“In conferences, everyone goes ‘health data is the future’ ”: an interview study on challenges in re-using EHR data for research in Clinical Data Warehouses

Sonia Priou, Guillaume Lamé, Marija Jankovic, Emmanuelle Kempf

"In conferences, everyone goes 'health data is the future' ": an interview study on challenges in re-using EHR data for research in Clinical Data Warehouses

Sonia Priou, MSc[1], Guillaume Lame, PhD[1], Marija Jankovic, PhD[1], Emmanuelle Kempf, MD, MSc[2,3]

[1]Université Paris-Saclay, CentraleSupélec, Laboratoire Génie Industriel, France; [2]Université Paris Est Créteil, AP–HP, Department of medical oncology, CHU Henri Mondor and Albert Chenevier, Créteil, France; [3]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, Paris, France;

Abstract

More and more hospital Clinical Data Warehouses (CDWs) are developed to gain access to EHR data. The rapid growth of investments in CDWs suggest a real potential for innovation in healthcare. However, it is still not confirmed that CDWs will deliver on their promises as researchers working with CDWs face many challenges. To gain a better understanding of these challenges and how to overcome them, we conducted a series of semi-structured interviews with EHR data experts. In this article, we share some initial results from the ongoing interview study. Two main themes emerged from the analysis of the transcripts of the interviews: the importance of infrastructures in terms of data and how it is generated, and the difficulty to make care, clinical research, and data science work together. Finally, based on the experts' experience, several recommendations were identified when using a CDW.

Introduction

The secondary use of Electronic Health Record (EHR) data is increasingly common in healthcare research. EHR data are composed of several data types, including structured encoded terminology (e.g. claims data) and unstructured text (machine-readable or scanned)[1]. There are many challenges when re-using EHR data for research as the data are not intended for that purpose[2]: privacy and regulatory concerns[3], representativeness of the population[4], accuracy of the data[5], data completeness[2], lack of control on the data generation process[6], inconsistent information between different EHR data sources[7], or difficulties in obtaining access to EHR data[8]. Several researchers have retrospectively analysed the challenges they faced when using EHR data[9,10], and expressed some guidelines[11] (e.g., take into account the context in which EHR data were generated[12,13], check for potential risk of bias[14,15], or follow data quality guidelines to improve the validity of the study[16]). Despite these guidelines, concerns about the quality of the studies have been raised and several checklists have been developed to help readers evaluate of the quality of research on EHR data[17,18].

In France, researchers can perform analysis on EHR data through various databases, including the anonymised national claims database[19], hospital Clinical Data Warehouses (CDWs)[20], and regional specialised registries. CDWs are the most recently developed infrastructure to access EHR data. CDWs are both promising and exciting, and fraught with challenges. Compared to claims databases, hospital CDWs enable researchers to access different formats of EHR data as well as other data such as national death registries. The main advantage of CDWs is the access to laboratory results and clinical text reports[21]. However, the process of extracting, transforming, and loading (ETL) EHR data into CDWs is not straightforward. Technical difficulties are common when building ETL queries, with concerns on how these may affect the quality of CDW data[22]. Recommendations on how to implement a CDW in the most efficient and correct ways have been raised[23]. A main concern about using EHR data from CDW is the correctness of the information available in comparison to the original data. A study compared the data from the Hospital Information System (HIS) and the data in the CDW to validate the ETL processes of the given CDW[24], but this approach is only possible if researchers can get access to un-anonymised data to do the comparison of the content of pre- and post-ETL data. This validation method is also very time consuming and only a limited number of data can be verified.

Regardless of the difficulties, CDWs keep attracting investment. Since 2018, the number of hospital-based CDW has been increasing rapidly in France[21], partly encouraged by the French government with the creation in 2019 of the Health Data Hub[25], which aims at simplifying the access to healthcare data. In 2022, the aspiration to make France a

leader in e-health was accompanied by a call for projects by the French Health Ministry to set up and develop hospital CDWs. As of date, there are 22 CDWs in France that collect and analyse EHR data[21].

