WHITE PAPER

Big Data Analysis of Clinical Records for Cancer Care

UNICANCER collaborates with Sword Group and Intel to drive cancer insights through next-generation analysis of clinical data

Executive Summary

Advances in big data analytics are giving cancer researchers powerful new ways to extract value from diverse sources of data. UNICANCER, one of Europe's largest cancer research organizations, has demonstrated a next-generation data analytics solution with the potential to improve clinical trials, health economics, clinician/researcher productivity, and potentially patient care. The solution—called Continuum Soins Recherche (ConSoRe*) or Continuum of Care Research—has been deployed in a pilot project at four UNICANCER centers:

- Centre Georges François Leclerc (CGFL) in Dijon
- Centre Léon Bérard (CLB) in Lyon
- Institut Curie in Paris
- Institut du Cancer de Montpellier (ICM) in Montpellier

The pilot solution aggregates diverse forms of structured and unstructured data extracted from 24 million documents from the privacy-protected records of 1.25 million patients from four cancer centers. The solution then uses natural language processing (NLP) to search the aggregated data and perform advanced data mining. UNICANCER is developing the ConSoRe solution in collaboration with Sword Group, an international consulting firm.

By enabling sophisticated mining of diverse data sources, ConSoRe aims to move cancer breakthroughs more quickly along the continuum from the laboratory bench to clinical treatments at the bedside or pharmacy.

ConSoRe runs on Intel® technologies and uses the Luxid* Annotation Server from Expert System, along with a variety of open source and other technologies. Development work is ongoing. The next phase will concentrate on deepening the search capabilities, refining the user interface, further optimizing performance, and expanding the number of users and locations.

"Cancer is not one disease but a multitude of orphan diseases... Looking at just the data from a single center, we will not be able to predict the evolution of each patient's disease. We need to look at the data from a lot of sites, or even look at a global level."

–Dr. Alain Livartowski, Oncologist and Information Systems Leader, Curie Institute

Table of Contents

Executive Summary 1
Introduction: Getting More Value from Healthcare Data 2
Cancer as a Big Data Challenge 2
A Tool for Creating Value from Data 3
Solution Requirements and Issues 3
ConSoRe Solution Architecture 4
Homogenizing the Data and Applying Natural Language Processing 4
Interacting with the User 5
Intel® Technologies and Expertise 7
Looking to the Future 8

"Before the test, our reference time for processing the whole corpus of patient data was a little over 11 days. After parallelizing our processing pipeline to harness the processing power of the cluster, ...we reached a velocity enabling us to process the whole corpus in less than four days. We are doing further work and believe we may be able to run the full body of patient data in less than a day"

–Frederik Joly
Project Director,
Sword Group

Introduction: Getting More Value from Healthcare Data

Accurate and complete information is essential for medical research, clinical care, resource allocation, and operational excellence. Yet healthcare is often hampered by incomplete and incompatible data. Because the industry lacks mature information standards, important information is often locked in diverse systems with limited interoperability. Patients' electronic medical record (EMR) systems typically include only a subset of relevant data, and crucial information is buried in text-based case notes, scanned documents, and other unstructured data sources. Different hospitals use different EMR systems, making it difficult to get a unified view of a single patient's health and even harder to assemble a cohort of patients for a clinical trial or to analyze a rare disease.

"The problem is not insufficient data," says Eddy Willm, director of information systems at ICM (Montpellier). "We spend a great deal of time entering information in a wide range of systems, and we have compiled massive volumes of potentially valuable data. The problem is that the data is fragmented and isolated. We need new tools that are more advanced in terms of knowledge—that bring the data together, show us the relationships between information, and provide us with answers. Then, the data can begin to deliver its full value."

Willm continues, "Approximately 80 percent of clinical data and patient healthcare information are text based and stored in text document format. If, with annotation and indexes, we put these data in columns of a database, this unstructured data become structured data. The clinical mining 'magically' appears and unstructured data becomes analyzable for research. That's what we did with our hospital medical records during the first step of the ConSoRe project at the four pilot sites."

