Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Harmonising Alzheimer's Disease Cohorts using a Common ETL Tool

João Rafael Almeida (joao.rafael.almeida@ua.pt)
University of Aveiro
Alejandro Pazos
University of A Coruña
José Luís Oliveira
University of Aveiro

Research Article

Keywords: Clinical studies, Data harmonisation, ETL, OMOP CDM, Alzheimer's Disease

Posted Date: January 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1145054/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

RESEARCH

Harmonising Alzheimer's Disease Cohorts using a Common ETL Tool

João Rafael Almeida[1,2*], Alejandro Pazos[2] and José Luís Oliveira[1]

*Correspondence:
joao.rafael.almeida@ua.pt
[1]DETI / IEETA, University of Aveiro, Aveiro, Portugal
Full list of author information is available at the end of the article

Abstract

Background: Many scientific studies have sought to obtain better understanding of specific medical conditions. Concerning Alzheimer's Disease, there is a lack of reliable diagnostics and this can be related to the availability of only small-scale ongoing biomarker studies and longitudinal cohorts including these subjects. Aiming to generate more substantial clinical evidence, researchers have started to perform multiple cohort analyses. While this is currently possible by harmonising these cohorts into a common data model, the migration pipelines are usually implemented using programming languages. Therefore, cohort owners may have difficulties in contributing during the validation stage of these pipelines.

Results: To reduce the dependency on technical teams' support when validating the data transformations, we propose the use of an ETL tool with visual features. BIcenter is a collaborative web platform designed to implement ETL tasks through the browser. These pipelines are constructed using drag-and-drop features and intuitive forms to customise the ETL steps. This tool is an open-source project and is accessible at `https://bioinformatics-ua.github.io/BIcenter/`.

Conclusions: Our methodology produces interoperable cohorts for multicentric disease-specific studies. Therefore, we validated this using Alzheimer's Disease cohorts from several countries, combining at the end 6,669 subjects and 172 medical attributes. The harmonised cohorts now enable multi-cohort querying and analysis, helping in the execution of new studies.

Keywords: Clinical studies; Data harmonisation; ETL; OMOP CDM; Alzheimer's Disease

1 Background

Observational studies consist of a type of medical research that investigates the effectiveness of treatments for a particular clinical condition. In these studies, medical researchers limit themselves to documenting the relationship between the exposure and outcome in the study without changing who is or is not exposed to the treatments [1]. These studies can be split into three categories: case-control studies, cross-sectional studies and cohort studies [2]. The work presented here is focused on the latter type of study.

A cohort is defined as a subset of subjects that share similar characteristics [3]. These studies have sets of inclusion and exclusion criteria for filtering the subjects involved, and specific medical attributes defined in the study design for subsequent analysis [4]. The patient data used in these studies are usually stored in the institutional Electronic Health Record (EHR) system, and can then be exported and analysed by medical researchers [5]. In studies focused on diseases with limitations

in finding subjects to conduct a study, it is common to reuse data obtained in previous studies [6]. One of the issues with this strategy is the dependency on technical teams to obtain the data from the EHR systems, which is usually a bureaucratic process that delays study execution.

Another issue regarding this strategy where researchers aim to combine the data from distinct institutions to conduct a multiple cohort study is the lack of interoperability between these datasets [7]. While this strategy is not a common practice due to the difficulties associated with the process, some situations require more subjects involved in the study in order to generate reliable evidence. Combining multiple cohorts may solve this lack of subjects, since it increases population size, the power of statistical evidence, and thereby the study's impact [8].

The potential impact of these studies has also motivated researchers to seek more robust and reusable solutions to aggregate knowledge from distributed health datasets. This leads to establishing organisations and methodologies to explore clinical databases by reusing existent data [6]. One of these efforts aims to create a strategy to reuse EHR databases using a homogeneous schema, in order to facilitate the interoperability between databases. This integration is currently possible and optimized to use open source frameworks to support the whole process [7]. However, harmonising medical concepts requires thorough knowledge of the data and how to map it into a standard definition. This results in collaboration between technical teams capable of implementing ETL (Extract, Transform and Load) pipelines and medical specialists in the data domain, who are usually the cohort owners. Although there are tools to help in this collaboration, the process of mapping and harmonising the cohort raw data into a standard data schema is time-consuming. Furthermore, the ETL pipelines need to be validated by cohort owners to ensure that information was not corrupted.

