EDHER-MED: Early Detection of Health Risks by Textual Analysis of Medical Documents

Juan Martinez-Romo[1,2,*], Lourdes Araujo[1,2], Arantza Casillas Rubio[3] and Aitziber Atutxa[3]

[1]NLP & IR group - Universidad Nacional de Educación a Distancia (UNED). C/ Juan del Rosal, 16, 28040 Madrid, España (http://nlp.uned.es/)

[2]Instituto Mixto UNED-ISCIII (IMIENS), Monforte de Lemos nº 5 – Pabellón 7 – 2ª Pl. 28029, Madrid, España, (https://www.imiens.es/index.php)

[3]HiTZ Basque Center for Language Technologies - Ixa (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia, España (http://www.hitz.eus)

Abstract

The EDHER-MED project focuses on the early detection of health risks through Natural Language Processing (NLP) and Artificial Intelligence (AI). Led by two research groups – Natural Language Processing and Information Retrieval (NLP&IR) group at National University of Distance Education and Hizkuntza Teknologiako Euskal Zentroa (HiTZ) group at University of the Basque Country – this initiative develops advanced computational models to extract medical insights from Electronic Health Records (EHRs), scientific literature, and patient narratives. The project introduces novel methodologies, including biomedical domain-specific language models, clinical argument mining, and temporal event detection to improve risk assessment in mental health, HIV, rare diseases, and cardiovascular conditions. Its main objectives include developing NLP tools, medical ontologies, and predictive models for early disease detection. The anticipated scientific, social, and economic impact includes enhanced clinical decision support, reduced healthcare costs, and improved patient outcomes, positioning EDHER-MED as a transformative AI-driven solution in healthcare research.

Keywords

Early Disease Detection, Clinical Decision Support, Medical Text Mining, Predictive Health Analytics

1. Scientific Proposal

The EDHER-MED project is a project funded by the Ministry of Science and Innovation in the 2022 call for R+D+i projects, within the State Program for Research, Development and Innovation Oriented to the Challenges of Society. EDHER-MED is a coordinated project between the Natural Language Processing and Information Retrieval (NLP&IR) group at National University of Distance Education (UNED) and Hizkuntza Teknologiako Euskal Zentroa (HiTZ) group at University of the Basque Country (UPV-EHU). In this new project, the research is related to a set of use cases aiming to discover new knowledge on health risks and supporting the early diagnoses of some illnesses and disorders, mainly by the use of Electronic Health Records (EHRs), and other sources (such as scientific literature, texts written by subjects patients).

The early detection (ED) of health risks is a rapidly evolving field in healthcare research, aimed at identifying the earliest signs or symptoms of diseases before they become severe. Traditional biomedical informatics has focused on diagnosing patients based on symptoms, biological markers, and comorbidities. However, with the digitalization of medical records, personalized medicine has advanced, allowing for a broader analysis of multiple pathologies.

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025

*Corresponding author.

†These authors contributed equally.

juaner@lsi.uned.es (J. Martinez-Romo); lurdes@lsi.uned.es (L. Araujo); arantza.casillas@ehu.eus (A. C. Rubio); aitziber.atucha@ehu.eus (A. Atutxa)

0000-0002-6905-7051 (J. Martinez-Romo); 0000-0002-7657-4794 (L. Araujo); 0000-0003-2638-9598 (A. C. Rubio); 0000-0001-9097-6047 (A. Atutxa)

The EDHER-MED project leverages state-of-the-art (SOTA) Natural Language Processing (NLP) techniques to analyze patient medical histories, identifying specific indicators that can alert healthcare professionals about a patient's risk of developing a disease or complication. ED is not designed to replace physicians but to support their clinical decision-making by providing alerts based on risk indicators. Physicians can choose to investigate or disregard these alerts based on their expertise. When combined with timely interventions, ED can help mitigate complications, improve treatment efficacy, and prevent long-term disease progression. This leads to better patient outcomes and a reduced burden on healthcare systems by decreasing the need for intensive treatments and long-term management.

