Article

Not peer-reviewed version

# Consore: A Powerful Federated Data Mining Tool Driving a French Research Network to Accelerate Cancer Research

Julien Guérin * , Amine Nahid , Louis Tassy , Marc Deloger , François Bocquet , Simon Thézenas , Emmanuel Desandes , Marie-Cécile Le Deley , Xavier DURANDO , Anne Jaffré , Ikram Es Saad , Hugo Crochet , Marie Le Morvan , François Lion , Judith Raimbourg , Oussama Khay , Franck Craynest , Alexia Giro , Yec'han Laizet , Aurélie Bertaut , Frédérik Joly , Alain Livartowski , Pierre Etienne Heudel

Posted Date: 26 November 2023

doi: 10.20944/preprints202311.1570.v1

Keywords: cancer research; cancer; natural language processing; data mining; data warehouse; big data

*Article*

# Consore: A Powerful Federated Data Mining Tool Driving a French Research Network to Accelerate Cancer Research

**Julien Guérin [1], Amine Nahid [2], Louis Tassy [3], Marc Deloger [4], Francois Bocquet [5], Simon Thezenas [6], Emmanuel Desandes [7], Marie Cécile Le Deley [8], Xavier Durando [9], Anne Jaffré [10], Ikram Es Saad [11], Hugo Crochet [12], Marie Le Morvan [3], François Lion [4], Judith Raimbourg [5], Oussama Khay [7], Franck Craynest [8], Alexia Giro [9], Yec'han Laizet [10], Aurélie Bertaut [11], Frederik Joly [2], Alain Livartowsk [1] and Pierre Heudel [12]**

[1] Institut Curie, Paris, France (julien.guerin@curie.fr, a-livartowski@unicancer.fr)
[2] Coexya, Saint-Didier-au-Mont-d'Or, France (Frederik.JOLY@coexya.eu, nahidm@tcd.ie)
[3] Institut Paoli-Calmettes, Marseille, France (TASSYL@ipc.unicancer.fr, LEMORVANM@ipc.unicancer.fr)
[4] Gustave Roussy, Villejuif, France (Marc.DELOGER@gustaveroussy.fr, francois.lion@gustaveroussy.fr)
[5] Institut de Cancérologie de l'Ouest, Nantes & Angers, France (Francois.Bocquet@ico.unicancer.fr, judith.raimbourg@ico.unicancer.fr )
[6] Institut régional du cancer de Montpellier, Montpellier, France (Simon.Thezenas@icm.unicancer.fr)
[7] Institut de cancérologie de Lorraine, Nancy, France (e.desandes@nancy.unicancer.fr, o.khay@nancy.unicancer.fr)
[8] Centre Oscar Lambret, Lille, France (F-Craynest@o-lambret.fr, m-ledeley@o-lambret.fr)
[9] Centre Jean Perrin, Clermont Ferrand, France (Xavier.DURANDO@clermont.unicancer.fr, Alexia.GIRO@clermont.unicancer.fr)
[10] nstitut Bergonié, Bordeaux, France (A.Jaffre@bordeaux.unicancer.fr, y.laizet@bordeaux.unicancer.fr )
[11] Centre Georges Francois Leclerc, Dijon, France (icharifi@cgfl.fr ,abertaut@cgfl.fr)
[12] Centre Léon Bérard, Lyon, France (Hugo.crochet@lyon.unicancer.fr, pierre-etienne.heudel@lyon.unicancer.fr )
* Correspondence: julien.guerin@curie.f

**Abstract:** Background: Real-world data (RWD) related to the health status and care of cancer patients reflect the ongoing medical practice, and their analysis yields essential real-world evidence. Advanced information technologies are vital for their collection, qualification, and reuse in research projects. Methods: UNICANCER, the French federation of comprehensive cancer centres, has innovated a unique research network : Consore. This potent federated tool enables the analysis of data from millions of cancer patients across eleven French hospitals. Results: Currently operational within eleven French cancer centres, Consore employs natural language processing to structure the therapeutic management data of approximately 1.3 million cancer patients. This data originates from their electronic medical records, encompassing about 65 millions of medical records. Thanks to the structured data, which is harmonized within a common data model, and its federated search tool, Consore can create patient cohorts based on patient or tumor characteristics, and treatment modalities. This ability to derive larger cohorts is particularly attractive when studying rare cancers. Conclusions: Consore serves as a tremendous data mining instrument that propels French cancer centres into the big data era. With its federated technical architecture and unique shared data model, Consore facilitates compliance to regulations and acceleration of cancer research projects.