One of the communities aiming to develop strategies to support large-scale observational studies in health care data is OHDSI (Observational Health Data Sciences and Informatics)[1] [6, 9]. Although other initiatives have similar goals, the OHDSI principles are currently well established, with strong adoption by health institutions to conduct observational studies. One of these principles is the availability of open-source tools to create observational databases and perform medical product safety surveillance using those databases [6]. One of the major outcomes of this community is the OMOP (Observational Medical Outcomes Partnership) CDM (Common Data Model), which is a database schema defined to standardise the content of healthcare databases for observational studies [10].

In Europe, a project inspired by the core principles of OHDSI was the European Medical Information Framework project (EMIF)[2]. One of its goals was to enhance access to patient-level data from distinct health institutions across Europe, while researchers could carry out distributed observational studies [11]. In one of the project's tracks, relevant cohort studies across Europe focusing on Alzheimer's Disease (AD) were connected. EMIF-AD aimed to accelerate the discovery and validation of new biomarkers to diagnose this disease in the pre-dementia stage and to predict the rate of decline [12].

[1] http://www.ohdsi.org/
[2] http://www.emif.eu

1.1 Motivation

In previous work [7], we proposed a methodology aiming to harmonise different cohorts into a standard data schema. This methodology followed the OHDSI principles, reusing the OHDSI Common Data Model, which is currently used to harmonise EHR datasets for observational studies. This methodology was applied and validated in two Alzheimer's Disease cohorts, combining at the end 6,669 subjects and 172 clinical concepts. Figure 1 gives an overview of the key operations during the ETL pipeline. This workflow is divided into three main stages, having two processes running in parallel. The first stage extracts cohort information and loads it into the ETL pipeline for processing. The transformation stage performs all the defined operations over the raw data, using an input file containing the mappings of the clinical concepts onto their standard definition. The final stage loads the data into the database and generates a report with possible warnings and errors that may have occurred during the migration procedure.

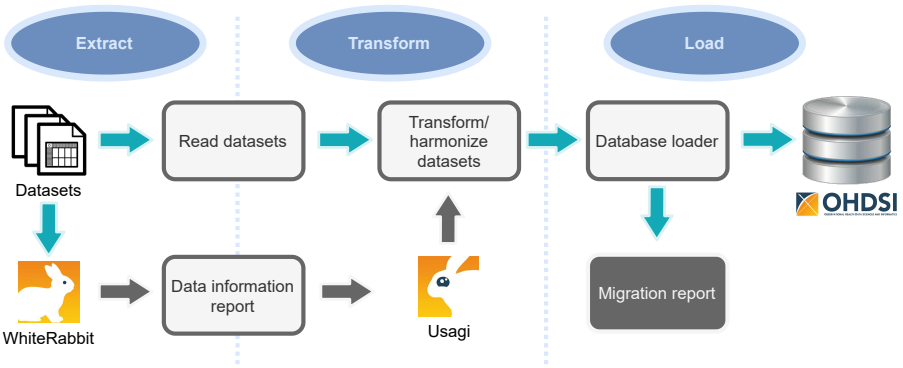

Figure 1: **Migration workflow.** Overview of the migration workflow from raw data to the OMOP CDM structure, using the methodology proposed at [7]. This workflow is divided into three main stages, having two processes running in parallel. The first stage extracts cohort information and loads it into the system. The transformation stage performs all the defined operations over the raw data using the mappings. Finally, the loading stage inserts the data in the database, producing a migration report which indicates all the problems with the original raw data.

The original concepts are mapped onto their standard definition using the Usagi tool, which is a desktop application from the OHDSI ETL toolkit. This tool provides an interface where data owners can link medical concepts existent in the cohort's raw data to a standard code. This is a process that requires human interaction and is usually iterative.