Most patients' health data is stored within Electronic Health Records (EHRs) in natural language making Natural Language Processing techniques crucial for the acquisition and extraction of relevant information. This project intends to make progress in the application and development of NLP techniques aiming to enhance the automatic processing of clinical reports and thereby improve the early detection of health risks in different scenarios. Note that the NLP technology implemented will be generalizable to all the scenarios faced in the project and by extension to many others contributing to build more and better systems and help in future use cases. The specific scenarios comprise diseases with a high social impact, specifically in Mental health, Human immunodeficiency virus, Rare diseases and Cardiovascular diseases.

The analysis of mental health in the youth population is the use case where both groups will be involved through two types of actions. On the one hand, the developed tools will detect high risk suicidal behaviors in children and adolescents by the automatic extraction of information from structured (questionary answers) and unstructured information (free text) to identify evidence of these situations according to the established psychiatric ideation-to-action framework [1]. On the other hand, we will deepen in the early detection of mental health risks in youth population by the stratification of young people (16-25 years) with mental health problems into subgroups or strata in terms of prognosis or response to treatment in emotional state, depression and suicide.

The early detection of Human Immunodeficiency Virus (HIV) will also be addressed by exploring different techniques to extract key indicators of HIV status. Both statistical techniques, machine learning and deep learning techniques will be explored. On one hand, we will analyze and classify as indicators of HIV different symptoms and diseases given by physicians. This classification will be completed by extracting relations from medical ontologies and graphs, to measure the relevance of indicators, which will be compared with HIV indicators from clinical text data. We will extract explicit and implicit indicators that can be combined with analytical laboratory values to early detect HIV.

Research will also be carried out to improve the characterization of rare diseases and their effects on the mental health and well-being of the child population. This is a little explored aspect in the field of diseases, which is already a field with great limitations in the available information [2]. The application of different techniques will be explored, including statistical methods and the most recent transformer techniques to identify clinical and mental signs relevant to the study, and their relationship with the disease under consideration, always considering the interpretability of the results.

To finish, the last use case covered in this project corresponds to automatically discovering potential risk factors associated with cardiovascular complications (for example, ischemic stroke, cardiomyopathy, or recurrence of Atrial Fibrillation (AF)) after a first episode. To that end we will use patients' Electronic Health Records (EHR), progress notes and electrocardiograms (ECG) and confront these new knowledge to the existing scores and guidelines used in hospitals to ameliorate them. In addition, by applying explainability algorithms, we aim at finding relevant information to detect AF signs at early stages.

2. Research Groups

The coordinated EDHER-MED project consists of two subprojects, each led by a different research group:

- ENIGMA (Early detectioN of hIGh iMpact diseAses through natural language processing). UNED subproject.

