'Where have my patients gone?': A simulation study on real-world data processing in Clinical Data Warehouses

Sonia Priou, Emmanuelle Kempf, Rémi Flicoteaux, Marija Jankovic, Gilles Chatellier, Christophe Tournigand, Christel Daniel, Guillaume Lamé

Health Policy and Technology 13 (2024) 100893

Contents lists available at ScienceDirect

Health Policy and Technology

journal homepage: www.elsevier.com/locate/hlpt

Original Article/Research

'*Where have my patients gone?*': A simulation study on real-world data processing in Clinical Data Warehouses

Sonia Priou [a,*], Emmanuelle Kempf [b,c,#], Rémi Flicoteaux [d], Marija Jankovic [a], Gilles Chatellier [e], Christophe Tournigand [b], Christel Daniel [f], Guillaume Lamé [a,#]

[a] *Centrale Supélec, Laboratoire de Génie industriel, Gif sur Yvette, France*
[b] *Department of Medical Oncology, Henri Mondor & Albert Chenevier Teaching Hospital, Assistance Publique – Hôpitaux de Paris, Paris, France*
[c] *Sorbonne Université, LIMICS, Paris, France*
[d] *Département d'Information Médicale, Assistance Publique – Hôpitaux de Paris, Paris, France*
[e] *Georges Pompidou Teaching Hospital, Assistance Publique – Hôpitaux de Paris, Paris, France*
[f] *Assistance Publique – Hôpitaux de Paris, Paris, France*

ARTICLE INFO

Keywords:
Data warehousing
Routinely collected health data
Electronic health records
Computer simulation
Secondary data analysis

ABSTRACT

Objective: To access Electronic Health Record (EHR) data, hospitals have implemented Clinical Data Warehouses (CDWs) using Extract Transform and Load (ETL) processes. While ETL performances are typically evaluated individually, our study examines the cumulative impact of ETLs on data availability.
Methods: Using a real multi-hospital CDW as a case study, we modeled EHR data processing from the software sources to the CDW's data store. We simulated a scenario where researchers aimed to reconstruct breast cancer care trajectories using EHR data. We calculated the size and characteristics of the data store population, and compared them to the original population.
Results: EHR data are recorded in various software depending on data category, hospital, and year, each requiring specific series of ETLs for integration in the CDW. Despite acceptable transfer rates for each ETL (range 73 %-100 %), cumulative losses led to study populations in the data store being up to 90 % smaller than anticipated when researchers required data exhaustivity for patients. Population size decreased steeply with the more data categories required. No difference was found in population characteristics between the data store and the original cohorts.
Discussion & Conclusion: Researchers should scrutinize data availability in CDWs as missing data could result from outsourced care, incomplete input, or underperforming ETLs. Integrating more data sources in CDWs increases the number of data routes, necessitating time for ETL implementation and maintenance, and increases data loss risks. Though commonly perceived as a "black box", data transformation can significantly influence the reliability of populations studied in CDWs.
Public interest Summary: To access data generated during care, researchers build Clinical Data Warehouses (CDWs). CDWs are infrastructures composed of a series of processing steps to extract the data from the data source, transform it according to the needs and load it into a data store. Usually, the performances of these processing steps are evaluated one a time. However, each data point goes through a series of processing steps before being made available for research. In this study, we aim to evaluate the impact of the entire data processing pipeline on the availability of data points in a CDW by simulating a study on breast cancer and evaluating the impact on the size and the characteristics of the final cohort. The cumulative losses of the processing steps resulted in a population 90 % smaller than anticipated. The characteristics of the final population showed no difference to those of the original cohort.

* Corresponding author: Centrale Supélec, Laboratoire de Génie industriel, 3 rue Joliot Curie, 91190 Gif-sur-Yvette, France.
E-mail address: sonia.priou@centralesupelec.fr (S. Priou).
# Guillaume Lamé and Emmanuelle Kempf contributed equally to this work.

https://doi.org/10.1016/j.hlpt.2024.100893

Available online 2 August 2024

Background and objective

Hospitals are increasingly building Clinical Data Warehouses (CDWs) to access clinical data (e.g., Electronic Health Record—EHR—data) for other purposes than patient care. A CDW is an infrastructure that collects healthcare data (including patient demographics, claims data, laboratory tests, medication and clinical notes) from various data sources (i.e., software applications of Hospital Information Systems (HIS)). Then data undergoes transformation such as deduplication, standardization, pseudonymization resulting in the creation of a subpopulation dataset in a target data store for researchers to access and analyze (Fig. 1) [1]. Like any data warehouse, CDWs rely on Extract Transform and Load (ETL) processes. A series of ETL processes converts the structure and semantics of the data from diverse sources into the target store (glossary available in Supplementary File). However, CDWs are particularly complex data warehouses, as EHRs come in various formats that are difficult to integrate [2–5]. The data that researchers analyze in the target store of a CDW are influenced by the ETL processes in place [6], and flawed ETLs can result in faulty data, potentially leading researchers to incorrect conclusions [7]. Therefore, the validity of a CDW relies on the validity of its ETL processes. However, most studies focus on biases caused by the way data is recorded and re-used beyond its original purposes [8,9], rather than examining the impact of data processing within the CDW infrastructure itself.

Engineers usually assess the performance of ETLs one at a time. Data quality checks aim at detecting potential violations of syntactic or semantic properties, while balancing tests compare the data before and after the ETL process [10]. These tests are time-consuming and predicting failure modes is challenging, so only a limited number of data points can be manually checked [11–13]. In addition, these tests are performed at the ETL level, despite the fact that data often undergoes multiple stages of processing from the HIS to the target data store. Consequently, what may have seemed like a good performance at the individual ETL level may not seem as good when multiplied across all ETL processes.