**Keywords:** cancer research ; cancer ; natural language processing ; data mining ; data warehouse ; big data

## 1. Introduction

Cancer is a primary cause of mortality globally. It was responsible for close to 10 million deaths in 2020 [1]. This disease presents a myriad of unique pathologies [2], adding a layer of complexity to its diagnosis, research, and the improvement of care. The exploration of a specific pathology often demands the identification of a significant number of patients that typically exceeds the count

 doi:10.20944/preprints202311.1570.v1



available at a single healthcare facility [3]. Therefore, data integration and interoperability across institutions becomes essential to enhance cancer research and care.

Patient identification is a critical yet labor-intensive process [4]. First step often require manual review and interpretation of electronic health record (EHR) data, a process that is both slow and financially taxing. Despite this, it remains a necessary step, considering that approximately 80% of the pertinent clinical information is encapsulated within the text of health records [5]. To tackle these challenges, the contribution of natural language processing (NLP) techniques is a key asset to help physicians and data experts identifying potential candidates for research projects.

Beyond being able to search available medical information, data must be structured and standardized to ensure their secondary reuse in research in accordance with the FAIR principles [6]. An initiative launched by eleven comprehensive cancer centres from the Unicancer network sought to enhance and expedite data sharing in oncology and has been involved in the creation of the OSIRIS model [7]. This common data model comprises a minimal set of clinical and genomic data, specific to oncology research, serving as the cornerstone of a larger initiative aimed at accelerating cancer research by simplifying the creation of cohorts of cancer patients with similar characteristics.

To navigate regulatory hurdles and reduce data flows, the initiative implemented a federated technical architecture, avoiding the need for a single centralised data warehouse. However, this federated network requires a high level of harmonisation among centres to supply the common model. To this end, the French federation of comprehensive cancer centres (UNICANCER) developed Consore, a unique network equipped with a powerful federated search engine. This tool enables the efficient and reliable identification of patient cohorts by digging into the electronic health records of millions of patients across eleven French comprehensive cancer centres.

The project addresses four major challenges: (i) the aggregation of tremendous amount of heterogeneous data; (ii) the semantic analysis of electronic health records, data standardisation, and modelling of the cancer disease; (iii) the technical implementation of a solution facilitating fast data querying at a national level; (iv) the development of ready-to-use services for clinicians and researchers. This paper presents an in-depth examination of each aspect of these challenges and their respective evaluations.

## 2. Materials and Methods

Overseen by UNICANCER, the federation of 18 French comprehensive cancer centres, the Consore project incorporates a multidisciplinary team of oncologists, bioinformaticians, project managers, data engineers and Information Technology (IT) engineers. The implementation of Consore was authorised in November 2016 (N°2016-331) by the French regulatory authorities (Commission Nationale de l'Informatique et des Libertés, CNIL). Despite this approval, it remains essential to inform patients individually and collectively about the potential reuse of their health data for cancer research in compliance with European and French regulations.

### *Data Aggregation*

Consore addresses various types of data coming from numerous disparate sources. This includes both structured and unstructured information in electronic medical records, demographic and administrative data, medical activity data derived from France's nationwide diagnostic-related group (DRG)-based information system - hospital discharge data from the programme for medicalising information systems (PMSI), biobanking data, tumor characteristics, lab results, pharmaceutical data and molecular alteration information.

### *Medical Concept Inference*

Medical reports provide a treasure trove of data for Consore. However, harnessing this data presents a significant challenge due to the voluminous mass of raw unstructured data they contain. Given the content is natural language text, it poses a formidable task for AI to process. To surmount this challenge, Consore is equipped with a specialised web service for natural language processing

(NLP) tasks. It receives medical reports, processes them via a spaCy NLP pipeline, and subsequently produces structured documents. This NLP pipeline carries out a series of operations on each report to extract as many medical concepts as possible. These operations include anonymizing personal information for regulatory reasons and text cleaning tasks such as stopwords removal, tokenization, and lemmatization.

The web service extracts several concepts from the reports using NER models, and structures them through entity linking tasks. The latter are operated according to the same standards and classifications used for structured data sources in Consore, such as:

- The "Classification Commune des Actes Médicaux" (CCAM), as the French classification for medical procedures [8];
- The 10th revision of the International Classification of Diseases (ICD-10) [9] and the 3rd edition of the ICD for Oncology (ICD-O-3) [10]

Another issue we address using Consore's NLP web service is redundancy and lack of precision. Indeed, for each patient, hundreds of reports are processed; most contain a redundant history section that is rarely updated. Moreover, in a single report, we tend to detect various occurrences of the same tumor, yet with different details: sometimes it is accompanied by the tumor location, its morphology, the associated biomarkers, the administered treatment, etc. These remain unitary concepts of different categories. Detecting them and representing them in a structured form does not solve all the challenges.

