from the HIS, through the multiple layers of the CDW and to the analysis platform was 90% [88 – 92] for PET imaging and pathology reports, and 93% [91 – 94] for other reports. The success rate for transmitting procedure and diagnostic codes is 93% [91 – 94] cases. For hospitals using the same laboratory results software, the median success rate for transmitting information to the analysis platform reached 93% [91 – 94]. For hospital A using another laboratory result software, no laboratory results are available in the analysis platform (the data flow does not exist, see figure 1).

Out of 1000 patients, in the "best case" scenario, 756 [743;770] have all the data elements necessary to rebuild their care pathway. In the "random distribution" scenario, 423 [367;483] pathways can be reconstructed (figure 3). Without considering laboratory results, 904 [884;923] and 556 [493;635] pathways can be reconstructed in the "best case" and "random distribution" scenarios, respectively.

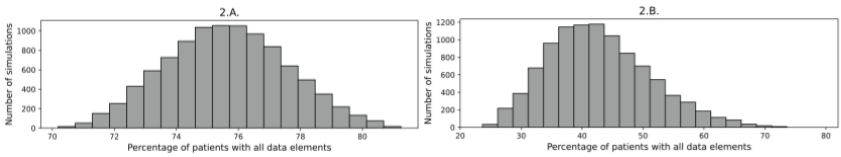

Figure 2. Percentage of pathways that can be fully reconstructed in the "best case" scenario (2.A) and the "random distribution" scenario (2.B)

The sensitivity analysis on the lower bound of the uniform distribution (between 0.90 and 0.99) shows a variation of the number of patients with all data elements available between 700 [673;726] and 804 [801;807], in the "best case" scenario, and 213 [160;282] and 717 [697;736] in the "random distribution" scenario (figure 3).

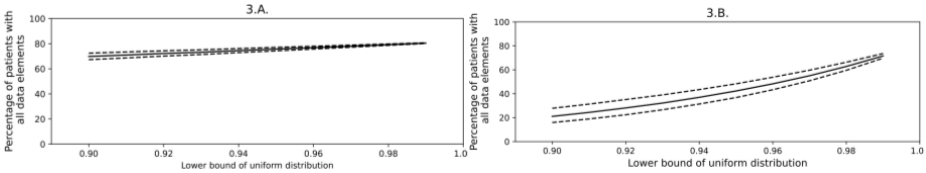

Figure 3. Sensitivity analysis on the lower bound coefficient of the uniform distribution in the "best case" Scenario (3.A) and the "random distribution" scenario (3.B)

4. Discussion and Conclusion

When working with data from a CDW, it is quite common to end up with data completeness issues where we did not expect to find any. This study highlights that, even when each data flow has a high success rate, the percentage of data available at the end can be surprisingly low. Each data element being entered in an independent software, the losses do not necessarily concern the same patients. In this configuration, the more data elements needed for a study, the more patients are going to be excluded because of missing data. More and more studies on data from CDW mention the ratio of patients for which a data element was found. As MDM are required by French law for all patients, one would expect to approach 100%. However, an AP-HP study on colorectal cancer showed that an MDM report was found for only 85% of patients [7].

This study enables to better understand the system behind CDW and the processes through which data are made available. CDW used for research purposes should not be "black boxes", as missing data may be source of biased or imprecise results. In this simulation, every data element was originally entered in the HIS. However, for real life data, the difficulty is not being able to explain the cause of the missing data. For example, is a report missing because it was not written, because it was never entered into the HIS or because of a malfunctioning data flow?

In conclusion, otherwise acceptable performance of individual applications translates to unacceptable overall performance when combined. The next step is to assess each level of data flow risk to better prioritize current and future data flows in the CDW.

References

[1] Tute E, Steiner J. Modeling of ETL-Processes and Processed Information in Clinical Data Warehousing.
[2] Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021 Mar 2;23(3):e22219.
[3] Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical Care. 2013 Aug;51(Supplement 8Suppl 3):S30–7.
[4] Ong T, Pradhananga R, Holve E, Kahn MG. A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation. 2017;5(1):16.
[5] Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018 May 29;20(5):e185.
[6] Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. International Journal of Medical Informatics. 2016 Oct;94:271–4.
[7] Kempf E, Priou S, Lamé G, Daniel C, Bellamine A, Sommacale D, et al. Impact of two waves of Sars-Cov2 outbreak on the number, clinical presentation, care trajectories and survival of patients newly referred for a colorectal cancer: A French multicentric cohort study from a large group of university hospitals. Intl Journal of Cancer. 2022 May 15;150(10):1609–18.