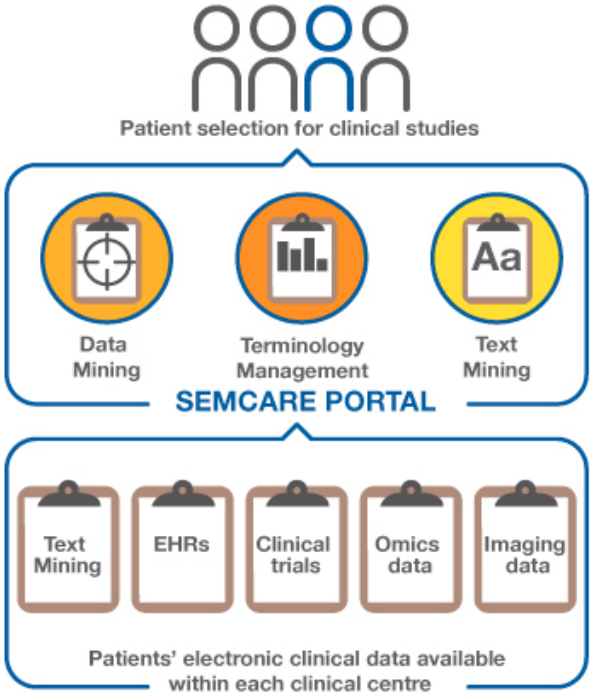
1. PUBLISHABLE SUMMARY

1.1. PROJECT SUMMARY OF SEMCARE (WWW.SEMCARE.EU)



The need for exploiting medical data for clinical trials and to monitor and improve healthcare delivery has grown tremendously over the last years . Aggregated patient-level data can support the identification of disease mechanisms and new discovery areas, improve drug safety surveillance through continuous monitoring, and decrease patient recruitment cycle times for clinical trials . Currently, almost 80% of clinical trials fail to meet their patient enrolment quotas on time, causing delays in bringing new drugs to market . Exploiting patient-level data can optimize clinical studies in several ways, e.g. by enabling the definition of appropriate study design or ensuring that inclusion/exclusion criteria map to an existing patient population . As large parts of patient-level data in electronic health records (EHRs) are only available as free text, language technologies are an indispensable prerequisite for this process.

SEMCARE aims to build a semantic data platform that is able to identify patient cohorts based on clinical criteria (e.g. age, gender, diagnosis, indication, symptoms, lab results) scattered in heterogeneous clinical resources. This platform combines the power of full-text search with text analytics and semantic web technologies for a hybrid semantic full-text search. Search and filter capabilities, such as faceted search, range queries etc. known from advanced search engines like Apache Solr, are made available. The platform enables semantic integration of heterogeneous, unstructured and structured data sources for information extraction, document retrieval, healthcare analytics, as well as data visualization and exploration. Users are able to pose complex queries without concern over how the data have been recorded.

SEMCARE integrates state-of-the-art text mining technologies and multilingual semantic resources (domain vocabularies, terminologies, nomenclatures, classifications, ontologies; in the following referred to as terminologies) to address specific idiosyncrasies of medical language like ambiguous terms, acronyms, compounds, derivations, spelling variants, uncorrected spelling errors, jargon, telegram style. It also focuses on the mining of quantitative data (e.g., lab results, drug dosages, dates/times) from unstructured text. Sentiment analysis will detect contextual clues for distinguishing facts from opinions and plans, identify grades of diagnostic certainty, and pinpoint

negations and temporal contexts. Interfaces based on open standards such as HL7 CDA will facilitate integration within existing hospital information systems. However, the basic functionality only requires the access to the text, regardless of its format (PDF, RTF, TXT…).

1.2. MAIN ACHIEVEMENTS IN YEAR 1

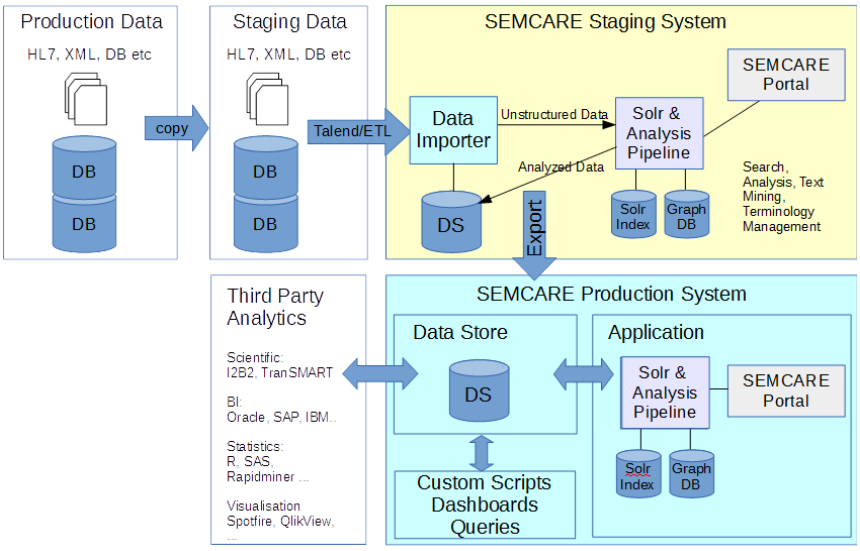
Our goal of implementing a semantic data platform for clinical research and clinical trial protocol feasibility in two years' time is an ambitious task that can only be achieved because many of the individual blocks are already available at the partners' sites. We envisage a two-phase approach:

- **Phase 1** (year 1) focuses on the implementation of the system;
- **Phase 2** (year 2) covers the evaluation and continuous improvement of the system in the hospitals.

1.2.1. Use-Case

In the beginning, we described in detail a representative use case for SEMCARE. The three participating European health centres have agreed on one first general use case on which they will focus during the project. The use case is called 'Risk Stratification and Differential Diagnosis of Patients suffering from transient loss of consciousness (T-LOC)'. A number of phenotypic features can help risk stratify T-LOC patients, most of which are available from routine assessment and investigations. The SEMCARE semantic data platform seeks to identify high-risk patient cohorts based on patient-level, criteria scattered in heterogeneous, predominantly textual EHR sources.

1.2.2. Architecture & Requirements



Next, we collected the functional and technical requirements, and designed and presented in detail a generic software architecture according to the objectives of SEMCARE and that conforms to the collected technical requirements, does not affect existing clinical information systems, and respects patient privacy.

1.2.3. Terminologies

Our terminology team addressed two main tasks. First, we generated an initial multilingual biomedical terminology for the SEMCARE platform. This terminology is based on the UMLS and

includes English, German, and Dutch terms. Software was developed to convert the UMLS data into the OBO format for easy integration in the SEMCARE terminology management system. Using the SEMCARE use case description, an assessment of term coverage in the different languages has been made. Also several approaches for automatic term acquisition in German and Dutch have been explored.

Second, we developed software to support the management of terminologies that are needed for the SEMCARE platform. The terminology management software includes routines to import and export terminologies, to easily browse and inspect the term hierarchies, and to add or modify concepts and terms. A prototype system and user manual is available.