In fact, not only do we detect single concepts, but we also need to infer the links between them. For instance, there is high interest in data structuring to find the dates linked to a concept in order to locate it in time. In addition to this example, we also tackle linking a treatment response to the involved treatment or recognising whether a tumor remains in its first stages or if a metastasis is diagnosed.

*Information Consolidation Tool Using PMSI Data*

The programme for medicalising information systems (PMSI) , is a French mechanism of the national health system, aiming to reduce inequalities in resources between health institutions (Ordonnance of 24/04/1996). It stores quantified and standardised hospital discharge data to measure the activity and resources of health care institutions. The PMSI is thus a data source, which contains the patients' stays in the healthcare facility with the reasons for hospitalisation, presented in a main diagnosis and associated diagnoses, both coded according to the ICD-10 as well as most treatments coded according to the CCAM.

This data source plays a crucial role in Consore, as it serves to validated identified diseases and treatments. Being of significant financial relevance to healthcare institutions, the PMSI provides a comprehensive dataset. However, Consore uses the PMSI to corroborate inferred medical information, because hospitalisation discharge data from the PMSI is suboptimal from a clinical perspective. While the data source is comprehensive, it lacks exhaustive documentation. For instance, for a given patient, we have the clinical diagnoses, but no information on the specific location or affected side, as well as the diagnosis date; Similarly, we can retrieve information on the occurrences of chemotherapy sessions, without any details on the administered medications.

*Pivot Model for Diverse Data Sources*

Consore handles a variety of data from several different sources that do not have a shared structure. Thus, a pivot model was designed in compliance with the OSIRIS model, capable of storing information from these diverse sources. This model addresses several hurdles such as partial data, conceptual links, volumetry and redundancy, and data sourcing.

- **Partial data**: we often deal with information about the morphology of a tumor, but we have nothing about the initial diagnostic or its location; or treatment responses without the involved treatment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2023 doi:10.20944/preprints202311.1570.v1



- **Concept links**: the pivot model must preserve, when applicable, the links between the detected concepts (e.g. the link between a metastasis and the primary tumor)
- **Volumetry and redundancy**: Consore might detect or receive thousands of concepts, often redundant, for a single patient.
- **Data sourcing**: for each data item, we need to identify its source and the date it was recorded in the information system.
- In order to organise and classify all the identified concepts, we developed a common model defining the cancer disease based on several main hierarchical classes (or layers): Cancers (all cancer recurrences for a given patient), Tumor events (primary tumor, local or metastatic relapse), Acts (treatments and/or analysis), and Documents (all the documents of a patient or available biological samples) (Figure 1).

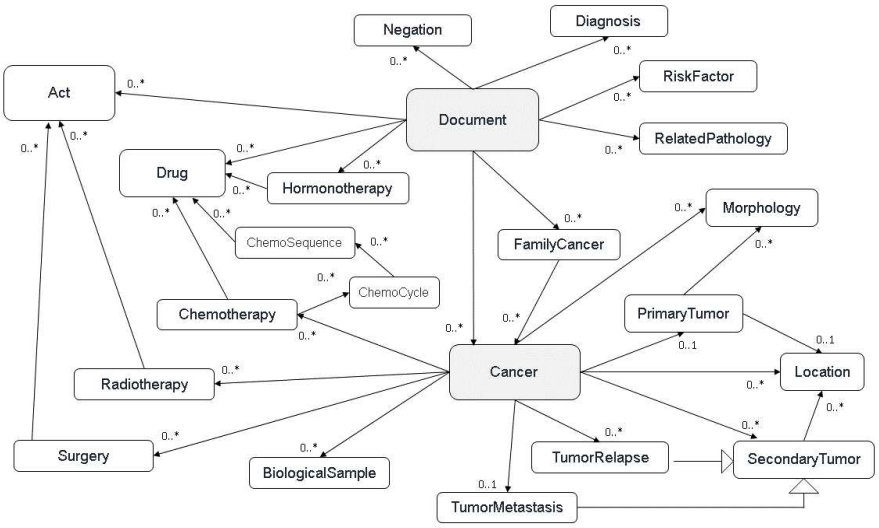


**Figure 1.** Consore's Elementary Data Model.

Upon completion of documents processing through various pipelines, each data source possessing its own, Consore carries out its primary function: data inference and structuring. The upcoming sections will delve into metastasis structuring as an example of data structuring within Consore.