Notwithstanding the enthusiasm and public push to develop CDWs, little empirical research has been conducted on how CDWs are used for research, what difficulties users encounter, and what advantages they derive from this infrastructure. Research on EHR data has been very fruitful and the limits of these analysis are well known. However, the added step of going through a CDW to access the data has seldom been discussed, even though the added layer of data treatment is not neutral. The question stands: do CDWs deliver on their promises, or do challenges outweigh them? This is a crucial question, as innovations like this need to build momentum with users to become sustainable. If challenges are too difficult to overcome, the enthusiasm for CDWs may fade away, and CDWs may linger in the so-called *Valley of Death*[26] that spreads between promising concepts and actual value creation by successful innovations.

Through interviews, we investigate how research teams deal with this new infrastructure and the challenges they encounter. The objective is to gather user experience, to identify common challenges, to pick-up good practices, and to highlight areas where efforts should focus to improve CDW's contribution to healthcare research. In this article, we report on the first part of a larger study.

Methods

We conducted a qualitative study based on semi-structured interviews with French-speaking experts in healthcare data analysis between October and December 2022. We focused on EHR data generated during patient care. Participants were able to withdraw from the study at any time. There was no patient involvement in the study. This study was approved by the Paris-Saclay University Ethics for Research Committee on May 13th, 2022 (Approval CER-Paris-Saclay-2022-033).

Potential participants were contacted based on their publications and conference presentations, involvement with key French institutions working with EHR data, and snowball sampling, whereby one interviewer suggests the name of other interesting candidates to contact. Participants were contacted via email. Interviews were conducted either face-to-face or online via videoconferencing, depending on the participant's preference. All interviews were conducted by the first author, following an interview guide. The interviews lasted between forty-five minutes and two hours. All interviews were recorded and transcribed verbatim, using QSR Nvivo's automated transcription service. Every draft transcript was carefully and entirely checked against the original recording, validated, and anonymised by the first author. All the interviews were conducted in French by the first author. All citations included in this article were translated to English by the first author.

The semi-structured interview guide started with a general question about the participant's background and their current position. The interviewee was then asked to choose one research project that they had worked on to use as a driver for the interview. Their research project had to rely on healthcare data. Interviewees were asked a short presentation of their research project (scientific question, expected results, population of interest, findings, publication). Then, we asked about the difficulties they faced during their project and we probed for certain types of difficulties : regulatory, access to data, data quality checks, performance of algorithms. The final part of the interview focused on the data used during the project and the methods applied.

The anonymised audio records were analysed using QSR NVivo release 1.6.2. Transcripts were coded inductively. The following steps were followed: familiarisation with the data by listening to the tapes and validating the transcripts, coding the records without following a predefined framework (i.e., conventional content analysis[27]), regrouping codes into higher-level themes. Final themes were reviewed collectively to confirm the interpretation of the data during a three-hour discussion meeting between the co-authors.

Results

In this first phase, we interviewed ten experts. Six worked on EHR data from hospital CDWs, two worked on data from cancer registries, and two on data extracted from the French National Claims Database. The interviewees were either epidemiologists, data scientists, clinicians, medical informatics experts or CDW directors. All had been involved in research projects using data from CDW. Three themes emerged from the data. "Infrastructure matters" and "Three worlds colliding: Care, clinical research and data science" detail the challenges encountered by the interviewees (Table

1). The third theme "Moving forward: how do we still make it work?" synthesises good practices and suggestions from interviewees.

Table 1. Themes, subthemes, and analysis of the challenges encountered during the interviews.

Theme	Subthemes	Analysis	Potential consequences
Infrastructure matters	Data transfer and privacy rules: why users do not always get all the data they expected	Data in CDW are restricted to hospital care for patients who come to hospital Data privacy laws deprive CDW of patients not informed of the potential re-use of data Not all the data from the HIS are available in the CDW ETL are complicated to design and can degrade the data	Not all studies are possible when working with CDW Biased selection of patients Missing data Unpublished studies due to data completeness issues
	The key role of Hospital Information System providers	HIS were not designed for data re-using Purposes HIS are still not designed for data re-using Purposes CDW are dependent on HIS and are prone to its malfunctions Software providers show no interest in helping CDW extract and prepare data	Hours of work to extract and prepare the data Incomplete data due to HIS bugs Discover HIS failure when verifying data quality
	Peculiarities in structured claims data	It is the most commonly used data format Their billing purpose make them more correct than other data formats Their billing purpose create specificities with no clinical meaning	Results are not always representative of clinical behaviour False clinical conclusions
Three worlds colliding: care, clinical research and data science	Managing expectations: from simple clinical questions to data nightmares	Define a research question that is relevant for clinical research and feasible with the data available A lot of back-and-forth between clinicians and data scientists are necessary Evolution of the clinical question because of the primary results	Loss of interest in the study
	Building on sand: dynamic, evolving data	The medical field is in constant evolution which necessarily impacts the data Temporality should be included when designing a study on retrospective data	Obsolescence of algorithms Maintenance of mapping vocabularies
	Dreams of automation: automated free text analysis	80% of the information is unstructured There is no common structure for reports Very elaborate writing from clinicians NLP methods are not performing as well as expected	Manually extracting features from free text Wishing for more structured data, which will not happen