Cancer as a Big Data Challenge

In its innovative work to aggregate and analyze unstructured data, the UNICANCER Group is taking steps to improve cancer care in France and chart a path for researchers around the world. UNICANCER is a network of 20 French Cancer Centers—private, non-profit health establishments located across France and devoted exclusively to fighting cancer. Each center has a threefold mission of patient care, research, and teaching, and 300 clinical trials are currently conducted within the group. UNICANCER has revenues of 2.1 billion Euros annually.[1]

Cancer is a leading cause of death in France, as it is in many countries. An estimated 371,700 people in France were diagnosed with cancer (excluding non-melanoma skin cancers) in 2012, with an age-standardized rate of 324.6 diagnoses per 100,000 people.[2] An estimated 120,000 cancer patients are hospitalized at UNICANCER centers each year.

Despite decades of work and significant investments, the specific causes of cancer are still poorly understood, and treatment planning often proceeds through trial and error. Part of the problem lies in the great variety of ways cancer attacks the body.

"Cancer is not one disease but a multitude of orphan diseases," explains Dr. Alain Livartowski, an oncologist and information systems leader at the Curie Institute. Based in Paris, the Curie Institute is a French Cancer Center with roots in the laboratory of Marie Curie. The Curie Institute has been a paperless hospital since 2004 and a filmless hospital since 2007, and Dr. Livartowski coordinated the launch of the EMR at the Institute.

"In every hospital, few cases of cancer will look alike," Dr. Livartowski continues. "Even in a large cancer center, we will have a multitude of patients with very different diseases from each other. Their vital prognosis will be very different as well. Looking at just the data from a single center, we will not be able to predict the evolution of each patient's disease. We need to look at the data from a lot of sites, or even look at a global level."

Adding to the data challenges, researchers need to examine cancer from many angles. Relevant data can include environmental and epidemiological studies, genomic profiles, cellular and molecular data, tumor samples, patient treatment histories, imaging studies, published studies of treatment results, and much more.

When research results in potential treatment breakthroughs, the massive volumes of fragmented data force researchers to manually search through enormous volumes of data to assemble a cohort of patients with the precise combination of factors to meet the criteria for a clinical trial. This process has helped make patient recruitment the most expensive aspect of clinical trials, responsible for 32 percent of the total trials cost.[3] In addition to driving up costs, these data problems cause delays in bringing potentially life-saving innovations to the patient.

A Tool for Creating Value from Data

UNICANCER's ConSoRe solution provides more meaningful data for cancer research by combining diverse data from multiple French Cancer Centers and using natural language programming to search the data. The pilot project combined data from 24 million documents for 1.25 million patients. UNICANCER estimates that ConSoRe will extend to a total of 4 to 5 million patients and at least 70 million documents when it is fully deployed in the 20 French Cancer Centers. Pilot data includes:

- Structured and unstructured information in patient medical records
- Demographic and administrative data
- Medical activity drawn from France's nationwide diagnostic-related group (DRG)-based information system known as the Medical Program Information System (Programme de Médicalisation des Systèmes d'Information or PMSI)
- Biobank data
- Tumor characteristics
- Pharmaceutical data

"ConSoRe is a tool for creating value from healthcare data," says Dr. Pierre Heudel, oncologist at Centre Léon Bérard (Lyon). "Identifying cohorts of patients for clinical trials is a primary way of gaining value. It is time consuming to create manually a database, and it is complicated to maintain it. Without this work, the database is quickly obsolete. We recently had a metastatic breast cancer research project where it took 30 people reviewing patient records for six months to assemble a cohort of patients who had been treated in one of the 20 French Cancer Centers. We believe ConSoRe will help us do that within a matter of hours or days."