End-users often perceive infrastructure as 'boring things' [14] and many data researchers may be tempted to regard CDWs as black-boxes systems and let CDW engineers deal with ETL issues. They should rather consider whether the ETL infrastructure influences research findings. In this article, we open the black box, asking these questions: how much data is lost between the point where healthcare professionals enter data in a software, and their utilization by researchers in the target store of a CDW? What impact does this have on the size and characteristics of the final cohort? (Fig. 1) The objective is to evaluate the impact of the entire CDW ETL infrastructure on the availability of data in the target store of a CDW. To answer this question, we simulated the data processing by the ETL infrastructure of a complex, multicenter CDW and applied it to a realistic study on breast cancer patient care pathway.

Materials and methods

We adapted the Dependency Structure Modelling (DSM) formalism from complex systems engineering [15] to build a simulation model of the routes of each data category from its origin software to the target data store in a CDW. DSM represents complex systems (cars, planes, information systems…) as a set of components connected by interfaces, in matrices. Each row or column of the matrix represents a component, and each intersection represents the interface between these two components. Standard matrix algebra (additions, multiplications) can then help understanding how information propagates through interfaces in the system. In our case, each row or column of the DSM represents a component of the CDW, and the matrix contains the transfer rate between these components, i.e. the performance of the ETL linking them.

We instantiated this model to simulate the ETLs process of the Assistance Publique - Hôpitaux de Paris's (AP-HP) CDW. AP-HP is a network of 38 university hospitals in Paris region, which makes it a complex real-world example that can illustrate the integration of data elements of different categories (in this study, restricted to patient demographics, claims data, clinical reports, and laboratory tests) from various data sources frequently impacted by software changes, and variations of information systems between hospitals. Having modelled the CDW, we simulated data availability in the target data store on a scenario where researchers want to reconstruct initial breast cancer patient care trajectories. We compared the original cohort composed of all patients available in the HIS with the final cohort, composed of the patients that have all the data categories needed for the study available in the target data store of the CDW (Fig. 1).

A dependency structure model of data routes

We considered a multi-software HIS from which data of different categories (patient demographics, claims data, clinical reports, and laboratory tests) are extracted, transformed, and integrated into a CDW. Researchers can then access the data in the target data store (Fig. 1). We retraced the route of data categories from each source to the target data

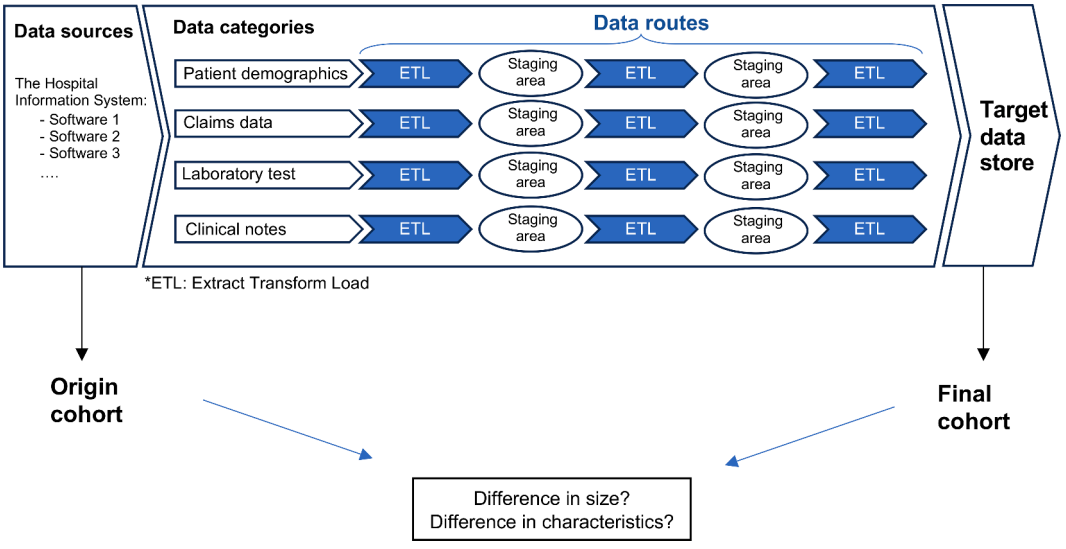

Fig. 1. Data route process between the data sources and the target data store for four data categories.

store for each piece of software of the HIS. Each ETL was qualified according to the type of transformation applied: extraction, data format conversion, feature extraction, sanity checks, standardization, and pseudonymization. In this model, we considered that the data available in the target data store was formatted to fit the Observational Medical Outcome Partnership data model (OMOP v5.0) [16].

To model data loss, we characterized the performances of each ETL by a success rate φ_{ETL}. The success rate is a continuous value between 0 (no data transferred to the next staging area) and 1 (all data transferred to the next staging area). We constructed a dependency structure matrix [15] with the data sources and data staging areas labelling the horizontal and vertical axes. The i^{th} row and j^{th} column correspond to the success rate $\varphi_{i,j}$ of the ETL extracting data from the label of column i and loading it to the label of column j (Fig. 2.a.). For a given data source, the global success rate of the corresponding data route φ_{route}corresponds to the multiplication of the success rates of all ETLs on the route. It was calculated by multiplying a 1-dimensional vector composed of zeros and one 1 located in the data source column by the dependency structure matrix powered to the length of the data route. This translates to the following expression:

$$\varphi_{route} = \prod_{ETLs\ on\ route} \varphi_{ETL}$$

The model was implemented using Python 3.9.

Evaluation of data availability

We considered a cohort of *p* patients for which *m* data categories are needed in the target data store by researchers. Using the dependency model, we calculated the global success rate of the data routes for each data category needed. For each patient, we built a random 1-dimensional vector size *m* with coefficients drawn from a uniform distribution over the interval [0;1]. Then, we compared the random vector to the vector of global success rates. If the draw was lower than the global success rate, then the data category was available in the target data store; otherwise, it was not available (Fig. 2.b.). We used a Monte-Carlo simulation on the random vectors and simulated 1000 times the availability of the data categories per patient in the target data store.