Although this methodology was able to solve the interoperability issue, the manual interactions and the lack of interconnection between teams may delay the migration pipeline. In Figure 1 there is a parallel flow shown by dark arrows. This flow is what currently makes the cohort harmonisation take longer than expected. For example, in the event of incorrectly mapping a concept, or detecting data with errors in the source data during the validation stage, data owners need to rectify these. This

new interaction would require an element of the technical team to execute the ETL pipeline again to generate a new migration procedure. This process would generate a new report that is analysed by the cohort owners, which may result in new adjustments.

Since this is an interactive process that requires active collaboration between these two teams, it is necessary to use strategies to coordinate these interactions and optimize the process. This was done using TASKA, a task/workflow management system designed to simplify the set-up of health studies, the management of participants and their roles, and the overall governance process [13]. This tool provides features for coordinating interactions in the different stages of the ETL pipeline.

However, there is another issue regarding the toolkit developed in this previous work [7]. The toolkit is operated without a graphical interface, and therefore, cohort owners have difficulties in collaborating in the construction of the ETL pipeline and its validation. In this methodology, cohort owners rely only on the ETL output and the migration report to understand if the harmonisation pipeline has produced the expected outcome.

1.2 Contribution

To avoid the dependency on technical elements to perform small adjustments in the ETL pipelines, we propose the use of a web platform to define these pipelines. This tool has an ETL Visual Editor that simplifies the implementation of ETL pipelines. The goal is to substitute old scripts with graphical flows that can be created using drag-and-drop features.

Although some steps would be easier to implement using a programming language, their validation would be harder. Therefore, this new paradigm for cohort harmonisation is helpful to allow cohort owners to understand what is happening with the data. This simplifies the collaboration between both teams in the stages of the pipeline: 1) design; 2) implementation; and 3) validation.

2 Methods

2.1 Methodology Overview

The methodology used to harmonise the cohorts is established in the ETL principles. It explores a similar strategy as represented in Figure 1. The extraction stage is responsible for gathering the data from their source. It can perform the connection to the database or load the data from CSV files when these were exported from the EHR system. The goal of this stage is to load the data into the ETL pipeline without interfering with normal use of the system. In scenarios like health databases, this requires extra attention, since the collection task may interfere with the EHR performance. In the case of cohort studies, the amount of data is reduced, which should not interfere with the system's normal behaviour. Furthermore, we validated this proposal using two real cohorts. Both use cases had the data exported to spreadsheets, and so we focused on loading the data from CSV files.

The transformation has an initial phase to transform the data to a homogeneous format, aiming to simplify the data harmonisation steps. This stage also uses the output of the WhiteRabbit tool, which is an OHDSI ETL to extract metadata information about the data sources. This output is used as input in the Usagi tool,

where cohort owners can map the original concepts to their standard definition. In the case of health databases, this mapping stage is fundamental to ensure interoperability between databases. In a subsequent phase of this stage, these mappings are uploaded in a component responsible for applying the transformations to the original data.

The loading, and final stage, loads the processed data into the target database. This database is compliant with OHDSI analytical tools. Therefore, the principles applied for observational studies on OHDSI can be reused in this scenario.

2.2 Collaborative ETL tool

Since the ETL pipelines for this use case may require the intervention of cohort owners, a tool with collaborative features may simplify their implementation. Kettle, also known as Pentaho Data Integration (PDI), is considered one of the most relevant and complete tools aiming to simplify the design and creation of ETL processes [14, 15]. The problem with this tool is the lack of collaborative features. However, BIcenter is the implementation of a reflexive software architecture that enables a simple and dynamic representation of ETL components. This is a web-based ETL tool that covers some limitations and problems currently found in building and managing ETL tasks in multi-institution environments [16].

One of the main features of BIcenter is the Visual ETL editor. This editor is illustrated in Figure 2, where an ETL pipeline with four simple steps was defined. This tool can fill some of the existent gaps in the ETL tools, namely related to collaborative environments. With BIcenter, cohort owners can participate actively in implementing the ETL pipelines, which would simplify the ETL design, implementation and validation.