Dr. Heudel identifies numerous other benefits. "ConSoRe will also help us make correlations between several patient cases across cancer centers, go deeper in analyzing a particular patient case, and conduct medico-economic or epidemiological studies with our data," he says. "Big data technology offers new opportunities to pool diverse sources of data and interpret it to optimize our understanding and treatment of cancer."

ConSoRe has demonstrated value even in its early pilot phase. "In one case, we received a safety alert notifying us that some specific breast implants were likely to be the cause of a rare form of lymphoma—breast anaplastic large-cell lymphoma or ALCL," Dr. Livartowski recalls. "Using ConSoRe, we were able to pool data from multiple sites and look at large enough numbers of cases to determine that the occurrence of ALCL with those implants was actually a very rare event. And the response of the ConSoRe service was very fast. Previously, it would have been practically impossible to get the answer."

Solution Requirements and Issues

The ConSoRe development team addressed a wide range of issues and requirements as it designed and implemented the solution. For example:

- **Data heterogeneity.** ConSoRe must aggregate different types of data sources from multiple hospitals. Hospitals use different software, vocabularies, and data values, so data would be in a variety of formats for a given source type. In the ALCL example cited above, one hospital might code the relevant data under "Breast," and another under "Lymphoma." Some use the plain name of a pathology and others use the ICD-10 code.
- **Patient privacy.** Big data solutions for healthcare must place a priority on protecting patient privacy while enabling researchers and clinicians to maximize the value of the data. ConSoRe had to be designed to ensure that confidential data is safeguarded and is accessed only by authorized individuals. The solution must also meet the varying access requirements of its user community. Clinicians need access to patient-identified data to ensure a personalized therapy for their patients. Researchers do not require access to specific patient identities, and can work successfully with de-identified data.

- **Text documents.** Some of the most valuable information in an individual's medical documents is found only in plain text documents. ConSoRe must analyze the "full text content of documents" and extract relevant meaning from scanned/handwritten clinical case notes and other unstructured data.
- **Data storage.** French Cancer Centers need to retain control of their own data. The team also wanted to avoid the expense of building and maintaining a central data warehouse that would import data from all hospitals.
- **Performance for massive data volumes.** A French Cancer Center typically has between 100,000 and 600,000 patient records, each containing up to 100 documents. ConSoRe would have to index several million documents per hospital, and this work would need to be performed when initializing the system, as well as periodically when new versions of natural language patterns, reasoning rules, and so forth are released.
- **Ease of use.** Despite the heterogeneity of sources and complexity of the data handled, interrogation must remain simple, flexible, and powerful for busy end users.

ConSoRe Solution Architecture

UNICANCER is developing the ConSoRe solution with Intel and Sword Group, an international company established in 2000 and offering consulting, services, and software. Sword Group operates in 50 countries, handling complex IT projects that help organizations make strategic use of information technology and evolve their business processes. Sword's expertise in natural language processing has been important to the success of the ConSoRe project.

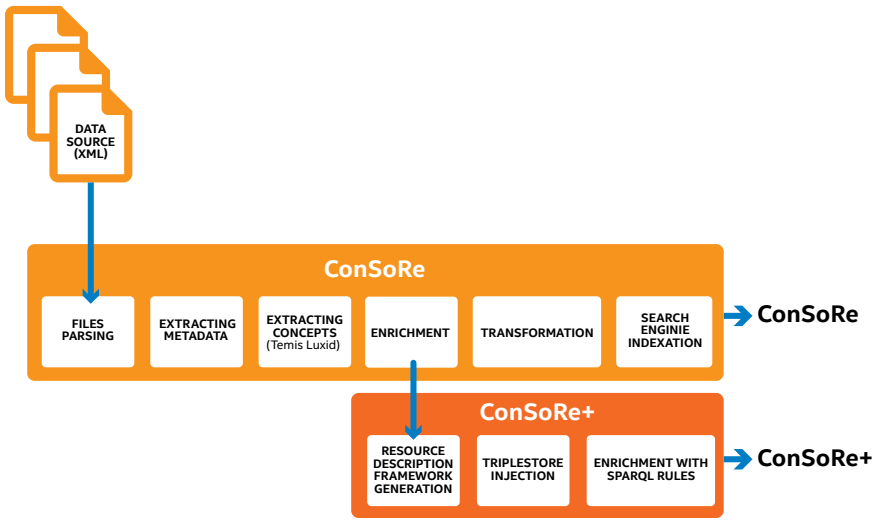

Figure 1. ConSoRe Processing Pipeline.