Abbreviations: CDW, clinical data warehouse; HIS, Hospital Information System, ETL, Extract Transform Load; NLP, natural language processing

Infrastructure matters

Interviewees shared frustrating experiences caused by often neglected, mundane issues of technical or administrative nature.

Data transfer and privacy rules: why users do not always get all the data they expected

CDWs regroup multiple data formats (e.g., claims data, reports, imaging, and laboratory results) but within a limited scope: researchers only have access to data concerning the hospital care of the patients who came to hospital. As a consequence, not all studies are possible with CDW data and researchers need to understand the scope of the data available in CDW and design studies accordingly.

When you work on a CDW, you only have access to patients that came to hospital. You only have sick people, which implies multiple methodological biases. (Epidemiologist)

I think that one of the first biases is to be sure that the person doing the study understands the data […] because if someone decides to do a study on sore throat with a hospital CDW, it is not the right place. You will have people with sore throats, but what you will find at hospitals are phlegmons, not anginas. (Epidemiologist)

Data privacy laws, and most prominently the European General Data Protection Regulation (GDPR), can also have a significant impact on the population of a CDW. All the interviewees brought up these privacy regulations that govern access to healthcare data. Under these regulations, patients must agree to their clinical data being re-used for research. In practice, patients who object to the re-use of their data for research purposes are quite rare. However, privacy laws also cover past data, generated before the GDPR era. Therefore, patients who were treated before this time, and were never notified of the potential re-use of their medical data, cannot be included in studies. Two solutions exist: leaving out all patients who have not come back to hospital since the data re-use information campaign of a given hospital was launched–introducing a selection bias–or informing patients by mail–which is very time consuming.

Once the selection bias of the initial population understood, interviewees were often surprised by the actual content available in the CDW. Their general assumption was that all the data from the HIS would be available to researchers, but the reality is far from that. Before being made available to researchers, the data needs to be extracted from the HIS to a datalake, transformed into a format suitable for future re-use, and loaded for researchers to access it. The development of ETL processes is complicated, takes time, and is specific to each data format and each HIS software component. When building a CDW, not all data flows can be processed at once. Some data formats are more difficult than others to integrate, giving priority to supposedly easier ETL queries. After that, all ETL processes must be validated and regularly checked as there can be integration issues.

What we try to do is not to damage the structured data for once. Laboratory results, claims data, they are not supposed to be degraded. We have checks for that because, even [integrating structured data] is not obvious. We had situations where we would just lose data. (Medical informatics expert)

For example, we didn't do bacteriology because it is complicated to model. For now, it is in the datalake but with a very low knowledge representation as there is no link between the antibiogram, the bacteria and the drug in the datalake. (Medical informatics expert)

These challenges in correctly processing data into CDWs can lead to quality issues. In particular, data completeness can be compromised. Several data scientists stated that missing data is a major issue, which can lead to smaller cohorts (diminishing the statistical power of the study) and could skew the characteristics of the selected population (leading to biased results). Interviewees explained that it can lead to smaller–biased–cohorts as patients with missing data are excluded.

It is quite difficult to shut down a research project with an investigator after having bragged about the CDW. They were very excited, spent several hours discussing the medical research question with you but in the end, you have to tell them that you won't be able to complete the study because of the data quality. (Data scientist)

[We had] a complete care trajectory for 15% of the patients. (Data scientist)

One data scientist added that, in the worst cases, they prefer not to publish their results when they are not sure of the data completeness.