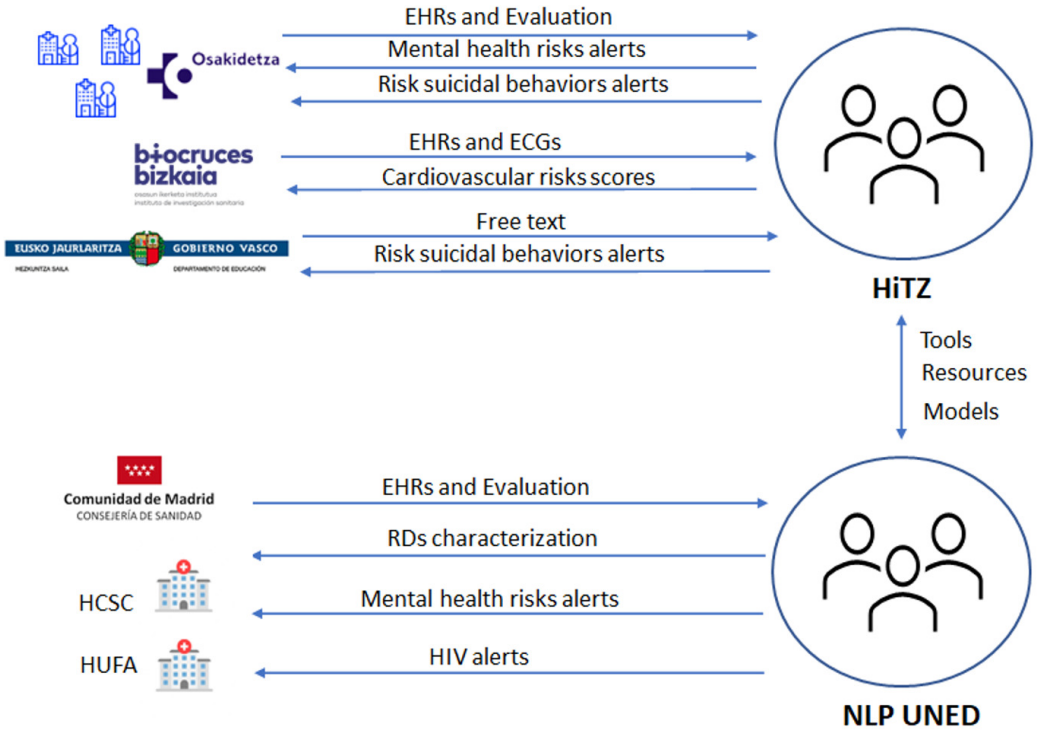

Figure 1: Collaboration Scheme among the subprojects led by the HiTZ Research Center at Univ. of the Basque Country UPV/EHU and the NLP Group at UNED. Entities involved: Osakidetza (Basque Health System), HCSC (Hospital Clínico San Carlos), HUFA (Hospital Universitario Fundación Alcorcón), Biocruces (Bizkaia Institute of Health Research), Health Department of the Community of Madrid, and Basque Government Department of Education (CSCM).

- EDHIA (Early Detection and Health rIsk identification with NLP and Argumentation). UPV-EHU subproject.

The project, which is multidisciplinary in nature, has the collaboration of two NLP research groups and several health-related institutions, with whom the use cases of each subproject will be developed. Figure 1 shows a scheme of the collaboration among the different participants.

Both groups will contribute with their experience, resources, and knowledge of NLP, and in its application to the medical field. The UNED team has proven experience in massive data crawling in both scientific publications and tweets, anonymization and acronym disambiguation on medical texts, and provides experience working with psychological disorders. The HiTZ team has large experience on annotation, and development of tools for topic extraction, named entity recognition, and other relation extraction tasks in medical texts, building multilingual and multimodal Large (LLM) and discriminative Language Models (LM), clinical narrative section/topic modeling and working in low data regimes. Both teams have experience with International Classification of Diseases (ICD) classification, Named Entity Recognition using different approximations, negation detection and the adaptation of language models for information extraction with transformers and deep learning in the clinical domain. The collaboration is therefore synergistic, as the teams will share their expertise and knowledge of different techniques, and methods to solve the medical domain NLP tasks required in the project.

3. General Objectives

The main objectives of the project will be presented below. Figure 2 shows the main elements of the project and their interactions.