UNICANCER and Sword are meeting their design objectives through a flexible architecture that uses semantic web tools and NLP to enable meaningful search of diverse document types. Figure 1 provides an overview of the ConSoRe processing pipeline, including functionality that is being added in the next phase of the project.

ConSoRe uses a federated data architecture, so that each center maintains control of its own records. This helps safeguard privacy, maintain the integrity of the data, and avoid the issues of a central data warehouse. Users can query one cancer center or several centers simultaneously to find patient records meeting the search criteria. To safeguard privacy, UNICANCER leaders held discussions with the Commission Nationale de l'informatique et des Libertés[4] (CNIL), the French data privacy agency, and established written best practice rules.

The solution architecture is implemented with open source technologies and leading-edge proprietary software, running on Intel® technology-based platforms. The resulting solution delivers the reliability needed for healthcare computing, and the intelligence, performance, scalability, and capacity to quickly process and analyze massive data sets.

Homogenizing the Data and Applying Natural Language Processing

A key challenge for ConSoRe is to standardize the various data types at the levels of structure, type, and content, including homogenizing the different vocabularies in place at different institutions or even among different users. To help accomplish this, ConSoRe transforms the hospitals' specific formats to a unique pivot format, which is used for subsequent manipulation and analytics. ConSoRe normalizes the values with a single thesaurus based on industry standards such as ADICAP, ICD-10, CIMO-3, and drug formularies.

Instead of the traditional Extract-Transform-Load (ETL) approach, ConSoRe uses an Extract-Transform-Enrich-Publish (ETEP) chain designed to handle both structured data (such as SQL and NoSQL databases, XML files, and Integrating the Healthcare Enterprise (IHE) formatted data), as well as unstructured text files. The implementation chain uses Apache* Camel,* an open source integration framework that performs routing and mediation.

In addition to standardizing the various field values, ConSoRe applies reasoning rules and NLP to analyze unstructured content, extract relevant information, and semantically enrich content with domain-specific metadata. ConSoRe's search engine pipeline is implemented by two primary elements. The Expert System Luxid* Annotation Server uses patented NLP technology to enrich the data with semantic meaning. The results are published to Elasticsearch,* a distributed, open source search and analytics engine. OpenLink Virtuoso* Universal Server is used for storing the medical standards and medical ontology.

Extracting medical concepts from unstructured documents is a complex process that must handle negations, variations, conjugated verb forms, and local variations. The development team implemented an iterative process to refine, test, and improve ConSoRe's NLP detection patterns. ConSoRe handles advanced forms such as negations, family history, and tumors.

Other rules are applied in the processing chain to enrich the data, such as creating links between concepts and consolidating data at the level of an individual patient (for example, calculating the cumulative doses of chemotherapy a patient received to help identify potential side effects for the treatment.) The ConSoRe solution can thus draw inferences and create new knowledge to enrich patient information. For example, if ConSoRe detects the symptom of diabetes and a prescription for typical diabetic medications, it marks the record as diabetic, even if patient's file mentioned only the symptom and treatment and did not include the formal diagnosis. ConSoRe users can then view the relevant files to confirm the inferred diagnosis.

Interacting with the User

Users interact with ConSoRe through a secure web portal based on Liferay* Portal, a leading open source solution. The ConSoRe portal offers a graphical user interface based on the Apache* Tomcat* open source implementation of the Java* Servlet.