All the process [of data extraction] had taken so long that we only wanted the data to start the analysis. We were very close to saying "never mind, we will work on 500 patients [instead of 2000]. (Data scientist)

We could have published but we were not confident, especially regarding data quality because we had a lot of missing data. (Data scientist)

The key role of Hospital Information System providers

For studies using CDWs, researchers only have access to a very specific portion of the population: patients who have presented themselves at hospital to seek treatment, and whose data has been successfully uploaded from the HIS to the research area of the CDW. This second part supposes a good interface between HISs and CDWs, which is not always the case. The data saved by the HIS can be partial and not relevant for research, and it can be difficult to adapt to a standardized format for re-use.

When working with intensive care data, you realise they are parse compared to the data from the intensive care unit. It's because the software is totally stupid and only keeps one data point every half hour [...]. That means that if the data point is not representative of what has happened in the past half hour or so, it is all wrong. (Medical informatics expert)

Moreover, for hospitals looking for a new HIS, it came as a shock that the feasibility of extracting the data from the HIS and shaping it to be re-used in the CDW is often not considered, including in recent products. CDW interfacing and data re-use are not on the specifications of what makes a 'good' HIS. One participant went so far as to explain that sometimes even gaining access to the data in the software can complicated.

I go to meetings every month on projects regarding the CDW. So they know the difficulties. But the priority is care in a healthcare centre. So all my considerations are wiped out in regards to healthcare issues. (CDW director)

This leads to far too many hours of work by engineers to sort out the data and prepare it for re-use. The general feeling is that medical software providers are not interested in CDWs and that all talks about re-using data are for show.

It's crazy the number of hours we can lose on this stuff. It should be correct from the start but that reflects the poor quality of the information system in France. It's a disaster. (Medical informatics expert)

But CDWs are positioned in a delicate spot as they are fully depending on HISs: if the HIS changes, fails, or updates, it directly affects the data extracted to the CDW. It is normal for HISs to be subject to changes, updates, or even bugs. However, several interviewees mentioned the lack of communication between the HIS and the teams preparing the data for the CDW. It seemed that often researchers would notice a system failure in the HIS because of questionable results in a research project. It is quite concerning to discover data quality issues due to HIS malfunctions during a study as it puts at risks the validity of the study and of previous studies.

Sometimes there are little updates, or little modifications and we only notice them latter "oh why isn't this working?". (Epidemiologist)

Discovering one year later something you should have known one year ago makes you doubt one year worth of work. And then you tell yourself "perhaps we published something completely wrong". Then you start doubting what you are doing now. You think that maybe you will discover something else in 6 months… (Data scientist)

Peculiarities in structured claims data

Claims data are one of the most commonly used structured data formats in healthcare. They are often considered to be accurate because they are used for billing and therefore are double-checked.

When you have legal obligations, it helps because it's mandatory, things are installed and are done correctly. (Data scientist).

This does not mean claims data are perfect, but their limits for epidemiology studies are well-known. Nonetheless, certain peculiarities of these data still surprised interviewees. As their main purpose is billing, claims data inherit properties that have no clinical meaning, and several interviewees shared very specific facts about claims data. With no correlation to care, these specificities lead to false conclusions if researchers are not familiar with them. For example, because billing closes at the end of the year for accounting reasons, all consultations and hospital stays are closed on this date, only to be reopened later on the next year. As easy as analysing claims data may be set to be, researchers need to be extremely careful when analysing them as they are very specific and are not always representative of care.

Always keep in mind that we are working with claims data that is not perfect and so we have to put up safeguards all the time. (Clinician)

Three worlds colliding: care, clinical research, and data science

Research on CDWs exists at the intersection of three worlds: care, where data is generated; clinical research, where many investigators come from; and data engineering and science, a key element of this nascent research stream. The interaction between these three worlds is not without challenges.

Managing expectations: from simple clinical questions to data nightmares

CDWs are sometimes oversold to prospective clinical researchers, and interviewees insisted on how important it is to be clear on what will be achievable before starting the study. For all interviewees, the research question for the CDW-based studies they had been involved in had been originally elaborated by a healthcare professional, who had defined the population of interest and the outcomes of interest. In this context, the role of the data scientist is to help the clinicians adjust the research question in light the data available. Their knowledge of the database and of the technologies at their disposal are key elements in deciding what will be possible and what will not.