*Data Inference and Structuring: an Illustration through Metastasis Structuring*

The main asset of Consore lies in its ability to infer and structure data. For each patient, after processing and storing all their data in the pivot model, Consore uses detected concepts to create a structured patient profile. This structuring process, involving the cleansing, merging, correcting, and selection of the most accurate data, results in a comprehensive document summarizing the patient's history. It enables the generation of a timeline detailing their condition, completed with events and observations since their initial examination. This structuring is complex, involving a series of rule execution in a specific order, with certain rules activated only under specific conditions.

Figure 2 shows the sequence of structuring rules leading to a patient profile within Consore's inferred model. This model, similar to the elementary model, is solely used to store inferred data where concepts are connected to the patient at the core of the model. In contrast, in the elementary model, the core is a single document.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2023 doi:10.20944/preprints202311.1570.v1



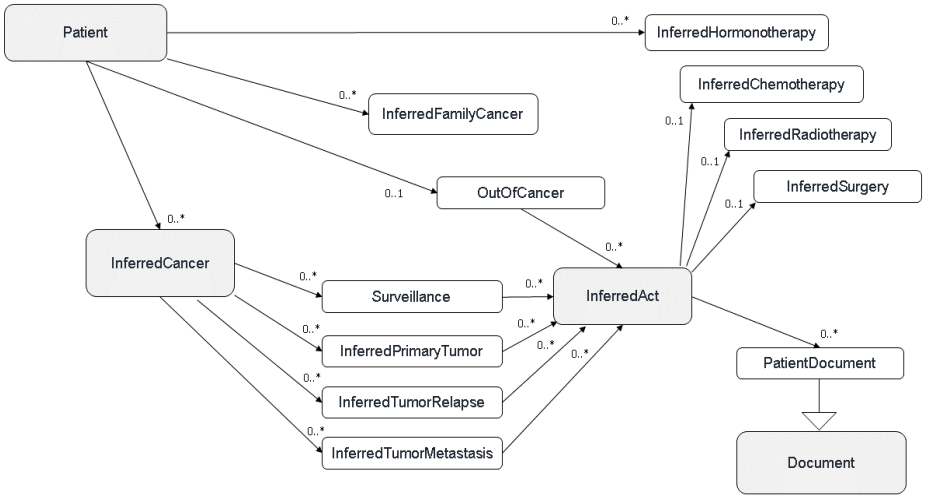


**Figure 2.** Consore's Inferred Model.

Here, we focus on the structuring rules for metastasis, a pivotal phase in cancer progression that is of great concern to oncology researchers. Identifying metastasis is technically challenging as it's often difficult to infer the primary tumor from which the metastasis arises.

The development of metastasis means in many cases the terminal stage of cancer. The primary task, therefore, is to accurately infer the starting date of the metastasis phase. However, as we deal with medical reports processed with NLP algorithms, a significant proportion of noise is introduced. These algorithms might pick up concepts of metastasis that are actually embedded in incorrect contexts such as an hypothesis or negation.

To address this challenge, we implemented a heuristic-based algorithm, summarised as follows:

1. Retrieve all detected occurrences of the concepts "metastasis" and "relapses" that are located further from the primary tumor within the patient's dataset. To maintain clarity, both types of occurrences will be referred to as « metastasis ».
2. Sort these concepts based on their dates, either the date provided by the algorithm (e.g., « metastasis diagnosed on 24th May 2022 ») or, in cases where no relevant date is found in the report, the date of the document itself.
3. Determine the relevant date within the corpus of metastasis concepts. This involves defining a heuristic time interval (potentially 3/6 months or a year), starting from the first occurrence of a metastasis.
4. Assign a weighting factor to each occurrence, considering the data source or the relevance of its associated date.
5. Calculate the cumulative weight of the concepts falling within the defined interval, obtaining the interval's total weight.
6. If the interval's weight exceeds a predefined empirical threshold, the start date within that interva is considered the commencement of the metastasis. If not, the process moves to the next interval, checks the same conditions, and repeats until a date is determined.

*Performance of the Consore tool*

Recall, precision and F1-score were used to determine the Consore effectiveness in patients with unknown primary site of cancer [11]. Recall was defined as the number of true positives (TP) over the number of true positives and false negatives (TP + FN), and precision as the number of true positives (TP) over the number of true positives and false positives (TP + FP). F1-score combined the two-competing metrics as the harmonic mean of precision and recall (2 X Recall X

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2023 doi:10.20944/preprints202311.1570.v1



Precision/(recall+precision)). A F1-score between 0.6 and 0.8 was considered as acceptable, a score between 0.8 and 0.9 as excellent, and more than 0.9 as outstanding [12].