With the ConSoRe portal, users do not have to be concerned about the format of the underlying data. From the same home page, they can simultaneously query both structured and textual data

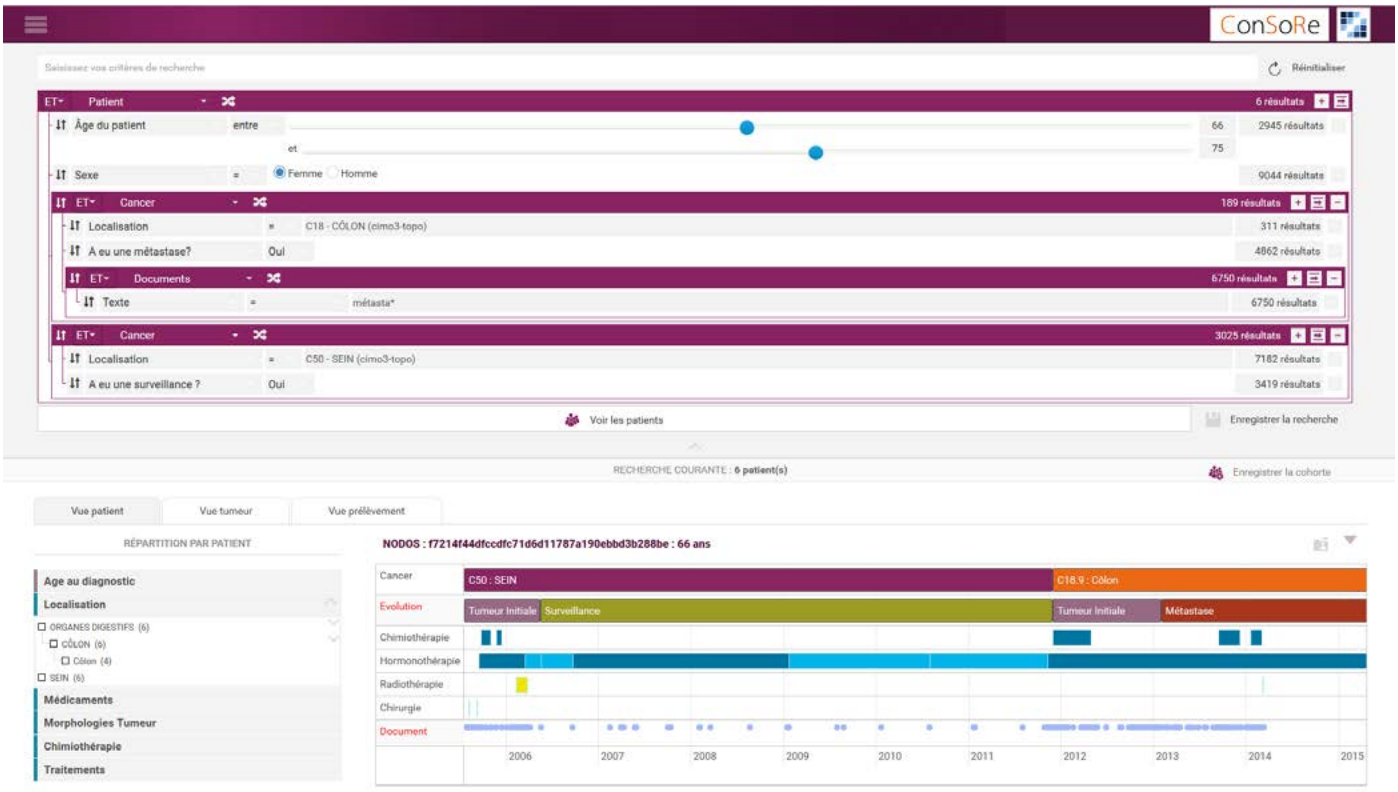

Figure 2. Query with results.