Sometimes it is like a Santa wish list. Some, mostly the head of departments maybe because they are more imaginative or older I don't know, but they say "so it's easy, you have this this this and that" and we answer "hold on, careful, because behind all this, we need feature extraction, classification, validation… And the data are not that clean". So we need to explain this to them, to slow them down and tell them that what they are asking for require time and/or money. (CDW director)

According to interviewees who worked with CDW data, more often than none, after the first extraction of the cohort, a few back-and-forths between the clinicians and the data scientist are needed despite the preliminary discussions. The delimitation of the research question when analysing the data is not always trivial and clinicians can end up with more (or less) patients in their cohort than they had anticipated. The selection criteria are not always as relevant as expected, and several attempts can be necessary to identify the right population.

The clinicians said: okay, it is easy, this is how you can identify the population: […]. But once it was computed, we realised we had almost no patients. It was simply because the information used to identify the population was not there. […] People create cohorts without considering the data available in the dataset, using features that are not appropriate at all. (Epidemiologist)

It is also quite common that the research question evolves as the data is being analysed as criteria that clinicians had not considered pop up. New data elements can be brought by the data scientist to specify the clinical question according to the data available, sometimes transforming a simple medical question into a very complicated data problem.

We could think the question is simple, but it isn't, and diabetes is one of the simpler ones because liver failure, kidney failure: is it moderate? severe? acute? chronic? (CDW director)

Building on sand: dynamic, evolving data

Care is an evolving eco-system. New molecules, new terminology, new treatment guidelines or even new diseases are prone to happen in any medical specialities. It is not surprising that clinical data evolves as well, as it captures the information of this eco-system.

This dynamic aspect of care impacts clinical researchers as they need to add a temporal aspect to their clinical research question when using retrospective data. If a CDW has access to data from several years back, it is important to understand how care has evolved between the beginning and the end of the time period of the data.

They can't know it all and there are evolutions. Now, young doctors tell us that they learn the term fibula for that bone in the leg, whereas we call it "péroné". (CDW director)

The evolution of clinical data begs the questions of the obsolescence of algorithms in time, and of the maintenance of mapping vocabularies. More and more mapping vocabularies are being engineered to help standardize CDW data into Common Data Models. However, these mapping vocabularies need to be checked regularly, and possibly updated to remain accurate.

The mapping, you must maintain it because over time vocabularies change and the structure of your initial dataset changes. (Epidemiologist)

What is really important is to frequently check the outputs of the algorithms. Frequently, a new validation process is needed because, as we know, the variables will be modified, things will change. And if there is no surveillance and therefore nobody working to produce data of good quality to re-validate the algorithms' outputs, it will not work. (Epidemiologist)

We have to be careful because, in every database […] the variables, over time, are abandoned in favour of others. […] What we realise is that every year, regularly, there are new molecules. So, if we don't train the algorithm again, we will bias the results. (Epidemiologist)

Dreams of automation: automated free text analysis

Two interviewees stated that 80% of healthcare data is unstructured. In this situation, all interviewees agreed on the benefits of extracting key patient information from free text to enrich structured data.

You can work strictly on structured data but you will quickly reach its limits. Almost every time we want to do a project, we are told we need this precise feature, and it is not in the structured data. (Epidemiologist)

However, those who tried to leverage free-text data had multiple concerns. First, as anticipated, free text reports have no pre-determined structure. The structure of a given type of document can differ significantly between hospitals, and even between clinicians from the same hospital, since clinicians are free to modify the structure of a report.

There is a framework, but it changes between departments, and clinicians can change stuff in the framework. So we cannot do something with an adhoc rule. We need something that can adapt to these changes and to these variations per department and per clinician. (CDW director)

A clinician told us "I know why your performances are a bit lower here, it's because we are not good. It's the clinicians. We know we don't describe this very well. It's not standardised". (CDW director)

Second, what really struck those working on free text was the content of the reports. The text in the reports are much more elaborated than expected, making it much more challenging to exploit than imagined. Natural Language Processing (NLP) technologies cannot always cope with this rich, flowering prose.