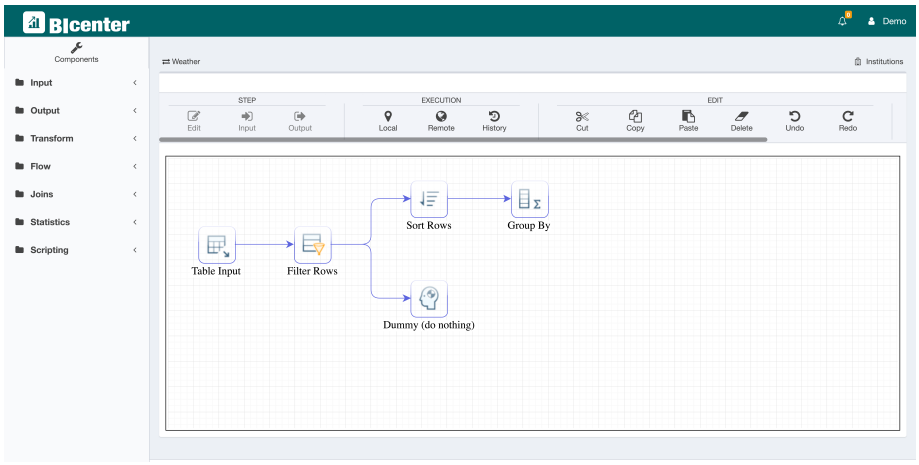

Figure 2: **BIcenter task editor.** ETL Task editor of BIcenter, where it is possible to define and execute ETL tasks. This is a simple example that uses four common steps.

This application implements a logic to segregate users by project or institution. Therefore, we defined that for each cohort, there is a set of users with permission to work in the ETL tasks. These tasks can be executed using a local database or private

and remote servers. In our case, we used the local database during the development of the ETL tasks, to then apply the same pipeline using the remote servers. These servers are based on Carte, which is a lightweight HTTP server available on Kettle that allows remote and parallel execution of ETL tasks. This approach aims to ensure data protection and isolation when dealing with sensitive patient data.

2.3 Usagi Mapper Component

Although BIcenter already includes a set of ETL operations, some flows can be optimized, namely by creating a new step. A component capable of applying the transformation defined on the Usagi tool directly in the data would reduce a set of operations in the diagram to a single step. This transformation would be able to identify the source concepts in the data and change them for the standard codes. Furthermore, this component would reduce the complexity of the ETL diagrams considerably and the cohort owners would only need to update the file with the mappings in each update.

Figure 3 illustrates the interface of the Usagi Mapper in BIcenter. This interface aims to be intuitive for the cohort owners, and the fields in this form can be easily understood by non-technical people. The "Variable" field is the column in the source data which would be applied to this transformation. The data in this column are matched with the mappings in the Usagi output, which are defined in the "Input Column" field. The new values for this transformation are defined in the same output but in a different column. This column is defined in the "Output Column" field. These options are compliant with the Usagi file structure.

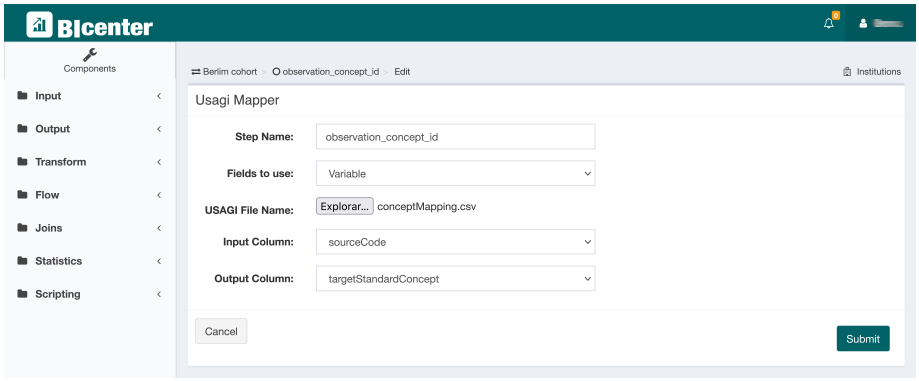

Figure 3: **Usagi Mapper component.** Configuration view for the Usagi Mapper component. The first field represents the step name in the ETL task. The second field selects the field in which the transformation would be applied. The remaining fields are for uploading the Usagi export file, and to select the input and output column.