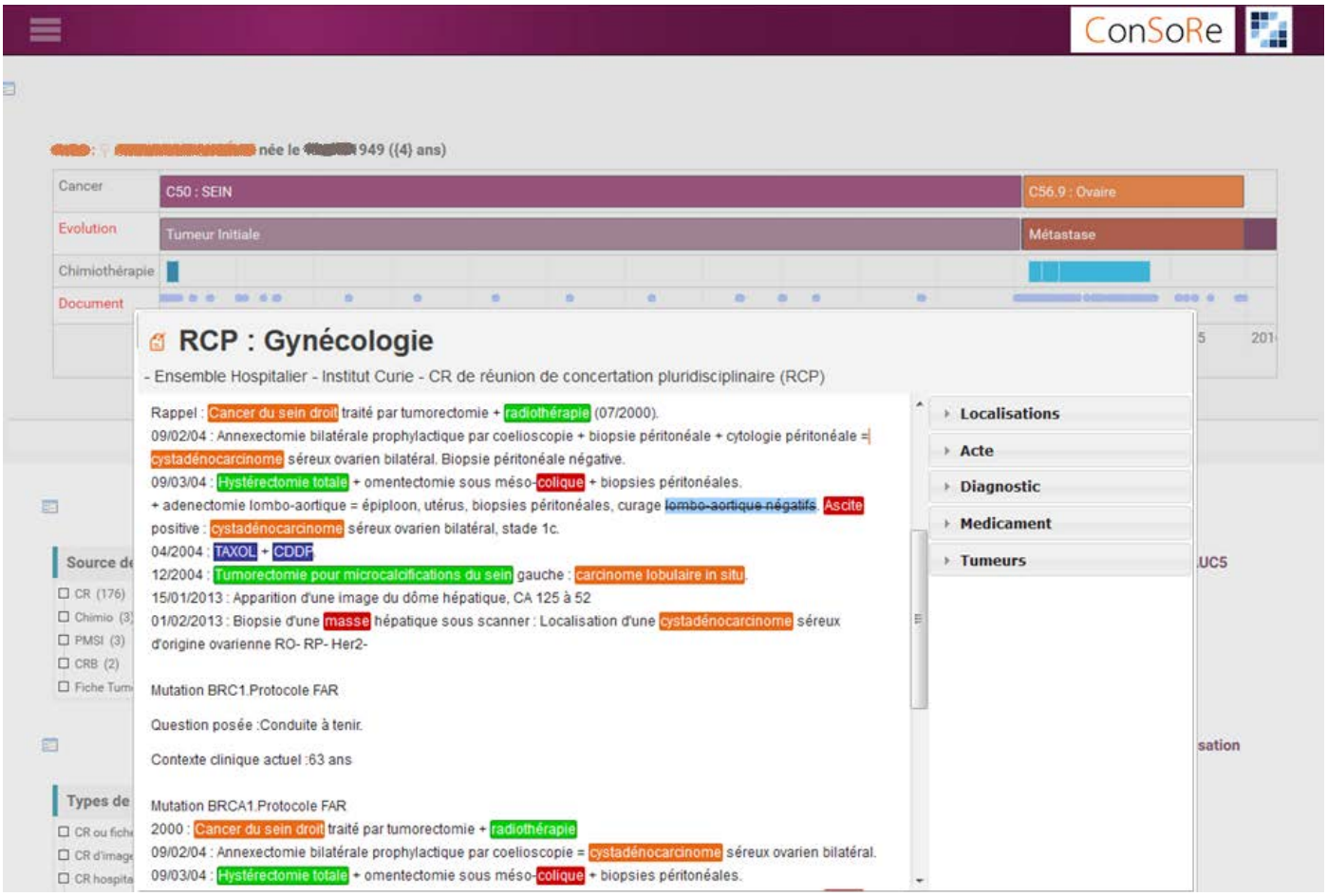

Figure 3. View of patient data.

coming from multiple databases. Users can select a set of search criteria that may include medical data (history, medications, tumor characteristics, etc.), marital status (age, gender), administrative details (such as the date of last visit), and other information. Search terms can also be chosen based on a free text search, and Boolean operators such as AND and OR can be used to narrow down the search criteria. Inferred data as well as structured data can be used as search criteria to select patients for a cohort. Figure 2 shows an example of the ConSoRe graphical user interface with a sample query and results.

