BigClin

Big data analytics for unstructured Clinical data

IRISA/Inria LinkMedia
IRISA/Inria Cidre & Dionysos
INSERM/LTSI Health Big Data
CNRS/STL

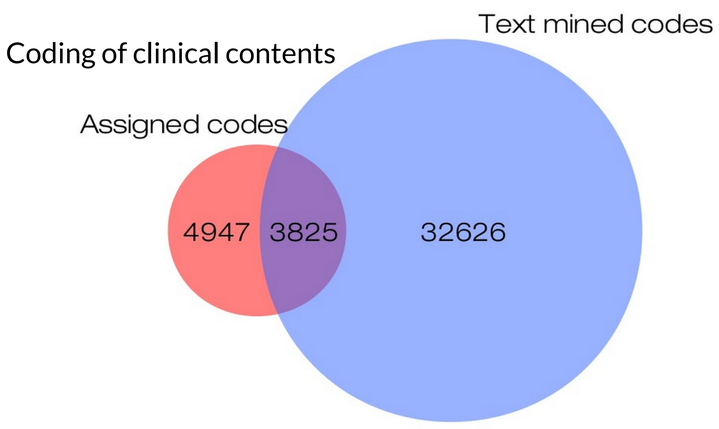
Follow us:

Context and Objectives

Project started fall 2016, related to modern clinical information systems:

- Electronic Health Records and Clinical Data warehouses
- accumulation of large amounts of clinical data
- structured, semi-structured and narrative data
- numerous sources of clinical data

The main objective is to **reuse available clinical data** for clinical and medical research, focusing on the richness of narrative texts.



Objective is to leverage challenges when using narrative data:

- need of robust, fine-grained **text mining and NLP** techniques dedicated to clinical narratives
- issues of **distributed systems** : scalability, management of uncertain data, privacy, stream processing at runtime

Approaches

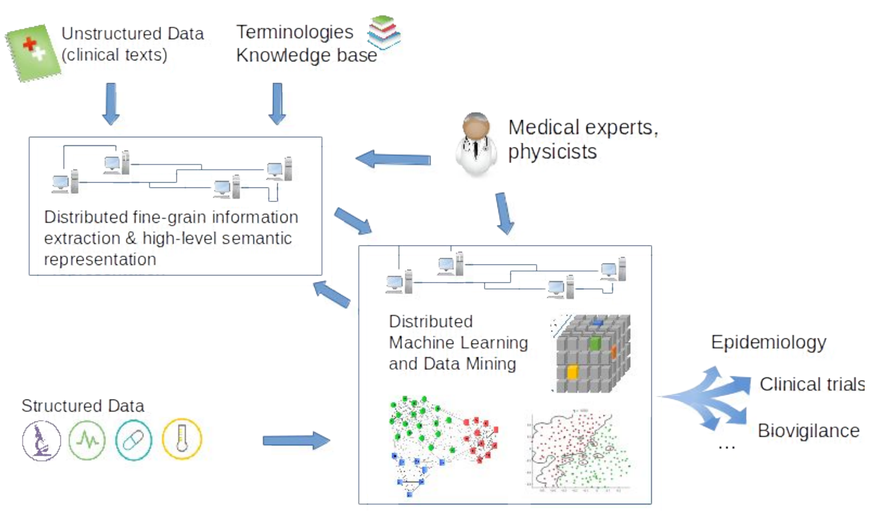
Hybrid system with several complementary approaches:

Information extraction: localize and extract relevant and precise information in narrrative texts, its status (present, absent, uncertain), its temporality, its attribution (the patient, his family)...

High-level semantic representation : represent the extracted information within suitable, semantically rich and processable structure (Word2Vec, Doc2Vec...)

Distributed and runtime computing : quickly process large amounts of data

Reference data building and evaluation : expertise from medical experts



Expected Results

Several applicational contexts addressed for secondary use of clinical data, related to patient security and medical efficiency. Such as:

- **clinical trials**: selection of patient cohorts according to the constraints of a given clinical trial (patient medical history, medications prescribed and taken, main disorder, co-morbidities, age and gender, allergies...);
- **health-care quality**: improve patient security related to medication prescription and intake (co-morbidities, allergies, weight, gender, pregnancy, other medications taken...), on the example of anti-coagulants

Expected benefits

For **health-care and clinical research** : methods for the secondary use of clinical data; to improve quality of care, to perform efficient feasibility studies, to identify targeted cohort and to improve patient recruitment in the perspective of a P4 (Predictive, Preventive, Personalized and Participatory) medicine.

For **teaching and training** : training of young researchers (Master and PhD students); use of proposed methods, resources and tools for training of Master students in various disciplines.

For **computer science**: creation of novel, robust and scalable text-mining analysis methods.

Partnership

Four teams with complementary expertise are involved in the project: (1) medical and health informatics, (2) distributed systems, and (3) NLP and text mining.

Medical and Health Informatics

HBD (Health Big Data) SEPIA-LTSI/INSERM is a medical informatics team closely associated with the Clinical Data Center of the Academic Hospital in Rennes (CHU of Rennes) and developing expertise in semantic interoperability and heterogeneous data integration in the health field. HBD leads or participates in different projects at national and international levels for several applications dedicated to secondary data reuse for clinical research: oncological recruitment support systems, EHR search and visualization, Clinical Data Warehouse infrastructures and networks. The team will bring its expertise in clinical research, semantic data integration, automatic reasoning within heterogeneous clinical data; clinical deidentified data; medical expertise to evaluate, as an end-user, the methods developed in the project.

Key persons: Marc Cuggia, Guillaume Bouzillé, Pascal Van Hille, Emmanuelle Sylvestre, Denis Delamarre

Distributed Systems

CIDRE-Dionysos/IRISA-INRIA is a team working on distributed algorithms and large-scale data stream analysis and processing. The team will bring its expertise and advices in the area of big data technologies and distributed computing. Since the current technologies used in clinical infrastructured are clearly inadequate to answer scalability concerns, the team will also work on data summaries and data streaming analyses, that may significantly enhance response time in real clinical research systems.

Key persons: Yann Busnel, Emmanuelle Anceaume
PhD students hired within the project: Richard Westerlynck, Vasile Cazacu

NLP and text mining

LinkMedia/IRISA-INRIA: team dedicated to the analysis of collections of multimedia documents, and more broadly to Multimedia analytics. LinkMedia researchers bring their expertise in several aspects of the project: semi-supervised machine-learning for NLP, information retrieval with spectral representations, biomedical NLP.

Key persons: Vincent Claveau, Ewa Kijak, Olivier Dameron

UMR 8163 STL (Savoirs, Textes, Langage)/CNRS is a multidisciplinary research team with skills in medical NLP related to several aspects related to biomedical NLP and relevant for the project: uncertainty, information extraction, terminology building and exploitation, detection and extraction of numerical values, temporality..,) and linguistics (semantics).

Key persons: Natalia Grabar, Fayssal Tayalati
PhD student hired within the project: Clément Dalloux