As we wished to evaluate the impact of missing data on the cohort, we excluded patients with at least one missing data category in the

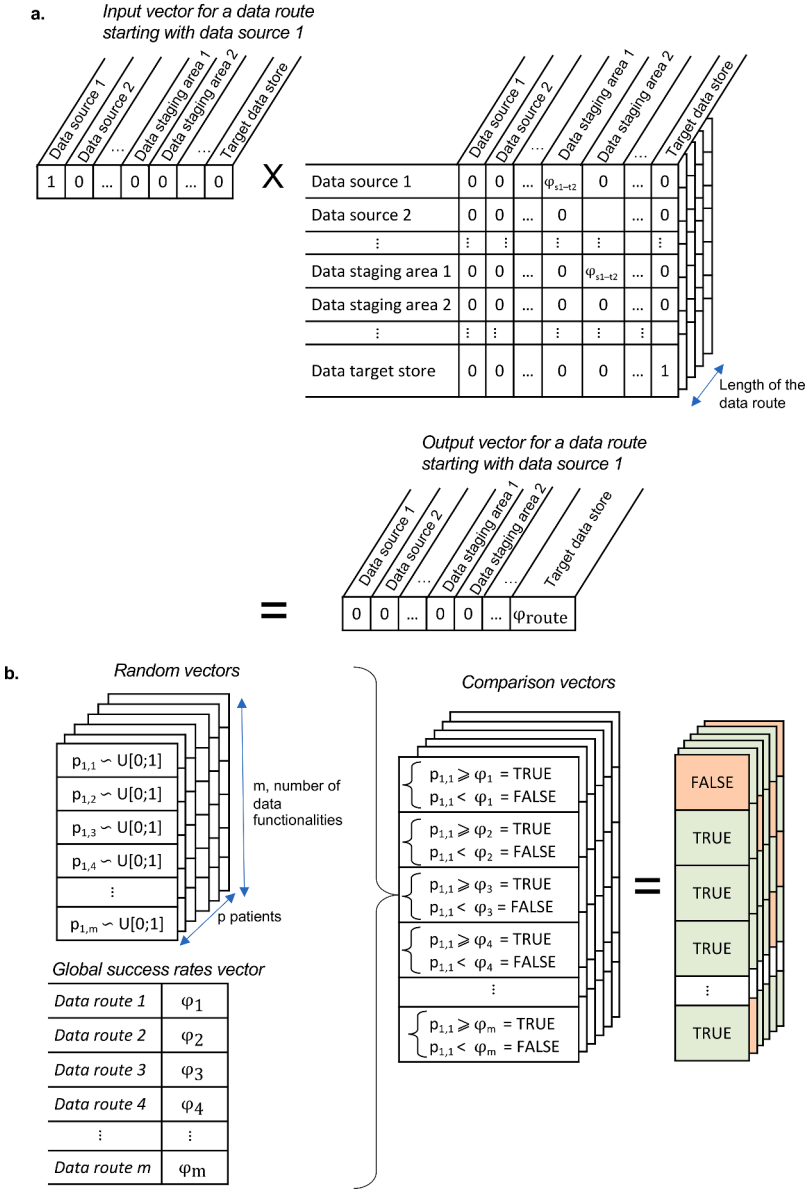

Fig. 2. Matrix calculation to evaluate the global success rate of data routes (a) and the number of data categories available in the target data store per patient (b).

target data store from the final cohort of our study. To evaluate the impact of the data routes on the size of the final cohort, we computed the median and interquartile range (IQR) of the number of patients for whom all the data categories were available in the target data store.

Then, to evaluate the impact of the data routes on the characteristics of the final cohort, we compared the characteristics of the origin cohort to those of the final cohort in the target data store using a Chi2 test for categorical data values and a z-test for continuous data values.

Finally, we assessed how the size of the final cohort varied according to the number of data categories required by researchers. We computed the median and interquartile range of the size of the final cohort for all possible combinations of data categories.

The AP-HP model

AP-HP is composed of 38 hospitals. Its HIS includes an integrated EHR software and specific software dedicated to information management in ancillary departments (e.g., radiology, medical biology, pathology, or genetics pathology). In 2012, AP-HP started to deploy a new integrated EHR software in all its hospitals. This new software enables the collection of different categories of data (e.g., patient demographic information, visits, claims data, and clinical reports) that were previously handled by specific solutions in the various hospitals. From that date, all hospitals and departments have progressively switched from their previous software solutions to this integrated EHR solution. However, the deployment has been slow and phased across hospitals and across departments and data categories inside each hospital [17]. Therefore, the new integrated EHR and the outdated solutions coexisted. After discussing with experts, we proposed to consider that, for each category of data, the switch to the integrated EHR software was made instantly at the beginning of a year. All software in the HIS are considered potential data sources for the CDW. We created an overall representation of the architecture of this information system.

Depending on the hospital and the year, the same data category can be input in the HIS through several data sources. We listed the software used by each hospital for every data category, for each year between 2016 and 2022, including identifying data routes that were not implemented in the CDW (See Supplementary Table 1).

Based on internal technical documents and calculations performed by AP-HP's data experts, where possible, we estimated the success rate of each ETL. When the information was not available, we used success rates from the literature. For example, the success rate of free text pseudonymization corresponds to the precision rate of the pseudonymization algorithm used at AP-HP [18].

Use case of new breast cancer pathways

We simulated a research project reconstructing care pathways for newly referred breast cancer patients. At AP-HP, breast cancer is mostly treated in five hospitals. One of these hospitals uses a completely different EHR software than the other four, and none of its data is integrated into the AP-HP CDW. For this reason, we decided to focus only on the other four hospitals, which we will name hospital A, B, C, and D. The study population included newly referred breast cancer patient at one of these four hospitals between January 2016 and December 2022 (origin cohort). We hypothesized that patients stayed in the same hospital during their care and that a patient's entire care pathway occurred in the same calendar year than their diagnosis.