Because of the federated data architecture, users can query their own center's data stores first, and then search those of other centers simultaneously as needed. The local search provides a list of subjects that fulfill the criteria, including a short summary of patient information and a list of suggested terms to help narrow down the search. Searches that include multiple centers provide a count of patients meeting the criteria and the centers where these patients are treated but do not divulge protected information.

Running on Intel technologies, ConSoRe's analysis to identify candidates meeting the user's criteria is very quick, with a list being reported back to the user almost instantly. Users can directly view the indexed documents or records to validate the system's conclusions. In the next generation of ConSoRe, information can be restructured and presented on a timeline, enabling users to understand a patient's history quickly as well as to review key documents if an extra level of scrutiny is needed. Figure 3 shows a sample view of patient data.

TABLE 1. HIGH-PERFORMANCE CLUSTER CONFIGURATION

Head Server Node (1)	Intel® Xeon® processor E5-2690 v4 (35M Cache, 2.60 GHz 135 W TDP 14 core M0 step QK8X) (2) 128 GB DDR4 (8x 16 GB) at 2400 MHz (1) SAS/SATA 1 TB Hard Drives in RSTe RAID 5 or RAID Z ~ 3 TB (4) Intel® SSD P3600 800 GB 2.5" NVME drives (3) Intel® Omni-Path Host Fabric Interface x16 PCI-E adapter card (1)
Execution Server Nodes (6)	Intel Xeon processor E5-2690 v4 (35 M Cache, 2.60 GHz 135 W TDP 14 core M0 step QK8X) (2) 128 GB DDR4 (8x 16 GB) at 2400 MHz (1) 300 GB SAS Hard Drives hardware RAID 5 or RAID Z ~ 3 TB (3) Intel SSD P3600 800 GB 2.5" NVME drives (1) Intel Omni-Path Host Fabric Interface x16 PCI-E adapter card (1)
Infrastructure (1)	48-port Omni Path Edge Switch (1)

Note. This table shows the configuration of the test cluster. The number and configuration of the SAS/SATA drives in Sword's recommended configuration may change depending on further testing, and are provided here as guidance.

Intel® Technologies and Expertise

ConSoRe's main processing pipeline is highly compute-intensive to execute the solution's sophisticated NLP algorithms and processes the enormous data volumes. UNICANCER and Sword have developed ConSoRe for Intel technologies, and ran ConSoRe on older-generation Intel® processor-based systems and smaller clusters.

With ConSoRe moving toward broader use, the development team wanted to develop a potential future infrastructure that would give their data-intensive search engine the performance and throughput advantages of the latest Intel technologies. Intel arranged for the team to test ConSoRe on a heterogeneous cluster powered by the Intel® Xeon® processor E5-2690 v4 family, which benefits from high core counts (14 cores and 28 threads per processor) and large memory capacity. Complementing the high-performance processors, the cluster also includes Intel® Data Center Family P3600 Solid-State Drives (Intel® SSDs) as cache accelerators. The Intel SSDs are balanced with the use of Intel® Omni-Path Host Fabric Interface and Intel® Omni-Path Edge Switch to enable future flexibility.

Sword developers reported dramatic performance benefits from running their processing pipeline on the Intel technology-based cluster. "The cluster provided by Intel exceeds by a few orders of magnitude the minimum requirements of the ConSoRe platform and even exceeds the power of our biggest installation, which is currently hosted at the Curie Institute," explains Frederik Joly, project director at Sword Group. "Before the test, our reference time for processing the whole corpus of patient data was a little over 11 days. After parallelizing our processing pipeline to harness the processing power of the cluster, we identified some significant bottlenecks in our current implementation of ConSoRe. We greatly improved the effectiveness of the whole processing pipeline, and reached a velocity enabling us to process the whole corpus in less than four days. We are doing further work and believe we may be able to run the full body of patient data in less than a day."

