

IASIS: Big Data for Supporting Precision Medicine and Public Health Policy-Making

Anastasia Krithara¹, George Paliouras¹, Anna Triantafillou², Maria-Esther Vidal³, Soren Auer³, Ernestina Menasalvas⁴, Mariano Provencio Pulla⁵, Maria Torrente⁵, Peter Garrard⁶, Alison Evans⁷, Roderic Guigo⁸, Louiqa Raschid⁹

¹ NCSR Demokritos, {akrithara|paliourg}@iit.demokritos.gr,

² Athens Technology Center, a.triantafillou@atc.gr

³ University of Bonn, {vidal|auer}@cs.uni-bonn.de

⁴ Universidad Politecnica de Madrid, ernestina.menasalvas@upm.es

⁵ Hospital Universitario Puerta de Hierro-Majadahonda, mtorrente80@gmail.com

⁶ ST George's Hospital Medical School, pgarrard@sgul.ac.uk

⁷ Alzheimer's Research UK, alison.evans@alzheimersresearchuk.org

⁸ Centre for Genomic Regulation, roderic.guigo@crg.eu

⁹ University of Maryland, louiqa@umiacs.umd.edu

1 IASIS in a Nutshell

IASIS is a 36-month H2020-RIA project that started in April 2017, with the vision of turning clinical and pharmacogenomics big data into actionable knowledge for personalized medicine and decision makers. This will be achieved by integrating and analyzing data from disparate sources, including genomics, electronic health records, and bibliography. The integration and analysis of these heterogeneous sources of information will enable the best decisions to be made, allowing for diagnosis and treatment to be personalized to each individual. IASIS will offer a common representation schema for the heterogeneous big data sources. The IASIS infrastructure will convert clinical notes into usable data, combine them with genomic data, related bibliography, image data and more, and create a global knowledge graph. This will facilitate the use of big data analytics in order to discover useful patterns and associations across different resources. Using semantic integration of big data will give the opportunity for paving the way to better care, as well as for the identification of more effective treatments. Data resources for two different disease categories will be explored: lung cancer and Alzheimer's disease.

2 IASIS Goals and Pilots

As precision medicine is practiced more widely, one of the main challenges is the integration and analysis of clinical data. First, genetic and genomic studies require precise phenotype definitions, but electronic health record (EHR) phenotype data are recorded inconsistently and in a variety of formats. Further, the richest information about a patient condition and progress is captured in unstructured clinical narratives. Additionally, patient movement in and out of a health plan (transience) can result in incomplete or fragmented data. IASIS aims at addressing these challenges and proposes an integrated system that will

collect, combine, and analyse disparate sources of big data. Personal data from electronic health records and genomic datasets will be enriched with biomedical ontologies, bibliographical, and biomedical databases. The tangible outcome of this knowledge fusion will be the IASIS integrated knowledge base. IASIS will provide several enabling technologies, such as a Question Answering system, to help experts during the interaction with the IASIS knowledge base in natural language. In addition, a pattern visualization tool will be developed, in order to give the possibility to the experts to explore and assess the extracted patterns. IASIS will enforce privacy- and trust-aware policies during data management and analytics to address the sensitivity of clinical data.

Pilot 1: Lung cancer is a heterogeneous disease, with characteristics that differ between patients and even among regions of the same tumor. A comprehensive characterization of lung cancer tumor signatures is critical to diagnosis and appropriate treatments. The aim of IASIS is to allow for the integration of anonymized clinical data and biological sample analysis, into the IASIS knowledge graph. Biomedical ontologies and taxonomies such as the Human Phenotype and the NCI Thesaurus, will be used for both terminology standardization and for semantically describing the lung cancer related datasets. Privacy-preserving methods will be used to perform high-level analysis to uncover patterns in the IASIS knowledge graph. These patterns will facilitate more accurate correlations between tumor spread, prognosis, and response to treatment, unravelling molecular mechanisms that predict response depending on the type and characteristics of the tumor. This will contribute to the discovery of personalized and more effective treatments.

Pilot 2: The recognition of dementia as an emerging global health problem of potentially devastating magnitude has come to public attention relatively recently. The Alzheimer's disease symptom heterogeneity presents difficulties for diagnosis as well as for the assessment of response to, and therefore evaluation of, new treatments. IASIS aims at harnessing the potential of big data sources to produce evidence-based clinical knowledge in highly novel and potentially powerful ways. A major scientific opportunity provided by IASIS will be a novel methodology for identifying or confirming associations with dementia types and responses to dementia treatment, as well as dementia prognoses and outcomes. IASIS will allow for identifying signals within noisy systems, but also to strengthen the empirical grounding of medical and lifestyle advice in relation to dementia risk and all stages of dementia severity.

3 Dissemination and Networking

In this phase, IASIS partners are identifying relevant data sources, main data and analytics requirements, as well as devising the IASIS platform architecture. The IASIS project invites health research centers, hospital, and health organizations to both contribute with pharmacogenomics and clinical data, and to profit from the data processing and analytics that the IASIS platform will offer. Contributing to IASIS will allow to generate knowledge that will support precise medicine and more effective treatments for lung cancer and the Alzheimer's disease.