## 3. Results

Consore is a software deployed within eleven major cancer centres in France (Figure 3).

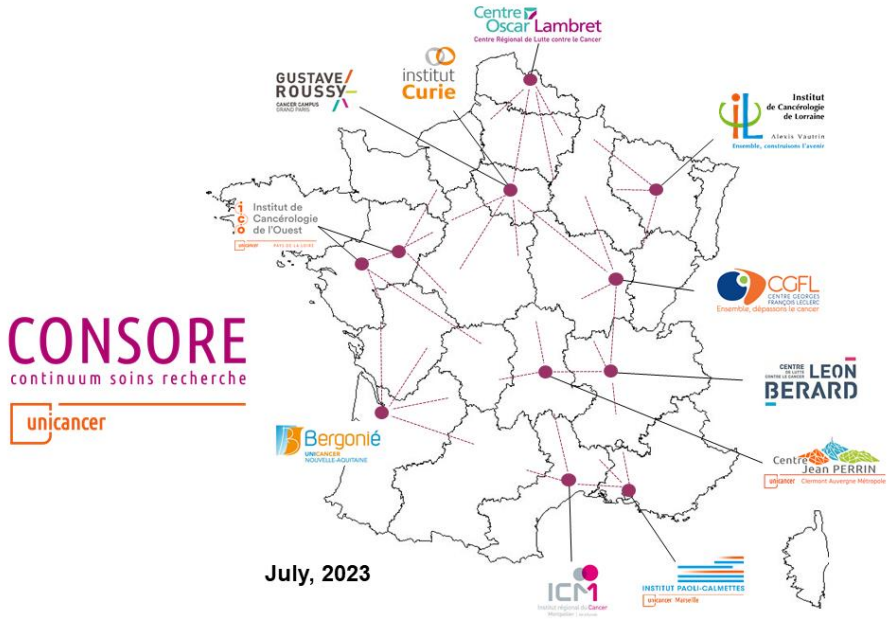


**Figure 3.** The Consore network in France (July 2023).

The following table summarises the number of patients taken cared of and whose data are included in Consore. This is to help represent the volume of data our software deals with.

The difference between the enumeration of patients and those with at least a cancer disease is due to the inclusion into Consore of data of patients who have benign tumors or those who came into a centre for clinical tests that concluded an absence of cancerous disease.

**Table 1.** Overview of data volumes in the different Consore data warehouses (*February 2023;** August 2023).

| French Cancer Centres | Nb of patients | Nb of patients with at least one cancer | Nb of patients with a metastatic relapse | Nb of medical records |
|---|---|---|---|---|
| Institut Curie* | 572 421 | 280 924 | 95 025 | 13 431 874 |
| Centre Léon Bérard* | 359 634 | 207 657 | 85 210 | 18 711 561 |
| Institut Paoli-Calmettes* | 347 415 | 136 500 | 43 767 | 4 464 580 |
| Gustave Roussy* | 399 665 | 237 132 | 96 074 | 12 856 023 |
| Institut de Cancérologie de l'Ouest | N/A (deployment in progress) | N/A | N/A | |
| Centre Oscar Lambret* | 182 436 | 118 506 | 57 784 | 5 865 404 |
| Institut du Cancer de Montpellier* | 176 257 | 79 601 | 34 138 | 3 401 825 |
| Centre Georges-François Leclerc* | 282 948 | 79 592 | 36 635 | 3 207 721 |
| Centre Jean Perrin* | 397 179 | 124 080 | 44 548 | 2 776 005 |
| Institut Bergonié* | 285 129 | 153 589 | 52 290 | 3 806 476 |
| Institut de Cancérologie de Lorraine** | 247869 | 63350 | 19105 | 1 096 485 |

 doi:10.20944/preprints202311.1570.v1



*Cancer of unknown primary*

Cancers of unknown primary (CUP) are metastatic cancers for which a diagnostic work-up fails to identify the site of origin at the time of diagnosis and account for < 5% of all cancers. It is a rare disease and often poorly documented in the EHR and for which the diagnosis is made by default, that is to say when all the primary cancer sites have been eliminated which can sometimes take several months.