I was working on pathology reports, and I had the impression that all [the clinicians] had literature degrees and that they enjoyed changing the formulation in each conclusion. […] Each formulation was different, and when it comes to negation, they are very inventive. (Data scientist)

I have been working for 10 years on [NLP methods], we've been talking about them for 10 years, and we are still at the same phase: NLP is still NLP. If you have a comma, a dot, an "a" instead of an "e", not the correct accent, it is very complex to analyse. (Epidemiologist)

Interviewees for whom NLP projects were not successful explained that the solution found in those situations was to extract the features by hand. Clinicians would offer to read the patients' files and extract manually the information required for the study.

[Clinicians] tell us: "We will find a solution, we will go through each patient's file one by one and extract it manually". (Data scientist)

The clinicians said that all the files were interesting so they would look at all of them one by one. An intern, when you tell them they will have to go through 400 files, they are not scared. (CDW director)

As a consequence, one of researchers' biggest dream is to have only structured or semi-structured data, which would be much simpler for analysis. But structuring the data is not a priority for clinicians, as it provides no benefit for patient care. Therefore, there is no point in trying to change the information clinicians enter in the HIS, e.g. by forcing them to fill forms rather than free-text fields. The question arose of clinicians who also do research: could they be incentivised to better structure their data, given the perspective of re-using it later for research? Even in this case, this did not seem to be a viable perspective.

When we read the reports, it was prose. It was very very difficult. And [the clinicians] all say: "what if we structured the information when we type it?". But when I ask if they would actually do it, the answer was no. (Epidemiologist)

Moving forward: how do we still make it work?

Despite the challenges mentioned above, interviewees still managed to complete projects using CDW data. Their experiences led them to formulate a few key recommendations for future projects (Table 2).

Table 2. Themes, subthemes, and recommendations based on interviewees' experience.

Theme	Subthemes	Recommendations
Moving forward: how do we still make it work?	Teamwork, from start to finish	Clinicians need to provide implicit rules about care and help understand the context of the hospital The context is important to interpreted the data and the results
	Bounded projects: define a precise perimeter and timeline for the study	Define a precise research question that is not too ambitious Set a precise timeline for each step of the project
	Knowledge management: define common guidelines for CDW projects	Sharing documentation to help new researchers Creating and sharing guidelines, safeguards and algorithms for research projects on CDWs Intervention of experts if needed

Teamwork, from start to finish

Collaboration between clinicians and data scientists is essential to make sense of the data analysed in healthcare. Pre-set, implicit rules flourish in healthcare, and they often need to be integrated into algorithms for them to be functional. Clinicians master this specific knowledge, which is not always deductible from the data. Therefore, close collaboration with clinicians is indispensable to data scientist when designing an algorithm.

The problem is usually that there is a lack of communication between the people in Artificial Intelligence, mostly informaticians or mathematicians, and the clinical world with the medical knowledge. There are a lot of companies […] that discover that there are plenty of implicit rules and that if you don't apply them, your work is worth nothing. (CDW director)

Clinicians also represent the hospital and are aware of how it runs. They can provide important insights about the context of the data generation. For example, simple changes in the workforce may explain changes in hospital caseload that would otherwise be noticeable in the data, but not explainable without further knowledge.

There are numerous potential answers to [missing data]. […] Here, the answer we know it. It's because a surgeon specialised in tumour removal left. The activity has almost completely disappeared because of it. (CDW director)

Without first-hand knowledge of the context in which data was first generated, the interpretation of a decrease in activity could have been completely different, and possibly wrong. Now that more and more researchers are gaining access to healthcare data, it is important that the collaborations between data scientists and clinicians is valued, in order to guarantee that the results of a study are always contextualised.

One of the biggest biases when you do a study on databases, it is not always the data, it is the people who exploit the data because they don't understand what they are looking for. (Epidemiologist)

<u>Bounded projects: define a precise perimeter and timeline for the study</u>

Interviewees agreed that research questions should not be overly ambitious. Often, research questions were too wide at the start of projects, which may jeopardise the whole project, or make it drag on to the extent that researchers might lose interest and become disillusioned by the possibilities offered by CDWs. Therefore, it is important to define a very precise research question.

We realized that [the research question] was too large, so we had to break it down cancer per cancer. We could not analyse all the cancers together as they each had their own specificity and mixing cancers with different survival rates made no sense. (Data scientist)

The project needs to be restricted in time to assure all involved researchers keep interested. This is even more important in the context of collaboration between different fields (healthcare, data science, research) as each party does not necessarily understand the time needed for each task. A feeling of not moving forward in the research can appear when tasks performed by others are not well understood, or not given proper consideration.