The characteristics of the origin cohort were computed using aggregated indicators at the hospital level, routinely computed for activity monitoring purposes. We considered patient characteristics (age, gender, social deprivation, Charlson comorbidity index and the Elixhauser comorbidity index) and hospital visit characteristics (length of a hospital visit, visits to the emergency room, severity of the visits and visits in cancer-specialized departments). Patients' social deprivation and visits' severity level are computed according to the French national guidelines [19] using International Classification of Diseases 10th edition coding [20]. We compared the percentage of patients with visits with severity level of three or four to those with visits of severity levels strictly below 3. The global population percentage was calculated for categorial values (gender, patient over 75 years old, social deprivation, visit in emergency room, severity of the visit above ¾, and visits in cancer-specialized departments). The weighted mean and standard deviation were calculated for quantitative values (Charlson comorbidity index, Elixhauser comorbidity index, and the length of a hospital visit).

With medical experts, we identified the main steps in breast cancer patient care and the features (e.g., dates, characteristics, and treatments) needed in each step to perform trajectory analysis. Then, we pinpointed the specific data categories required to extract each feature. As we wished to evaluate the effect of ETLs and not the quality of data input by clinicians, we hypothesized that for all newly referred breast cancer patients, the entire care pathway was conducted at AP-HP and that all the data categories were input into the HIS. We identified data routes from the HIS to the CDW for each data category required and calculated their global success rate. Since data input by clinicians is not the main concern of this study, we considered that all the data categories needed for the breast cancer care pathway research were available in the HIS for the origin cohort. We evaluated the proportion of patients from the origin cohort for whom all data categories were still available in the CDW's data store. This defined a second population: the final cohort composed of the patients that could be included in the breast cancer care pathway study performed on the CDW's data store.

We calculated the size of the final population, and we computed its characteristics using the hospital level aggregated indicators of the origin cohort weighted on the size of the final population for each hospital and each year. The characteristics of the origin cohort and the final cohort were compared using a χ^2 test for categorical values and a *t*-test for quantitative values.

Uncertainty quantification and sensitivity analysis

We evaluated the uncertainty of the median percentage of patients in the final cohort with respect to the uncertainty of the success rates φ_{ETL}. We modelled the uncertainty of success rates estimated by AP-HP experts or found in the literature by a symmetric triangular distribution [21] with a mode at φ_{ETL}, a lower bound set to l, and an upper bound set to L (Supp Table 2), with:

$$l = \varphi_{ETL} \times \left(1 - \min\left(0.10;\ \frac{1-\varphi_{ETL}}{\varphi_{ETL}}\right)\right)$$

$$L = \varphi_{ETL} \times \left(1 + \min\left(0.10;\ \frac{1-\varphi_{ETL}}{\varphi_{ETL}}\right)\right)$$

For uncertainty estimation, we performed a Monte Carlo Simulation on the success rate of the ETLs and simulated them 1000 times and calculated the variance of the 1000 outputs.

Secondly, we conducted sensitivity analysis using Sobol's method [22] to identify which data sources had the most impact on the median percentage of patients with all data categories available in the target data store. We only calculated Sobol's first-order and total-order indices to limit computing time. The first-order indices evaluate the impact of each success rate of ETLs individually by calculating the contribution of each success rate to the variance of the output of the model. The total-order indices evaluate the importance of one success rate and its relation with other success rates on the output. The success rates with the higher Sobol's indices have the most impact. We performed a Monte Carlo Simulation on the success rate of the ETLs. We evaluated the model *100 x (number of success rates + 2)* times [23]. We used Python's SALib library [24,25] to calculate Sobol's indices in which the total-order indices are estimated [26]. If the estimators of the total-order indices were negative with their confidence interval overlapping zero, then we treated them as zero.

Results

Models of the CDW and the patient care pathway

Each data category can be input in the AP-HP HIS through various software (the integrated EHR software, an outdated software, a new specific software), depending on the hospital and the year. The general structure of data processing in the CDW is shown in Fig. 3. A data route is composed of several ETLs, whose performances for each type of ETL are estimated in Table 1. The detailed data routes per hospital and per year are described in Supplementary Table 1.

Each data category has a specific data route from the HIS to the target data store depending on its data source. The success rate of a data route depends on the combination of ETLs on the route. For example, the global success rates for imaging reports are not the same if the report was input in the imaging software or in the outdated imaging software (Table 2). The global success rate of data routes varies between 0.69 and 0.99 for data sources integrated in the CDW (Table 2).

An epidemiology study on breast cancer patient trajectories in hospitals would identify five main steps in the patient's care pathway. Data sources for each feature needed to analyze the care pathway are presented in Fig. 4.

The final cohort

Data availability in the target data store

The median percentage of patients from the origin cohort that have all the data categories needed for the study in the target data store is 11.3 % [IQR 11.1–11.5]. This percentage varies by hospital and by year, depending on which software was used at the time (Fig. 5). None of the patients from hospital D are included in the final cohort as all five laboratory results are never available (Supp Table 3). If we remove the need for laboratory results in our epidemiology study, then 20.3 % (IQR; 20.0 – 20.5) of the origin cohort would be available for the study in the target data store.

To obtain a larger final cohort, researchers can choose to restrict the number of data categories needed. Depending on which data categories are needed, the median percentage of patients with all data categories available in the target data store can vary (Fig. 6). For example, if a researcher requires only three data categories, the percentage of patients available varies from 26.6 % to 50.6 % depending on which triplet of data categories is required.

The final cohort is also impacted by the choice not to integrate outdated software. If we restrict the origin cohort to patients referred to AP-HP between 2020 and 2022, a period where more hospitals had switched to the integrated EHR software, the final cohort represents 16.7% [IQR 16.4 – 17.0] of the origin cohort (27.9 % [IQR 27.5 – 28.3] if laboratory results are excluded).

Characteristics of the origin and final cohorts

The characteristics of the final cohort did not differ from the origin cohort (Table 3). Apart from reducing the size of the population in the final cohort, there are no significant differences between the two cohorts.

Uncertainty and sensitivity analysis

The triangular distributions used to model the uncertainty of the success rates of each ETL are presented in Supplementary Table 2. After 1000 simulations, the median percentage of patients with all data categories available in the target data store varies between 5.7 % and 20.5 %, with a variance of 2.4 % (Supp Figure 1).