*Results at the "Centre Léon Bérard" (CLB)*

After analysis of an "institutional" database and the Consore query, we have a total of 145 CUP at the "Centre Leon Berard". The selection criteria used in Consore were as follows: patients initially diagnosed with de novo metastatic cancer and for whom the mention "unknown primary" was present in the electronic health record since January 1, 2010. The difficulty of this query lies in the fact that it is based both on a textual search but also on data structured by Consore (here the dates of diagnosis and the metastatic stage). Consore offers 2577 patients but with 1 patient from the institutional database not found by Consore. Focusing only on cancer diagnoses between January 1, 2019 and June 30, 2021, we manually reviewed 121 electronic medical records. Concerning these 121 patients, 69 had a CUP while 52 patients finally had a cancer whose primary was identified. The large number of false positives is linked to the period of diagnostic uncertainty which can last several months with cancers of unknown primary (the diagnosis of cancer of unknown primary being ultimately excluded when a primary tumor is identified). So over this period, Consore finds 69 additional diagnoses but detects at the same time 52 false positives.

- To assess the performance, recall, precision [11] and F1 score were calculated as followed [12]:
  - Recall = 99%;
  - Precision = 57%;
  - F1-score = 0,66.
  - calculation of inverse recall is not possible because manual control of all EMRs to identify true negatives is not possible.

*Results at the "Institut Curie" (IC)*

Based on the Consore query first designed at the "Centre Léon Bérard", we identified a selection of 133 CUP at Institut Curie. The primary comparison with the institutional TransCUPtomics [13] study (48 CUP patients) showed an identification of only 8 patients (17%). This poor result is mainly due to the initial criteria : in the TransCUPtomics study, 15 patients were diagnosed before 2010 and 33 patients had a metastatic relapse after the initial diagnosis. However, there was also a lack of keywords to properly identified CUP patients in Consore. In a second step, we have optimized the Consore query by adding the following list of French keywords : "ACUP", "primitif inconnu", "de primitif inconnu", "sans primitif retrouve", "sans primitif connu", "d'origine indéterminée", "pas de primitif retrouve", "d'origine inconnue", "recherche de la tumeur primitive", "autre primitif" (Figure 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2023 doi:10.20944/preprints202311.1570.v1



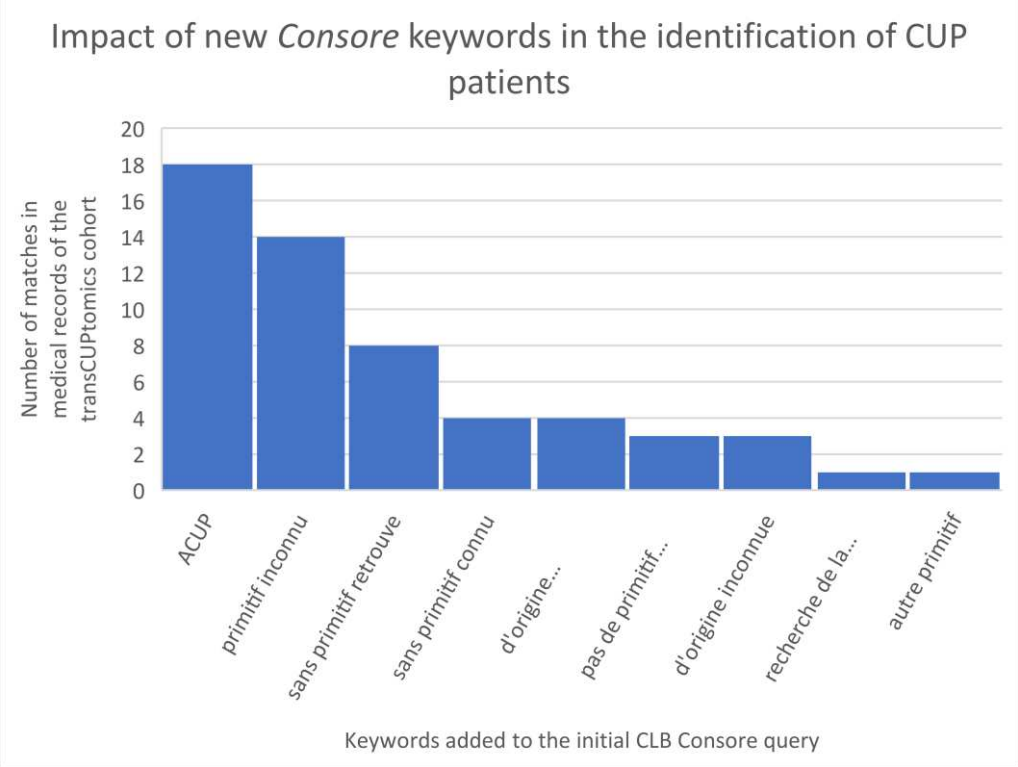


**Figure 4.** Impact of new Consore keywords in the identification of CUP patients at the "Institut Curie" (based on the initial query provided by the "Centre Léon Bérard").