For medications, we don't have structured columns with the number of capsules per administration or the grams of drug per capsule. This information is in the title, so we had to extract it. It is a lot of work. It took a lot of time and the clinicians had the impression that we weren't making any progress. (Data scientist)

<u>Knowledge management: define common guidelines for CDW projects</u>

When re-using healthcare data, a learning period is necessary. With experience, researchers gain knowledge about the data they are analysing. However, during the interviews, several interviewees mentioned that they would have appreciated more documentation at the start.

Most of the competences are developed within the teams that have been working for X years on the data. Because when you work on a research project, you realise that something is not right. You investigate why that is. And then, in the next study, you tell yourself: "this time I will check for this". That is how you enrich your catalogue of competence and expertise. (Epidemiologist)

There was a general consensus among interviewees that only little documentation was shared across institutions. Creating and sharing guidelines, safeguards, and algorithms could be interesting and valuable.

I feel like we are missing a common toolkit. It is a shame because I imagine that there are a lot of users that have to do the same things. If we have a toolkit managed by [an institution], it would simplify everyone's life. It must exist somewhere. It is not an absurd request. (Clinician)

Finally, all interviewees mentioned the importance of punctual expert intervention for specific issues that neither the clinicians nor the data scientist can answer. These experts were mostly claims data experts, and regulation experts for privacy-related issues.

Discussion

The aim of this study was to investigate the challenges faced by several EHR data experts during research projects. We presented the preliminary results of this on-going study. In the first theme, interviewees highlighted the importance of infrastructure and its impact on the data made available in the CDW. The data in CDWs are very specific and

therefore do not allow all studies. The restriction to hospital patients as well as data privacy regulations make the population of interest of CDWs very specific. Data completeness in CDWs is also a challenge because of how long and fastidious data flows are to elaborate. Finally, the complexity of data integration in CDWs also relies on HISs and their providers. HIS providers seem to show little interest in HIS data re-use for research, and do not appear to prioritise data interoperability. Consequently, hours of work are needed to extract, transform, and load each data format into the CDW. The regulated environment of claims data makes them more reliable, but also dependant of the specificities of billing. The second theme detailed the interaction between healthcare, clinical research, and data science. CDWs aim at aligning these three different fields to produce interesting studies. However, the interaction between healthcare and data science is complicated: transforming a clinical research question into a research question answerable via CDW data is not always easy nor feasible. It can be sometimes so complicated that clinicians lose interest in the study because it takes too long. Besides, the evolution of the medical field, and consequently of the data it generates, raises questions about the obsolescence of algorithms trained on this data and how to maintain vocabulary mappings. Using clinical reports to enrich structured data is a key component of CDWs, but NLP methods do not seem to work as well as expected yet.

Despite these challenges, interviewees all managed to successfully conduct studies on EHR data by collaborating closely with clinicians from start to finish, restricting their projects to narrower clinical questions than initially envisaged, and keeping to a tight timeline. They highlighted the need for more exchanges between CDW users in different institutions, in order to share knowledge and build a common documentation.

In this study, we interviewed people from various backgrounds and institutions. The semi-structured interviews gave them scope to express ideas outside of our predefined interview guide. Nonetheless, there are several limits to our approach. First, this is an interim report, and given the broad nature of our questions it is likely that new themes will emerge from future interviews[28]. Second, as is usual in qualitative research, we strove for diversity in our sample of participants, but made no attempt to obtain a statistically representative sample of CDW users. Finally, this study is restricted to France and Belgium, and there is no guarantee that our findings apply elsewhere. In a second series of interviews, we hope to broaden the scope of the analysis by interviewing researchers from other countries.

Our results open avenues for future research. A recent qualitative study sketches the current landscape of the French CDW ecosystem.[29] Our research aims at highlighting the common challenges faces by these CDWs' users. A key question is how researchers handle the scientific responsibility of their studies and results in this context of uncertainty on the data they use. This could be addressed in our next set of interviews. During the next phase, we wish to interview ten more experts. The structure of the interview will stay the same. We will diversify the profiles of the interviewees and incorporate companies specialized in CDW development. Empirical studies on information loss in CDW infrastructure would also be useful to better quantify the impact of the issues we identified. Finally, in future interviews, we will try to better specify which challenges emerge from the specific nature of CDWs, rather than more generic challenges in EHR data re-use.