To calculate the Sobol's indices for the 14 success rates, 2000 simulations of the model were computed. (Supp Figure 2). We observe that first-order indices for the success rates of extraction of measurements and standardization to OMOP of measurement are the highest. These two ETLs impact the most the variability of our model. As the total-order indices are larger than the first-order indices (Supp Figure 2), there may be interactions occurring between success rates that influence the variance of the output of our model. This was to be expected, as ETLs are chained on data routes. Our model is a complex system where all ETLs matter in the global performances. No particular ETL is responsible for a

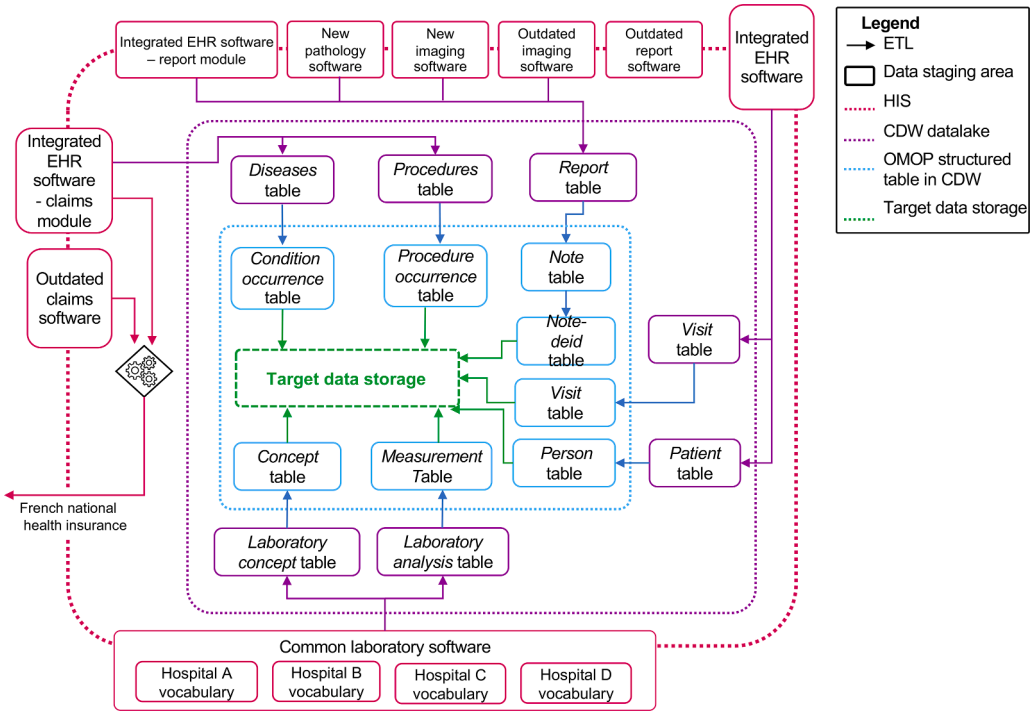

Fig. 3. Model of the data routes of a subset of data sources from the HIS to the target data store restricted to patient information, claims data, clinical reports, pathology analysis, imaging, and biology.

Table 1
Qualification of ETLs for each data category.

Data categories	Data staging area origin	Data staging area destination	Type of transformation	success rate $n_{transfer}$	Estimation method
Consultation report	Integrated EHR software – report	*Report* table	Extraction of consultation reports	0.99	Experts
MDM[a] reports	Integrated EHR software – report	*Report* table	Extraction of MDM reports	0.93	Experts
Imaging report	Imaging software	*Report* table	Extraction of imaging report	0.97	Experts
	Outdated imaging software	*Report* table	Extraction of imaging report	0.89	Experts
Pathology report	Pathology software	*Report* table	Extraction of pathology reports	0.96	Experts
All reports	*Report* table	*Note* table	Transformation to free text	0.97	Experts
			Classification of document with feature extraction	0.85	Experts
			Standardization to OMOP	1.00	Literature [1]
	Note table	*Note- deid* table	Pseudonymization via NLP[b]	0.94	Literature [2]
Claims data	Integrated EHR software – claims	*Diseases & Procedures* tables	Extraction of claims data	1.00	Experts
	*Diseases t*able	*Condition occurrence* table	Standardization to OMOP	0.90	Literature [1]
	*Procedure t*able	*Procedure occurrence* table	Standardization to OMOP	0.99	Literature [1]
Laboratory results	Integrated laboratory software	*Laboratory analysis* table	Extraction of measurements	1.00	Experts
	Laboratory analysis table	*Measurement* table	Standardization to OMOP	0.73	Literature [1]

[a] MDM: Multi-Disciplinary Meeting.
[b] NLP: Natural Language Processing.

Table 2
Global success rates of data routes depending on the data categories and the data source.

Data categories	Data source	Global success rate
Consultation report	Integrated EHR software	0.77
	Outdated report software	0.00
MDM report	Integrated EHR software	0.72
	Outdated report software	0.00
Imaging report	Imaging software	0.75
	Outdated imaging software	0.69
Pathology report	Pathology software	0.74
Claims data - diseases	Integrated EHR software	0.90
	Outdated claims software	0.00
Claims data - procedures	Integrated EHR software	0.99
	Outdated claims software	0.00
Laboratory results (albumin, hemoglobin, leukocytes, platelets, and bilirubin)	Common laboratory software	0.73

good or a poor outcome. To improve the global outcome, all ETLs need to be considered.

Discussion

In this paper, we modelled the data routes of different data categories from the data source (HIS) to a target data store in AP-HP's CDW. Each data category follows its own data route composed of several ETLs. Most ETLs have transfer rates that seem acceptable. However, during the route from the HIS to the CDW, the losses of each ETL multiply and, in the end, the success rate of a data route is low. This leads to much smaller cohorts than anticipated, up to 89 % smaller in our example study if researchers want data exhaustivity. However, there seem to be no statistical differences in patient and hospital visit characteristics of the final cohort analyzed by researchers compared to the original cohort.