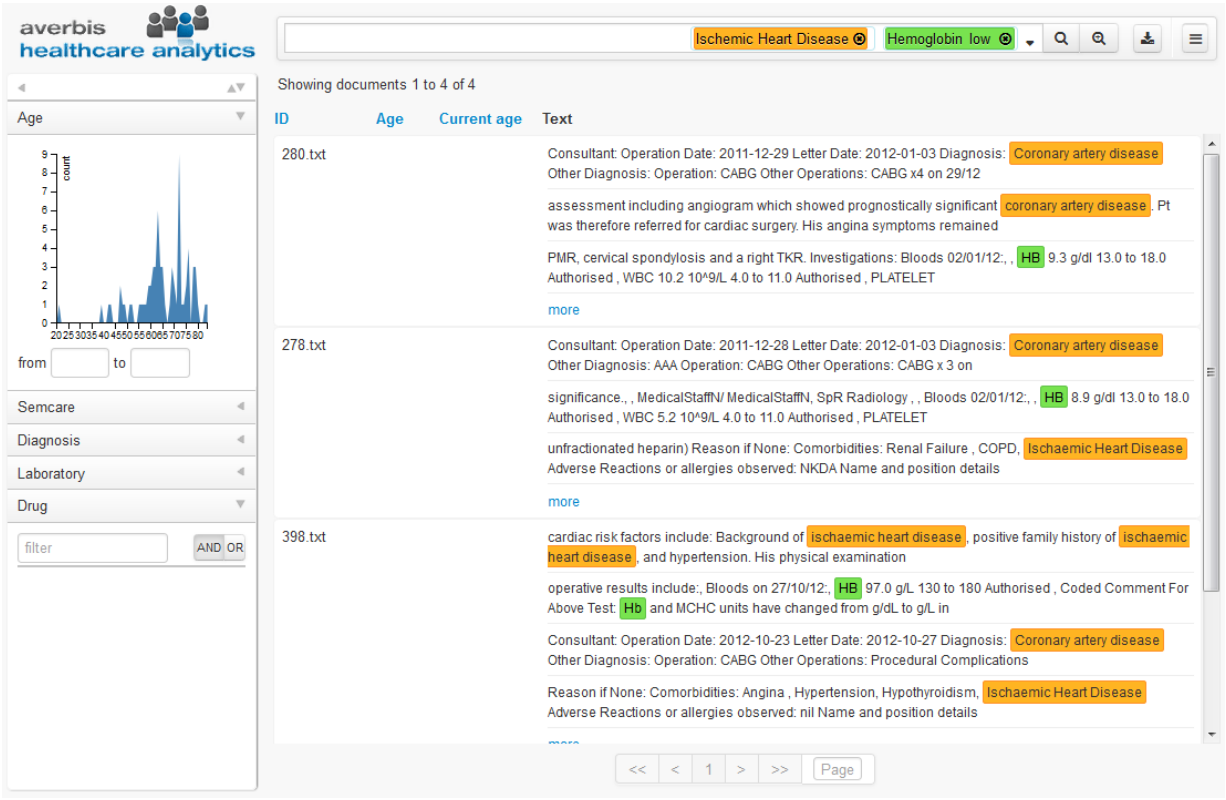
1.2.4. Text-Mining & Search

In the SEMCARE platform the Averbis Extraction Platform (AEP) analyses the unstructured text to extract information unit such as facts, evaluation of facts and relations. Apache UIMA (Unstructured Information Management Architecture) is used as text-mining framework. Integral part of the Averbis Extraction Platform is the Averbis Type System – a UIMA compliant annotation type system covering all levels of text analysis including structural, syntactic and semantic processing of textual documents. For SEMCARE we extended the Averbis Type System with a «medical extension pack». We defined the data types diagnosis, qualifier, observables, drug, medication, laboratory values and ECG. All these data types are subtypes of the data type concept. To each data type a terminology is assigned which is used for the mapping of free text phrases to concepts (terminology binding).

Over the last year, we developed and improved various medical Text mining taggers (Analysis Engines, AE):

- The Medication AE creates mappings between concepts of the provided drug terminology and free-text phrases. Included in this annotation are the unit, the dosage, the ingredient and trade name, and any comments if available.
- Laboratory AE: Detects Lab Values together with numerical values and units
- Qualifier AE: descriptive or evaluative qualifiers that are defined in the terminology SEMCARE Modifier, are mapped to concepts from the SEMCARE terminology found in the free-text phrases.
- ECG AE: This AE is handling different ECG data from a patient: the type of ECG performed, the position and type of recording, the observed heart frequency and heart rhythm, electrophysiological findings and resulting diagnosis.

The results of the text analytics platform are stored in the Averbis Search Platform, an SOLR based search engine. It provides a faceted search which means that the search results are organized according to a faceted classification system, thus allowing the user to explore a collection of information by applying multiple filters. In the SEMCARE interface provides the following facets: Age, SEMCARE terms, Diagnosis, Laboratoy, Medication. The search results are highlighted per facet in a different background colour.



1.2.5. Quality & Ethics

Quality management is a priority in the SEMCARE consortium. High quality standards have been established throughout all steps undertaken within the project. They are applied on the technical level by creating and monitoring the use of formal procedures for quality assurance on all activities in the project, as well as by developing early in the project quality guidelines that affect all work and procedures that need to be implemented.

Quality assurance also entails the continuous monitoring of the project with respect to the degree of fulfilment of its objectives and validation of its scope. Additionally, we continuously evaluate the strengths and weaknesses of the project as they evolve during its duration. In that sense, they will also be related to the risk management tasks.

Due to the sensitive nature of the personal health data it is important for SEMCARE to be fully aware of ethical and regulatory aspects and to implement all reasonable measures to ensure compliance with ethical and regulatory issues on privacy.

The local authorities in the hospitals play an important role for the definition of security needs within the clinic. The project SEMCARE assumes that corresponding actions for privacy protection are in place in the data providing clinics by means of already running clinical systems. The locally installed SEMCARE services will follow these existing rules.

Furthermore, additional actions are taken in order to protect the processed data, especially with regards to the project related infrastructure. These are e.g. rules for data deletion or an additional protection of system access and data transfer. Additionally, a de-identification of the patient data takes place at the sites in order to assure data privacy.

1.2.6. Expected final results, potential impact and use

In Phase 2 we plan to test and improve the prototype in clinical routine. Terminologies will be adapted to the clinicians needs, and the text-mining components shall be improved in quality and coverage.

Utmost importance will be laid on a successful implementation of a clinical use case. As hospitals are moving towards a fully digital health record, the SEMCARE software will be ideally placed to study using real world data the patterns, presentation and outcomes from different diseases on a scale that has been hitherto difficult to study in a cost effective manner. The outputs of this will inform healthcare nationally and internationally and have implications for the understanding of diseases in developing countries. In the longer term, hospitals will have a well phenotyped patient population which they can screen for eligibility and recruitment into clinical trials involving drugs and devices to reduce the overall cost of doing such trials.

Our goal is to make the semantic data platform prototype ready to reach the market soon after the project ends. Our long term goal is to build a pan-European supported diagnosing and phenotyping platform for various, especially rare diseases.