The complexity of updating the ETL mappings in BIcenter is reduced to the operation of uploading a new file. This simple task does not require programmatic knowledge, and it can be easily executed by the non-technical users collaborating in the cohort harmonisation. In the case of cohorts with non-English medical attributes, we can use an adaption of the Usagi tool prepared for multi-language [17].

This solves an important issue since it is common to have the original data in a non-English form.

3 Results

3.1 Research Application

Designing trials in pre-dementia of Alzheimer's Disease is challenging due to the difficulty in identifying subjects with this condition. The lack of reliable diagnostics for this condition can be related to the availability of only small-scale ongoing biomarker studies and longitudinal cohorts including these subjects. In the EMIF project, researchers connected relevant cohort studies performed across Europe. This linkage of information led to new strategies for studying Alzheimer's Disease [12].

The Alzheimer's Disease track on the EMIF project focused on harmonising the cohorts' raw data from institutional partners into a common data schema. The goal was to perform large-scale analysis associated with disease course, early diagnosis and risk factors for decline, and identify and validate biomarkers using measurements from sets of medical conditions, such as DNA, cerebrospinal fluid and plasma samples. This effort resulted in a strategy for conducting large-scale studies on biomarkers and risk factors for neurodegenerative disorders.

In the previous work [7], we proposed a methodology for harmonising Alzheimer's Disease cohorts. This was validated using two synthetic datasets. In a post-validation stage, it was used to harmonise two real cohorts from patients of different health institutions. Aiming to validate the ETL tasks using BIcenter, we reused the same cohort raw data and the mappings already validated by medical researchers. Those cohorts are the Berlin Memory Clinic (BMC) cohort related to the Charité University Hospital in Berlin, and another from the BioBank Alzheimer Center Limburg (BBACL) related to the memory clinic of the Maastricht University Medical Centre.

3.2 Methodology in Practice

In BIcenter, using PDI steps and the Usagi component, we were able to implement a methodology that resulted in the same successful migrations as presented in the work [7]. In these examples, we had the cohort raw data stored in CSV files, which did not require connecting to any database. However, these datasets had a heterogeneous format. To simplify the ETL flows and reuse parts of the transformation stage, we firstly reorganised the cohort raw data into a similar format. This operation has the ETL tasks divided into two: 1) a task designed for the cohort to transform the data to a pre-harmonised format; and 2) applying the transformations in the cohort using the data from this new structure.

The pre-harmonised format stores the data into a key-value structure. Both key and values of this format are tuples. Therefore, the fields composing the key are: 1) the patient identifier, 2) the visit date, and 3) the exam or cohort attribute. The value would be the entry for that attribute, and in the last stage, the mappings for the cohort attribute and its value. This format, and how the data is reorganised in it, is illustrated in Figure 4. The first table represents an example of cohort raw data. The second table is in the pre-harmonised format, and the coloured boxes represent the reorganisation of the columns and fields in the new format. This first

transformation is the most complex and ad hoc in the pipelines since it requires the use of several PDI steps to generate this transposition. The third table addresses the mappings of the concepts. Each row in these tables represents a medical attribute collected in a follow-up visit for a single patient. This structure simplifies the harmonisation stage since the medical attributes and their values are clearly identified in the structure.

Patient ID	Visit date	...	MMSE Total Score	Clock Drawing Test	...
10424	15-01-2013	...	16	1	...
10424	24-02-2013	...	20	3	...
...	...	...	...	...	...