The HIS is composed of multiple independent data sources. This specificity leads to two issues. First, for each patient, their data of different categories are spread across multiple independent data routes. One faulty route can lead to a data category being unavailable in the target data store and consequently the patient not being included in the final cohort for the epidemiology study. The more data routes, the more likely that at least one data category won't be available in the target data store. Secondly, multiple data sources entail multiple ETLs. As ETLs are very time-consuming to implement, it is possible that choices are made, and some outdated data sources are not integrated into the CDW. The difference in the availability of historical data across hospitals can have a real impact on the studied population. CDWs are sometimes presented as a leap forward in observational research. Indeed, with numerous data sources made available in one data store, it is quite normal to be greedy and aspire to complex epidemiology studies with multi-source data. However, as the number of data categories (and consequently data sources) goes up, the size of the final cohort quickly goes down.

Technical challenges of implementing ETLs are often overlooked when discussing data quality but have become more and more of a concern in recent years. A qualitative study classified the difficulties of ETL implementation according to three themes: challenges linked to the source data, the technical difficulties, and the knowledge generation, recording and maintenance [27]. Mapping local vocabulary to common vocabulary is an essential task, usually done manually, which questions its sustainability in the long run [28]. As EHR systems can be very complex and their configuration can vary over time, implementing ETLs to access EHR data can be challenging [29]. The high variability between pieces of software regarding data storage and format adds to the difficulty of developing ETL algorithms and can lead to incomplete datasets [30]. Researchers need to understand the operational workflows that enable the data to be accessible in order to interpret it correctly [29]. The lack of standardized ETL modelling is a barrier in making end-users understand the operational workflow [31].

This study enables researchers to understand better how clinical data generated for patient care are made available in a CDW and how the ETL process can impact data availability. The use of a simulation model enabled us to compute different scenarios, compare large populations and perform a sensitivity analysis. Data processing is often seen as a black box, which we tried to open using a simulation model based on parameters estimated from real data. Data experts from AP-HP were involved in the design of the data routes and the estimation of the parameters. Simulation is a key methodology in the study of complex systems, with multiple interacting components [32]. We based our model on a multisite CDW, fed by multiple software. Our study also

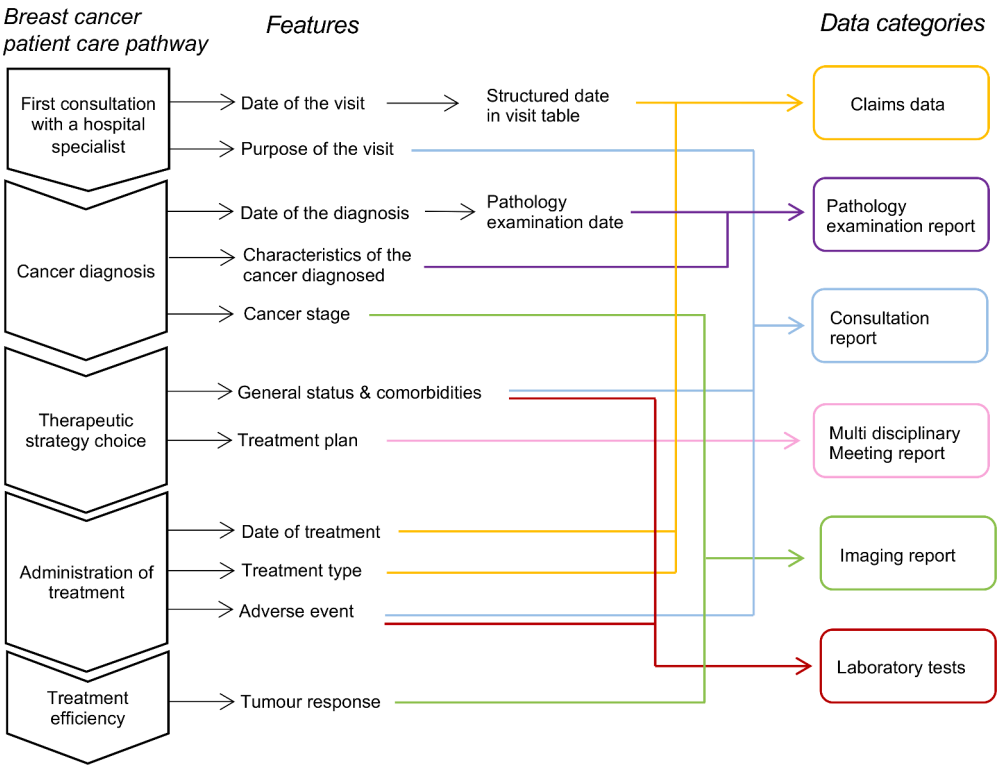

Fig. 4. Features and data categories needed to identify the steps in the care pathway of patients with breast cancer.