By basing their big data server and storage infrastructure on Intel technologies, healthcare organizations such as UNICANCER are assured of outstanding performance, high reliability, and a roadmap of innovations going forward. In addition to developing its product technologies, Intel works with health and science leaders to understand application requirements and design its hardware to meet them. Intel also engages in a broad range of collaborations to identify and address critical challenges in healthcare, the life sciences, and data analytics. These include optimizing widely used life science research codes, disseminating best practices, leading and participating in standards organizations, advancing open-source and open-platform solutions.

"We recently had a metastatic breast cancer research project where it took 30 people reviewing patient records for six months to assemble a cohort of patients who had been treated in one of the 20 French Cancer Centers. We believe ConSoRe will help us do that within a matter of hours or days."

–Dr. Pierre Heudel
Oncologist,
Centre Léon Bérard

"Using ConSoRe, we were able to pool data from multiple sites and look at large enough numbers of cases to determine that the occurrence of ALCL with those implants was actually a very rare event. And the response of the ConSoRe service was very fast. Previously, it would have been practically impossible to get the answer."

–Dr. Alain Livartowski, Oncologist and Information Systems Leader, Curie Institute

Looking to the Future

Because of solutions such as ConSoRe, the next big breakthrough in cancer treatment may owe as much to big data analytics as to the lab or petri dish. By bringing together diverse data and enabling researchers to derive more value from their data, ConSoRe and other big data solutions can help move breakthroughs more quickly from laboratory bench to the bedside, where they can improve diagnosis, treatment planning, and patient care. ConSoRe specifically shows great promise to improve epidemiological studies, accelerate patient trials, and reduce the high costs of patient recruitment.

"ConSoRe will help us improve pharmaceutical trials, health economics, clinician and researcher productivity, and physician training," says Dr. Livartowski. "As we learn more about the way cancer has progressed in each patient and about the effectiveness of each treatment, there will be less need for trial-and-error approaches. We will be able to say with greater certainty when treatment is necessary and which treatment will be most effective. This will help us avoid unnecessary expense and suffering."

Collaboration continues among UNICANCER, Sword, and Intel to deepen ConSoRe's search capabilities and refine the user interface. Future capabilities may also include the ability to provide treatment recommendations. UNICANCER plans to deploy ConSoRe to all French Cancer Centers and, in the longer term, to extend its use for collaboration to university hospitals in France and worldwide.

Learn More

- Explore Intel® technologies for big data in healthcare.
- Read about UNICANCER and Sword Group.
- Follow us on Twitter: @IntelHealth, @GroupeUNICANCER, @Sword_Group
- Join the conversation in the Intel Health and Life Sciences community.

Intel appreciates the following people who contributed to this paper:

- **Alain Livartowski,** Curie Institute
- **Pierre Heudel,** Centre Léon Bérard
- **Eddy Willm,** Institut du Cancer de Montpellier
- **Christophe Jamain and Emmanuel Reyrat,** UNICANCER
- **Edouard Barthuet, Guillaume Darves-Bornoz, and Frederik Joly,** Sword Group
- **Valere Dussaux, Kristina Kermanshahche, Malcolm Linington, and Jamie Wilcox,** Intel

[1] UNICANCER Group, Key Figures. http://www.unicancer.fr/en/unicancer-group/key-figures

[2] http://www.cancerindex.org/clinks5f.htm

[3] Cutting Edge Information, Patient Recruitment and Clinical Vendor Fees Top Clinical Trial Cost Drivers, http://www.cuttingedgeinfo.com/2011/clinical-trial-cost-drivers/.

[4] For information about CNIL, see https://www.cnil.fr/en/home.