Patient ID	Visit date	Original exam	Value	Harmomised Exam	Harmonised Value
10424	15-01-2013	MMSE Total Score	16		
10424	24-02-2013	MMSE Total Score	20		
...	...	...	...		

Patient ID	Visit date	Original exam	Value	Harmomised Exam	Harmonised Value
10424	15-01-2013	MMSE Total Score	16	2000000166	16
10424	24-02-2013	MMSE Total Score	20	2000000166	20
...	...	...	...	...	...

Figure 4: **Data transformations.** Illustration of cohort raw data (first table) and its representation during transformation stage. The blue box represents the concepts that identify the patient's visit. The green box represents the new position of the cohort's exams. Both of these fields represent the key of the key–value structure, for the value of the exam (orange box). The yellow box represents the fields that would receive the harmonised concept codes.

The BBACL cohort was stored in a single CSV file with 313 columns and 262 rows, in which some of them correspond to the same subject but at different follow-up visits. On the other hand, the BMC cohort contained more subjects and was structured differently. This cohort was split into 5 CSV files with a total of 89 columns for 6 583 subjects. Therefore, by reorganising these cohorts into the previously described structure, the second stage of the ETL pipeline would be similar between cohorts.

Using the BBACL as an example to demonstrate the ETL flow, the first step of this methodology was to identify which columns would constitute the key of the pre-harmonised structure. These fields should be capable of representing the patient in a follow-up visit and correlated to the medical attributes. The next step was reorganisation of the raw data into the pre-harmonised structure, as explained before. Once the cohort is in this structure, the transformation and loading stages are similar for all cohorts. The Usagi component loads the mappings and applies to each mapped concept the transformation for all the entries. This transformation was applied to the exams and their results since in some cases, the exam output was a valued possibility of being mapped to a standard concept. The last part of the ETL task gathers the structure and reorganizes the data in order to fit into the data schema of the OMOP CDM database.

4 Discussion

Applying a graphical ETL tool to design the cohort migration pipelines provides some advantages. Although some parts of the tasks presented would be simpler using a programming language, this may not be the best option when non-technical people need to understand what is happening with the data. In this section, we discuss the collaborative features of our proposal, the strategy adopted to have interoperability between the resulting databases and how data privacy is ensured.

4.1 Collaborative Features in Multi-institutional Environment

BIcenter was initially developed to have different roles belonging to different institutions. This strategy allows the use of a single installation to define the migration pipelines of all cohorts with the possibility of segregating users by institutions or cohorts. Therefore, the existing rule-based access control (RBAC) mechanisms maintain sets of permissions to access the different features of the application. For instance, it allows specific users to visualise the results of each transformation, or write them in the target database.

The mechanisms to access and manage the ETL tasks and institutions can be characterized in four distinct types of users: data analyst, task manager, resource manager and administrator. The data analyst is the most limited role in the system. Users with this role can inspect task execution history, namely the aggregations of resulting data, execution logs and performance metrics. These users cannot execute the ETL pipelines. Therefore, the medical teams that only contribute to the ETL validation have this role. The task manager is the entity capable of building and executing ETL tasks within a specific institution. Some elements of medical teams have this role when they collaborate more actively with the technical teams during the ETL implementations. The resource manager is the entity responsible for managing the private data sources and execution servers at a deeper level than the task manager. Finally, the administrator is responsible for moderating the platform.

The collaborative environment is centralized in the ETL Task Editor. This workspace allows definition of the ETL pipelines. Therefore, users with permission to edit an ETL task can work collaboratively in the same workspace. Although BIcenter does not create real-time working sessions, the system provides a user-friendly environment where multiple users can work collaboratively.

4.2 Datasets Interoperability

One of the key points of harmonising cohort data is the use of a common data schema as the output of this procedure. By using the OMOP CDM schema, we were able to apply a well-established data schema that is currently used to store EHR information in an interoperable format for observational studies. Alzheimer's Disease cohorts can be mapped to this structure without any adaptations in the original data schema. This ensures that the resulting databases are compliant with OHDSI principles, and cohort owners can use the OHDSI analytical tools to interact with the data.