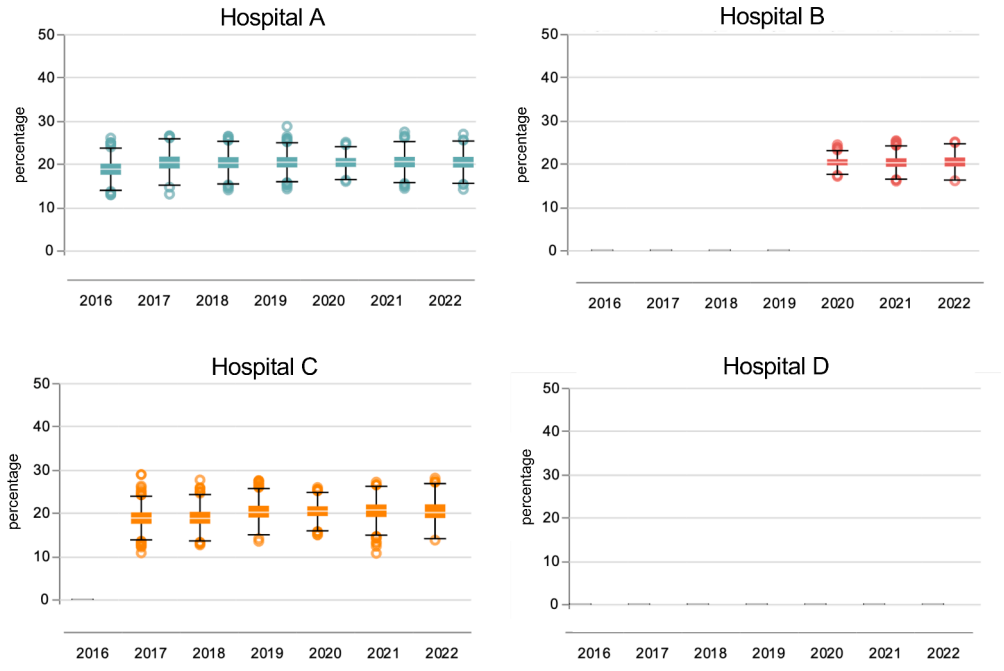

Fig. 5. Box plot of the percentage of patients with all the data categories available in the target data store per hospital per year. *Hospital D is set to 0 % every year as its laboratory results are not integrated into the CDW.*

takes account of longitudinal dynamics. Yet, our model remains simple and easily understandable.

However, this model has limits. First, this simulation is based on a simplified model of the integration and processing system of clinical data at AP-HP. We considered that data routes from the HIS to the target data store were independent, where, in reality, interactions between sub-tables are necessary, especially for data standardization. We did not consider the fact that data can be edited in the HIS after it has been processed to the CDW. We decided that the switch of each software would occur at the beginning of a year for an entire hospital. Reality is more complicated, with deployment happening per department and with an overlap period during which both pieces of software can be used simultaneously. This can lead to difficulties in determining the precise date of deployment [17]. Thirdly, the success rates of ETLs are estimated

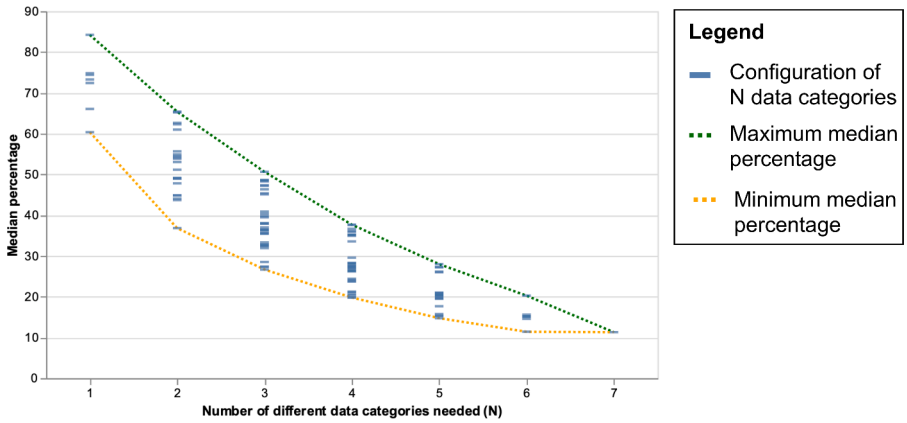

Fig. 6. Median percentage of patients with all data categories available in the target data store depending on the number of data categories needed (N).

Table 3
Characteristics of the origin cohort and of the final cohort.

	Origin cohort	Final cohort	p-value
Patients' characteristics			
Population, N	14,200	1602	
Sex ratio, F (%)	99.10	98.69	0.14
Patient over 75 years old (%)	15.56	15.33	0.84
Precarious status (%)	8.79	9.13	0.69
Charlson Comorbidity Index (mean [sd[a]])	4.55 [3.16]	4.69 [1.07]	0.07
Elixhauser Comorbidity Index (mean [sd])	9.15 [7.10]	9.32 [2.39]	0.33
Visits' characteristics			
Length of a hospital visit (days) (mean [sd])	7.37 [10.74]	6.93 [3.70]	0.10
Visits to the emergency room, N (%)	14.56	12.93	0.09
Severity of the visits (> 3/4), N (%)	15.14	14.94	0.86
Visits in cancer-specialized departments, N (%)	96.91	96.43	0.33

[a] Standard deviation.

using values found in the literature or by internal analysis performed by experts at AP-HP. We modelled uncertainty with triangular distributions and 10 % spread from the average, but this choice could deserve more investigation. Our model only considers data availability, but ignores data quality issues related to HIS data input [33] and the difficulty of extracting information from clinical reports., through NLP [34]. Finally, we took a really crude approach to missing data, by excluding patients with missing data. Various methods exist to tackle this issue [35], and we also ignored the fact that the same information can sometimes be obtained from several sources [36]. Accounting for missing data management methods and data redundancy would have increased the size of the final cohort.

The results of this study point out the importance for data users to understand how data is made available to them. For multi-site CDWs, researchers should question data availability per site. A sanity check of cohort sizes should always be performed to make sure no issue in data integration was missed. Finally, researchers should carefully weigh the number of data categories they wish to work on. This research focuses primarily on AP-HP's CDW. As it is a multi-hospital network, one can imagine translating this research to other complex structures such as multi-CDW structures. The variability in data sources when combining data from different infrastructures has been shown to generate new challenges [37]. Each CDW has its own specificities with its own data routes. An epidemiology study based on a multi-CDW network should analyze data availability before starting the analysis.

Conclusion

When using clinical data in a CDW, researchers should question the completeness of the data available. Missing data can be a result of outsourced or out-of-hospital care, incomplete data input in the HIS, or faulty ETLs from the HIS to the CDW. In the two first situations, researchers re-using healthcare data generated for care are mostly aware of these difficulties. The third situation is much less considered, as data transformation is seen as a "black box". Yet it deserves attention if we are to produce reliable results from CDWs.

Funding

None.

Ethical approval

Not required.

Acknowledgments

The authors wish to thank Perceval Wajsbürt, Christel Gérardin, Ariel Cohen, Alice Calliger, Adam Remaki and Véronique Letort-Le Chevalier for their support.