With this new query, we identified 2871 CUP patients on the overall database where we retrieved 45 of the 48 TransCUPtomics patients (94%) : 2 patients remained not found because of a lack of EHRs (less than 5 documents available) , 1 patient had a CUP status unclear with minimal description in the EHR. In order to assess the positive predictive value (PPV), we reviewed a sample of 119 patients diagnosed between January 1, 2019 and June 30, 2021. Concerning these 119 patients, 58 had a CUP while 26 patients finally had a cancer whose primary was identified (false positives linked to the period of diagnostic uncertainty). So over this period, Consore finds 58 additional diagnoses but detects at the same time 61 false positives.

- To assess the performance, Recall, Prevision and F1-score were calculated as followed:
  - Recall = 94%;
  - Precision = 56%;
  - F1-score = 0,7.
  - calculation of inverse recall is not possible because manual control of all EMRs to identify true negatives is not possible.

## 4. Discussion

Considering the growing impact of RWD for clinical practice and research, the development of digital solutions to gather and use such data becomes essential. Beyond the description of this innovative tool, our work underlines the major role of collaborative efforts in constructing a federated technical architecture and agreeing on a unified data model.

With the creation of the OSIRIS model, French cancer centres have laid down a foundation of a minimal data set, which serves as an essential starting point for any collaborative cancer research [7]. Consore, to date, has facilitated numerous research projects on large cohorts of cancer patients, including timely studies on topics such as immunotherapy or COVID-19 [14]. The tool's role in RWD research has been bolstered through collaboration with the Health Data Hub [15], through a project initiative aimed at cross-referencing Consore-provided data with data from the national health data

system [16]. A dialogue has also been initiated to foster communication between this data model and international initiatives, such as OMOP [17].

Numerous French and international consortium are aiming to structure information within medical records to facilitate their reuse for research purposes. Examples include Dr Warehouse [18] and the EHOP warehouse [19] in France, and CancerLinq [20] and FLATIRON [21] in the United States. In this ecosystem, Consore stands out as one of the few solutions dedicated to oncology with a federated architecture, employing natural language processing.

However, this project isn't without limitations. While the federated architecture is a strength, it brings also some issues. Each cancer centre maintains its own EHR, which contains heteregeneous data sources in different formats and varying data quality levels. Although the OSIRIS model allows data format standardisation, quality variability remains a challenge that must be acknowledged, quantified, and considered in the analyses and outcomes of multicentre projects. The question of completeness and quality of the data source as well as the results persist for each research work.

Another identified limitation deals with the common data model, which relies on the primary tumor identification. While suitable for most solid tumors, such as breast or lung cancer, it presents more challenges and therefore errors, for haematological cancers, sarcomas, and skin cancers. For these types of neoplastic diseases, the histological type is indeed more critical than the primary tumor location. Efforts are underway to improve these results.

As an expert tool, Consore is dedicated to oncology and its NLP algorithms have been exclusively trained on French language-based medical reports. A wider dissemination of Consore, at an international level or for other chronic disease (such as chronic obstructive pulmonary disease or chronic renal failure) would require substantial modifications such as the development of an English version or significant data model transformations.

The previously highlighted limitations suggest several potential areas for future enhancement. Special emphasis will be placed on improving the comprehensiveness and quality of the results, as well as bridging the data model into the OMOP-CDM format (Observational Medical Outcomes Partnership - Common Data Model) [16] to facilitate international engagement. The pivotal aspect, thus far, revolves around the update on data quality. Despite our awareness of the inherent constraints of Real-World Data (RWD), we have established a dedicated working group to systematically assess this quality and to enhance the overall robustness of the results.

## 5. Conclusions

Consore stands as a powerful tool dedicated to oncology and capable of modelling patients' neoplastic histories by structuring data from EHRs. Despite the discussed limitations, Consore has already demonstrated its potential in accelerating the identification of patient cohorts and the implementation of research projects. With its federated architecture and the use of a common data model, Consore is playing a key role in the development of multicentric projects in the field of oncology at a national level. The next steps will aim to enhance interoperability between Consore and other international networks as well as the development of next generation NLP algorithms based on large language models (LLM) [22].

