Representation learning of asplenic patient data for disease risk prediction [1]

Teresa Cappuccio[†], Maddalena Casale[‡], Laura Casalino[†‡], Maurizio Giordano[†], Marcella Vacca[†‡], Ilaria Granata[†]

[†]Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy

[‡]Haematology and Oncology Pediatric, Department of Woman, Children and General and Specialist Surgery, University of Campania "Luigi Vanvitelli", Naples, Italy

[†‡]Institute of Genetics and Biophisics "Adriano Buzzati-Traverso", National Research Council, Naples, Italy

teresa.cappuccio@icar.cnr.it, ilaria.granata@icar.cnr.it

In recent years, the storage and transfer of clinical data within electronic health records (EHRs) have represented a significant step forward for research, providing direct and almost instantaneous access to information useful for patient care.

However, despite their potential, extracting generalisable knowledge from these records remains a challenge. The lack of uniformity in the data they contain and the absence of standardisation in their formats hinder the direct use of EHR data for training predictive models of disease-associated risks.

Access to the information contained in EHRs is crucial for identifying issues or risk factors and is essential for developing new therapies. In this context, word embedding algorithms play a crucial role in standardising and analysing clinical data within EHRs, representing medical terms as numerical vectors capable of capturing semantic and syntactic similarities between terms. This approach facilitates the extraction of clinically meaningful patterns, revealing possible hidden relationships between features useful to improve the quality of healthcare.

A concrete example of the application of these techniques is a model of embedding tested on a dataset of clinical records of asplenic patients provided by the Italian Network for Asplenia (INA), a collaborative network of more than 60 Italian hospital centres. This model was used

[1]The work is supported by the European Union - Next Generation EU, Mission 4 Component 2, Investment 1.1. PRIN 2022
ABCare: The Asplenia Biobanking Community: from Analytes to Therapeutic Decision-Making
Macro-sector LS - Life Sciences, ERC Sector LS7 "Prevention, Diagnosis, and Treatment of Human Diseases"
Project Code 2022Y59MHL_LS7_PRIN2022

to identify possible relationships between clinical features, predict potential issues related to the disorder, and identify the most suitable therapies for each patient.

Keywords: *Asplenia, word embedding algorithm, electronic health records (EHRs), clinical data analysis, predictive models.*