References

1. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*. 2019;29(4):354-361. doi:10.1016/j.semradonc.2019.05.010
2. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*. 2013;51(Supplement 8Suppl 3):S30-S37. doi:10.1097/MLR.0b013e31829b1dbd
3. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. *JAMA*. Published online May 22, 2014. doi:10.1001/jama.2014.4228
4. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol*. 2021;21(1):234. doi:10.1186/s12874-021-01416-5
5. Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. *BMC Health Serv Res*. 2017;17(1):766. doi:10.1186/s12913-017-2697-y
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*. 2005;58(4):323-337. doi:10.1016/j.jclinepi.2004.10.012
7. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on translational bioinformatics, 2010, 1*.

8. Taylor JA, Crowe S, Espuny Pujol F, et al. The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement. *BMJ Open*. 2021;11(8):e047575. doi:10.1136/bmjopen-2020-047575
9. Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth D. Big data in healthcare– the promises, challenges and opportunities from a research perspective: A case study with a model database. *AMIA Annual Symposium Proceedings*. (Vol. 2017, p. 384). American Medical Informatics Association.
10. Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Annals of Internal Medicine*. 2020;172(11_Supplement):S79-S84. doi:10.7326/M19-0873
11. Konrad R, Zhang W, Bjarndóttir M, Proaño R. Key considerations when using health insurance claims data in advanced data analyses: an experience report. *Health Systems*. 2020;9(4):317-325. doi:10.1080/20476965.2019.1581433
12. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. Published online April 30, 2018:k1479. doi:10.1136/bmj.k1479
13. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*. 2013;20(1):117-121. doi:10.1136/amiajnl-2012-001145
14. Nguyen VT, Engleton M, Davison M, Ravaud P, Porcher R, Boutron I. Risk of bias in observational studies using routinely collected data of comparative effectiveness research: a meta-research study. *BMC Med*. 2021;19(1):279. doi:10.1186/s12916-021-02151-w
15. 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;20(5):e185. doi:10.2196/jmir.9134
16. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *Egems*. 2017;5(1).
17. Kohane IS, Aronow BJ, Avillach P, 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;23(3):e22219. doi:10.2196/22219
18. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. Published online March 20, 2020:l6927. doi:10.1136/bmj.l6927
19. Boudemaghe T, Belhadj I. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). *International Journal of Epidemiology*. 2017;46(2):392-392d. doi:10.1093/ije/dyw359
20. Khalaf Hamoud A, Salah Hashim A, Akeel Awadh W. Clinical data warehouse: a review. *ijci*. 2018;44(2). doi:10.25195/2017/4424
21. Doutreligne M, Degremont A, Jachiet PA, Tannier X, Lamer A. *Entrepôts de Données de Santé Hospitaliers En France.* Doctoral dissertation, HAS. 2022.
22. 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).
23. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What You Need to Know Before Implementing a Clinical Research Data Warehouse: Comparative Review of Integrated Data Repositories in Health Care Institutions. *JMIR Form Res*. 2020;4(8):e17687. doi:10.2196/17687
24. 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;94:271-274. doi:10.1016/j.ijmedinf.2016.07.009
25. HDH : Health Data Hub [Online] Available: https://www health-data-hub.fr/.
26. Klitsie JB, Price RA, De Lille CSH. Overcoming the Valley of Death: A Design Innovation Perspective. *Design Manag J*. 2019;14(1):28-41. doi:10.1111/dmj.12052
27. Three Approaches to Qualitative Content Analysis - Hsiu-Fang Hsieh, Sarah E. Shannon, 2005. Accessed July 21, 2023. https://journals.sagepub.com/doi/abs/10.1177/1049732305276687
28. Malterud K, Siersma VD, Guassora AD. Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qual Health Res*. 2016;26(13):1753-1760. doi:10.1177/1049732315617444
29. Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: A case study in France. Yoon D, ed. *PLOS Digit Health*. 2023;2(7):e0000298. doi:10.1371/journal.pdig.0000298