**Authors contributions :** Writing – original draft, Julien Guérin, Amine Nahid, Frédérik Joly and Pierre-Etienne Heudel; Writing – review & editing, Louis Tassy, Marc Deloger, François Bocquet, Simon Thézenas, Emmanuel Desandes, Marie-Cécile Le Deley, Xavier DURANDO, Anne Jaffré, Ikram Es Saad, Hugo Crochet, Marie Le Morvan, François Lion, Judith Raimbourg, Oussama Khay, Franck Craynest, Alexia Giro, Yec'han Laizet, Aurélie Bertaut and Alain Livartowski.

**Acknowledgments:** This project received special support from grant ANR-10-EQPX-03 (Equipex), Inca-DGOS-4654 (SiRIC), INCa-DGOS-Inserm_12554 (SiRIC). We thank the ICGEx team (Institut Curie) and the SiRIC-Curie program for their commitment.

## References

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; 2020 (https://gco.iarc.fr/today, accessed on May 2023).

2. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022 Jan;12(1):31-46. doi: 10.1158/2159-8290.CD-21-1059. PMID: 35022204
3. Lainé A, Hanvic B, Ray-Coquard I. Importance of guidelines and networking for the management of rare gynecological cancers. Curr Opin Oncol. 2021 Sep 1;33(5):442-446. doi: 10.1097/CCO.0000000000000760. PMID: 34172594
4. Wilke RA, Berg RL, Peissig P, Kitchner T, Sijercic B, McCarty CA, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. Clinical Medicine & Research. 2007; 5: 1–7.
5. Hersh WR, Weiner MG, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013. PMID: 23774517
6. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
7. Guérin J, Laizet Y, Le Texier V, Chanas L, Rance B, Koeppel F, Lion F, Gourgou S, Martin AL, Tejeda M, Toulmonde M, Cox S, Hess E, Rousseau-Tsangaris M, Jouhet V, Saintigny P. OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology. JCO Clin Cancer Inform. 2021 Mar;5:256-265. doi: 10.1200/CCI.20.00094. PMID: 33720747
8. CCAM. Available online : https://sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/financement/financement-des-etablissements-de-sante-10795/financement-des-etablissements-de-sante-glossaire/article/classification-commune-des-actes-medicaux-ccam, accessed on October 2023
9. World Health Organization. (2004). ICD-10 : international statistical classification of diseases and related health problems : tenth revision, 2nd ed. World Health Organization
10. A. Frizt, C. Percy, A. Jack, K. Shanmagaratnam, L. Sobin, D.M. Parkin, S. Whelan, International Classification of Diseases for Oncology. Third edition. First Revision, World Health Organization, Geneva, 2013
11. Fraser, Alexander & Daniel Marcu (2007). Measuring Word Alignment Quality for Statistical Machine Translation, Computational Linguistics 33(3):293-303
12. Mandrekar J.N. Receiver operating characteristic curve in diagnostic test assessment. J Thoracic Oncol. 2010;5(9):1315–1316
13. Vibert J, Pierron G, Benoist C, Gruel N, Guillemot D, Vincent-Salomon A, Le Tourneau C, Livartowski A, Mariani O, Baulande S, Bidard FC, Delattre O, Waterfall JJ, Watson S. Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics). J Mol Diagn. 2021 Oct;23(10):1380-1392. doi: 10.1016/j.jmoldx.2021.07.009. Epub 2021 Jul 26. PMID: 34325056.
14. Heudel P, Favier B, Solodky ML, et al. Survival and risk of COVID-19 after SARS-COV-2 vaccination in a series of 2391 cancer patients. Eur J Cancer. 2022 Apr;165:174-183. doi: 10.1016/j.ejca.2022.01.035. Epub 2022 Feb 10. PMID: 35245864 Free PMC article.
15. Health data Hub. Available online: https://www.health-data-hub.fr/page/faq-english , accessed on May 2023
16. Health data Hub, UNIBASE results.. Available online: https://www.health-data-hub.fr/annonce-laureats-unibase , accessed on May 2023, French version only.
17. OHDSI. Available online: https://www.ohdsi.org/data-standardization/the-common-data-model/ , accessed on May 2023.
18. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, Munnich A, Burgun A, Rance B. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018 Apr;80:52-63. doi: 10.1016/j.jbi.2018.02.019. Epub 2018 Mar 1. PMID: 29501921
19. Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML, Delamarre D, Raimbert V, Lemordant P, Cuggia M. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019 Aug 21;264:1536-1537. doi: 10.3233/SHTI190522. PMID: 31438219
20. CancerLinq. Available online: https://www.cancerlinq.org/ accessed on May 2023.
21. Flatiron. Available online: https://flatiron.com/ accessed on May 2023.
22. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8;9:e45312. doi: 10.2196/45312. PMID: 36753318