The interoperability lies in the use of the original OMOP CDM data schema. This schema is fully detailed at [18]. However, in this case, we only required to populate three tables, namely the "Person", "Observation" and "Observation Period" tables.

The "Person" table can store the patient's personal information, *i.e.* gender, date of birth, race and ethnicity. However, we did not require all of these fields in the Alzheimer's Disease cohorts. The "Observation" table maintains all the measures made during the study, which we defined previously as exams. Each entry in this table contains: 1) a numerical entry for patient identification, generated during the ETL procedure and only used in this database; 2) the standard code for the observation concept, *i.e.* specific exam conducted during the patient's visit; 3) the standard code for the observation type concept, which characterises the measure/exam done on the patient when it can be represented using a standard code; 4) the date and value of the observation. This value can be characterised by its type, *i.e.* it can be numeric, text or a code. The "Observation Period" table contains the time interval each patient was under observation, starting from the date of the first entry in the cohort and ending with the date of the last follow-up visit.

4.3 Data Privacy

The level of anonymity using OMOP CDM is dependent on the organization's privacy policies. The OMOP CDM can store patients' information without exposing sensitive data. In the case of sensitive attributes that would affect this directly, these were discarded during the migration. This was a manual procedure, in which the cohort owners identified the patients' attributes that did not contribute to studying the disease, but could identify the patient. The idea of this operation was to hide these attributes and aggregate the necessary fields in generic groups of data. For instance, we used patients' age and discarded their date of birth, which did not affect the data value.

We end up with databases containing harmonised patient information in a standard format. Although the data was anonymised, the institutions kept the data isolated and inaccessible without supervision. However, the people interested in querying the databases can define their study request, send it to the cohort owners and wait for the results. The cohort owners can execute the SQL against the database and analyse whether they can reveal the results. Currently, this methodology for performing distributed studies is used by the OHDSI community at the EHR database level.

5 Conclusions

Conducting a multi-centre cohort study is currently possible and easier due to existing efforts to migrate cohort raw data into a common data model. However, such ETL procedures require collaboration between a technical team and cohort owners, who are usually people with a medical background. Development of these procedures requires the above-mentioned collaboration during the design, implementation and validation of the ETL, due to the data scope.

BIcenter is a web collaborative ETL tool capable of reproducing the components of Kettle using a responsive HTML interface. This tool provides a workspace where both teams can work and understand what is happening with the data. The goal is to have a platform to set the ETL pipelines without using programming languages, which are not understood by the medical peers involved in the process. This simplifies some phases of the pipelines, reducing time, and ensures a deeper validation of what is happening with the data during each stage.

Ethics approval and consent to participate
All methods were carried out in accordance with relevant guidelines and regulations.

Consent to publish
Not applicable, same reasons as the previous item.

Availability of data and materials
Documentation about BIcenter is available at `https://bioinformatics-ua.github.io/BIcenter`. The source code is publicly available at `https://github.com/bioinformatics-ua/BIcenter`.

Competing interests
The authors declare that they have no competing interests.

Funding
This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. João Rafael Almeida is funded by the National Science Foundation (FCT), under the grant SFRH/BD/147837/2019.

Authors' contributions
JRA participated in the implementation of the current version of BIcenter. AP and JLO supervised this work. All authors have contributed to the manuscript, and have reviewed and approved this final version.

Acknowledgements
We are grateful to Leonardo Coelho for the initial developments on BIcenter.

Author details
[1]DETI / IEETA, University of Aveiro, Aveiro, Portugal. [2]Department of Information and Communications Technologies, University of A Coruña, A Coruña, Spain.