Declaration of competing interest

None declared.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.hlpt.2024.100893.

References

[1] Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: a case study in France. PLOS Digit Health 2023;2(7):e0000298.
[2] Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:1–9.
[3] Inmon B. Data warehousing in a healthcare environment. The Data Administration Newsletter; 2007.
[4] Callahan A, Shah NH, Chen JH. Research and reporting considerations for observational studies using electronic health record data. Ann Intern Med 2020;172(11_Supplement):S79–84.
[5] Khalaf Hamoud A, Salah Hashim A, Akeel Awadh W. Clinical data warehouse: a review. Iraqi J Comput Inform 2018;44(2).
[6] Rijnbeek PR. Converting to a common data model: what is lost in translation?: Commentary on "Fidelity assessment of a clinical practice research datalink conversion to the omop common data model". Drug Saf 2014;37(11):893–6.

[7] Homayouni H, Ghosh S, Ray I. An approach for testing the extract-transform-load process in data warehouse systems. In: Proceedings of the 22nd international database engineering & applications symposium; 2018. p. 236–45.
[8] Ni K, Chu H, Zeng L, Li N, Zhao Y. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. BMJ Open 2019;9(7):e029314.
[9] Madandola OO, Bjarnadottir RI, Yao Y, Ansell M, Dos Santos F, Cho H, et al. The relationship between electronic health records user interface features and data quality of patient clinical information: an integrative review. J Am Med Inform Assoc 2023:ocad188.
[10] Homayouni H. Testing extract-transform-load process in data warehouse systems. In: International symposium on software reliability engineering workshops; 2018. p. 158–61.
[11] Quiroz JC, Chard T, Sa Z, Ritchie A, Jorm L, Gallego B. Extract, transform, load framework for the conversion of health databases to OMOP. Deserno TM, editor. PLoS ONE 2022;17(4):e0266911.
[12] Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform 2016;94:271–4.
[13] Golfarelli M, Rizzi S. Data warehouse testing: a prototype-based methodology. Inf Softw Technol 2011;53(11):1183–98.
[14] Star SL. Infrastructure and ethnographic practice: working on the fringes. Scand J Inf Syst 2002;14(2):6.
[15] Lindemann U, Maurer M, Braun T. Structural complexity management: an approach for the field of product design. Springer Science & Business Media; 2008. p. 247.
[16] OHDSI – Observational Health Data Sciences and Informatics [Internet]. [cited 2023 Oct 2]. Available from: https://www.ohdsi.org/.
[17] Remaki A., Playe B., Bernard P., Vittoz S., Doutreligne M., Chatelier G., et al. Adjusting for the progressive digitization of health records: working examples on a multi-hospital clinical data warehouse. medRxiv. 2023 Aug 21;.
[18] Tannier X, Wajsbürt P, Calliger A, Dura B, Mouchet A, Hilka M, et al. Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. Methods Inf Med 2024: s–0044-1778693.
[19] Guide Méthodologique de Production des Informations Relatives à l'Activité Médicale et à sa Facturation en Médecine, Chirurgie, Obstétrique et Odontologie [Internet]. [cited 2023 Nov 28]. Available from: https://www.atih.sante.fr/sites/default/files/public/content/4219/guide_methodo_mco_2022_6_bis_version_provisoire_2.pdf.
[20] World Health Organization. International statistical classification of diseases and related health problems. 10th ed. 2019 [Internet]. [cited 2023 Nov 28]. Available from: https://icd.who.int/browse10/2019/en.
[21] Johnson D. The triangular distribution as a proxy for the beta distribution in risk analysis. J R Stat Soc Series D 1997;46(3):387–98.
[22] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 2001;55(1–3):271–80.
[23] Pianosi F, Wagener T. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environ Modell Softw 2015;67:1–11.
[24] Iwanaga T, Usher W, Herman J. Toward SALib 2.0: advancing the accessibility and interpretability of global sensitivity analyses. Socio-Environ Syst Modell 2022;4. 18155–18155.
[25] Herman J, Usher W. SALib: an open-source python library for sensitivity analysis. J Open Source Softw 2017;2(9):97.
[26] Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput Phys Commun 2010.
[27] Ong T, Pradhananga R, Holve E, Kahn MG. A framework for classification of electronic health data extraction-transformation-loading challenges in data network participation. EGEMS 2017;5(1):16.
[28] Oja M, Tamm S, Mooses K, Pajusalu M, Talvik HA, Ott A, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. JAMIA Open 2023;6(4):ooad100.
[29] Holmes JH, Beinlich J, Boland MR, Bowles KH, Chen Y, Cook TS, et al. Why is the electronic health record so challenging for research and clinical care? Methods Inf Med 2021;60(01/02):032–48.
[30] Ferrão J, Oliveira M, Janela F, Martins H. Preprocessing structured clinical data for predictive modeling and decision support: a roadmap to tackle the challenges. Appl Clin Inform 2016;07(04):1135–53.
[31] Tute E, Steiner J. Modeling of ETL-processes and processed information in clinical data warehousing. eHealth 2018;8.
[32] Lamé G, Simmons RK. From behavioural simulation to computer models: how simulation can be used to improve healthcare management and policy. BMJ Simul Technol Enhanc Learn 2020;6(2):95–102.
[33] Schorer AE, Moldwin R, Koskimaki J, Bernstam EV, Venepalli NK, Miller RS, et al. Chasm between cancer quality measures and electronic health record data quality. JCO Clin Cancer Inform 2022;(6):e2100128.
[34] Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020;8(3):e17984.
[35] Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Canad J Cardiol 2021;37(9):1322–31.
[36] Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. J Am Med Inform Assoc 2010;17(1): 49–53.
[37] Yu Y, Jiang G, Brandt E, Forsyth T, Dhruva SS, Zhang S, et al. Integrating real-world data to assess cardiac ablation device outcomes in a multicenter study using the OMOP common data model for regulatory decisions: implementation and evaluation. JAMIA Open 2023;6(1):ooac108.