Understanding Asplenia-Related Risk through Semantic Embedding

(1) Teresa Cappuccio; (1) Maurizio Giordano; (2) Maddalena Casale; (3) Laura Casalino; (3) Marcella Vacca; (1) Ilaria Granata

(1) Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy
(2) Haematology and Oncology Pediatric, Department of Woman, Children and General and Specialist Surgery, University of Campania "Luigi Vanvitelli", Naples, Italy
(3) Institute of Genetics and Biophisics "Adriano Buzzati-Traverso", National Research Council, Naples, Italy

Correspondence: teresa.cappuccio@icar.cnr.it

Motivation. The spleen plays a fundamental role in the human body, contributing to immune defence through the production of B and T lymphocytes and participating in the removal of damaged or senescent red blood cells. The term "asplenia" refers to a condition characterised by the anatomical absence of the spleen or a severe reduction in its functionality. Asplenia is associated with an increased predisposition to infections and thrombotic complications. In particular, asplenic patients are mostly vulnerable to encapsulated pathogens, requiring targeted prevention and monitoring strategies.
The absence of standardised medical-surgical guidelines for the management of asplenic patients - especially given that asplenia can result from diverse genetic, haematological, immunological, and oncological conditions - significantly hinders advances in patient care. In this context, splenectomy - the surgical removal of the spleen - often remains the first-line intervention, despite its association with serious complications.
This study aims to develop a computational pipeline and algorithm to identify risk factors associated with the most common complications in both surgical and functional asplenia, ultimately serving as a clinical decision support tool.
An in-depth analysis of patients' electronic health records (EHRs), supported by artificial intelligence techniques, can significantly contribute to the early identification of risks associated with this condition, enabling personalised therapeutic strategies and improved long-term clinical management.

Methods. The proposed analysis is based on the use of embedding techniques, aimed at semantically representing clinical events recorded in patients' EHRs. The core idea of such techniques is to transform categorical events such as diagnoses, treatments or symptoms into continuous numerical vectors capable of capturing latent relationships between concepts frequently associated in similar clinical contexts.
This type of representation is particularly valuable as it allows the identification of semantic similarity between different events that may be clinically related, facilitating better generalisation compared to traditional encoding approaches such as One-Hot Encoding or Label Encoding, which treat each event as an independent entity. Moreover, embeddings help to reduce dimensionality while maintaining a more compact and expressive informational structure. This, in turn, enables predictive models to learn meaningful patterns even in the presence of incomplete or noisy data.
Our approach was tested on a dataset comprising approximately 1,800 asplenic patients, whose information was collected and harmonised through the collaboration of more than 60 Italian hospitals members of the INA (Italian Network for Asplenia).

The vectors obtained via embedding were first used to construct an occurrence matrix, enabling the estimation of how frequently each clinical condition (e.g. therapies, surgical procedures) appears in every electronic health record (i.e., per patient). Subsequently, the same vectors were used to build a risk matrix capable of quantifying, for each patient, the strength of the relationship between observed clinical events and the occurrence of specific target conditions such as infectious events. The aggregated values derived from the risk matrix were then used as input to a supervised classifier, with the aim of predicting the probability that a patient might develop the target condition based on their clinical profile.

Results. The preliminary results obtained from the proposed approach suggest that the model is capable of effectively distinguishing between patients who develop the target condition and those who do not. This indicates its potential applicability in predictive analysis settings, where early identification of at-risk patients is crucial.
The combined use of embeddings for the semantic representation of clinical events and risk-based feature construction has yielded encouraging outcomes, further supporting the model's suitability for the automated interpretation of clinical records and the early anticipation of onset or progression of these pathological conditions.

References

Cappuccio T, Casale M, Casalino L et al. Prediction medium and long-term risks in asplenic patients: a precision medicine approach [version 1; not peer reviewed]. F1000Research 2024, 13:1450 (poster) (https://doi.org/10.7490/f1000research.1120036.1)

Jun Wen, Hao Xue, Everett Rush, Vidul A. Panickan, Tianrun Cai, Doudou Zhou, Yuk-Lam Ho, Lauren Costa, Edmon Begoli, Chuan Hong, J. Michael Gaziano, Kelly Cho, Katherine P. Liao, Junwei Lu, Tianxi Cai, DOME: Directional medical embedding vectors from Electronic Health Records, Journal of Biomedical Informatics, Volume 162, 2025, 104768, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2024.104768 (https://www.sciencedirect.com/science/article/pii/S1532046424001862)