References

1. Ranganathan, P., Aggarwal, R.: Study designs: Part 1–an overview and classification. Perspectives in clinical research **9**(4), 184 (2018). doi:10.4103/picr.PICR_124_18
2. Lu, C.Y.: Observational studies: a review of study designs, challenges and strategies to reduce confounding. International journal of clinical practice **63**(5), 691–697 (2009). doi:10.1111/j.1742-1241.2009.02056.x
3. Ranganathan, P., Aggarwal, R.: Study designs: Part 3-analytical observational studies. Perspectives in clinical research **10**(2), 91 (2019). doi:10.4103/picr.PICR_35_19
4. Carlson, M.D., Morrison, R.S.: Study design, precision, and validity in observational studies. Journal of palliative medicine **12**(1), 77–82 (2009). doi:10.1089/jpm.2008.9690
5. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G.: Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of biomedical informatics **42**(2), 377–381 (2009). doi:10.1016/j.jbi.2008.08.010
6. Hripcsak, G., Duke, J.D., Shah, N.H., Reich, C.G., Huser, V., Schuemie, M.J., Suchard, M.A., Park, R.W., Wong, I.C.K., Rijnbeek, P.R., *et al.*: Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics **216**, 574 (2015). doi:10.3233/978-1-61499-564-7-574
7. Almeida, J.R., Silva, L.B., Bos, I., Visser, P.J., Oliveira, J.L.: A methodology for cohort harmonisation in multicentre clinical research. Informatics in Medicine Unlocked **Volume 27**, 100760 (2021). doi:10.1016/j.imu.2021.100760
8. Brown, C.H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burkhart, G., Vigna-Taglianti, F., Howe, G., Masyn, K., Wang, W., *et al.*: Methods for synthesizing findings on moderation effects across multiple randomized trials. Prevention science **14**(2), 144–156 (2013). doi:10.1007/s11121-011-0207-8
9. Hripcsak, G., Ryan, P.B., Duke, J.D., Shah, N.H., Park, R.W., Huser, V., Suchard, M.A., Schuemie, M.J., DeFalco, F.J., Perotte, A., *et al.*: Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences **113**(27), 7329–7336 (2016). doi:10.1073/pnas.1510502113
10. Overhage, J.M., Ryan, P.B., Reich, C.G., Hartzema, A.G., Stang, P.E.: Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association **19**(1), 54–60 (2011). doi:10.1136/amiajnl-2011-000376
11. Almeida, J.R., Fajarda, O., Pereira, A., Oliveira, J.L.: Strategies to Access Patient Clinical Data from Distributed Databases. In: HEALTHINF, pp. 466–473. SciTePress, ??? (2019). doi:10.5220/0007576104660473
12. Bos, I., Vos, S., Vandenberghe, R., Scheltens, P., Engelborghs, S., Frisoni, G., Molinuevo, J.L., Wallin, A., Lleó, A., Popp, J., *et al.*: The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. Alzheimer's research & therapy **10**(1), 64 (2018). doi:10.1186/s13195-018-0396-5
13. Almeida, J.R., Gini, R., Roberto, G., Rijnbeek, P., Oliveira, J.L.: Taska: a modular task management system to support health research studies. BMC medical informatics and decision making **19**(1), 1–9 (2019). doi:10.1186/s12911-019-0844-6
14. Kherdekar, V.A., Metkewar, P.S.: A technical comprehensive survey of etl tools. International Journal of Applied Engineering Research **11**(4), 2557–2559 (2016). doi:10.37622/IJAER/11.4.2016.2557-2559
15. Biswas, N., Sarkar, A., Mondal, K.C.: Efficient incremental loading in ETL processing for real-time data integration. Innovations in Systems and Software Engineering **16**(1), 53–61 (2020). doi:10.1007/s11334-019-00344-4
16. Almeida, J.R., Coelho, L., Oliveira, J.L.: BIcenter: A collaborative Web ETL solution based on a reflective software approach. SoftwareX **16**, 100892 (2021). doi:10.1016/j.softx.2021.100892

17. Almeida, J.R., Oliveira, J.L.: Multi-language Concept Normalisation of Clinical Cohorts. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 261–264 (2020). doi:10.1109/CBMS49503.2020.00056. IEEE
18. Hripcsak G, Ryan P, Madigan D, Kostka K, Schuemie M, DeFalco F, et al: The Book of OHDSI: Observational Health Data Sciences and Informatics, (2019)