1. Basic NLP resources necessary to achieve the rest of the general objectives.

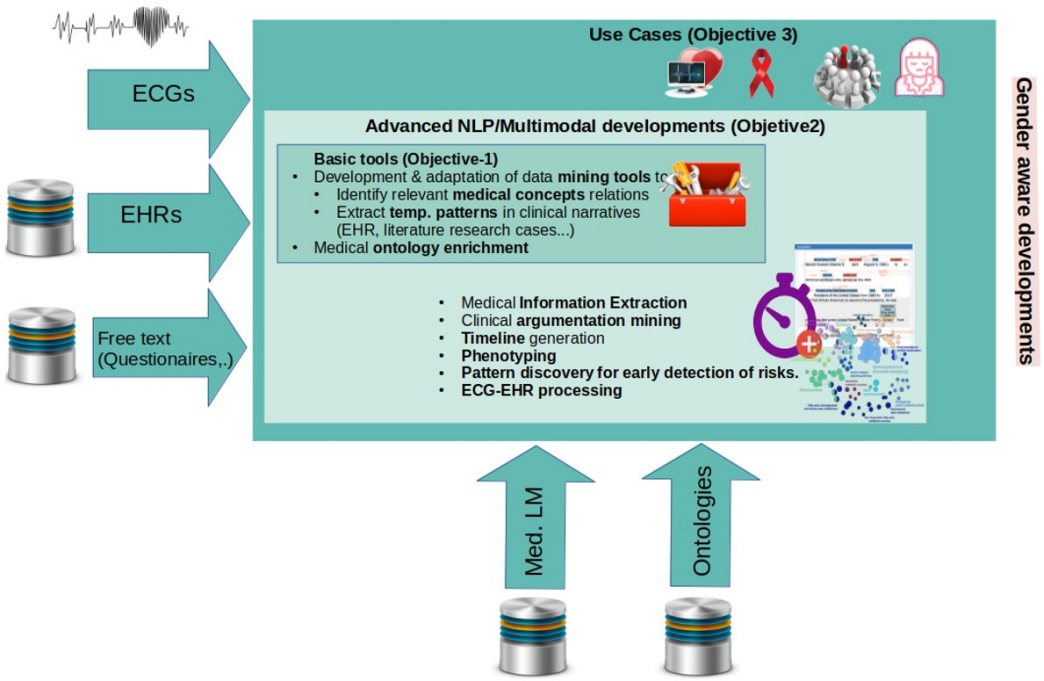

Figure 2: General architecture of the project.

- Development and adaptation of data mining tools to identify medical concepts and relevant relations between them, temporal pattern extraction in patient EHRs, creation and annotation of different corpora and development of medical LLMs.
- Medical ontology enrichment.

2. Development of NLP/multimodal technology to support early detection of health risks.
   - Medical Information Extraction (timeline generation, diagnoses encoded as ICD, phenotyping, ECG-EHR joint processing).
   - Pattern discovery for early detection of risks.
   - Clinical Argument Mining.
3. Use of the tools and technologies developed in the previous objectives for their application to specific use cases focused on health problems with high social impact.
   - Early detection of mental health risks.
   - Early detection of HIV.
   - Early detection of mental health and social problems in children affected by rare diseases.
   - Potential risk factors identification to prevent cardiovascular complications.

Given that our research project involves the use of electronic health records (EHRs) and is conducted in collaboration with various healthcare institutions, adherence to ethical and legal standards concerning patient data has been a fundamental priority from the outset. In compliance with national and institutional regulations governing the use of health-related data for research purposes, the project has been submitted for ethical review and approved by the corresponding ethics committees of each participating healthcare institution. These evaluations were essential not only to authorize access to the EHRs but also to ensure that the procedures for data handling align with established norms of biomedical research ethics. Importantly, the data obtained for the study were subject to anonymization protocols implemented at the institutional level prior to their provision to the research team. This approach ensures that all personal identifiers are removed at the source, thereby safeguarding patient confidentiality. Furthermore, the research team has implemented additional data protection measures throughout the project lifecycle, including secure data storage, restricted access, and compliance with

applicable data privacy regulations. These combined efforts guarantee that the privacy and integrity of patient information are rigorously protected, in accordance with both ethical obligations and legal requirements.

4. Expected impact

The extraordinary advances in AI and NLP that have revolutionized information processing technology need to be adapted and evaluated in the different applications and types of data in the medical domain. In particular we will explore improvements related to the following points:

- Mental health disorders and NLP are mainly joined by two scientific-technical perspectives: computational methods for the detection of people with mental health risk and the generation of specific data sources that feed the computational methods. Our main contribution will be the generation of annotated Spanish corpora and the development of a model for the detection of young population with risk suicidal behaviors.
- Progress in the knowledge of rare diseases is essential to improve their prognosis. Often the scarce information available on a Rare Disease (RD) is scattered in different databases, with different formats, in many cases unstructured. Moreover, even the information on the same patient is scattered, since the diagnosis may involve multiple clinical centers, specialities and investigations. We expect to improve the knowledge about RDs and their relationships with mental health and social factors, especially in the case of the child population.
- Our main contribution regarding cardiovascular diseases will be to advance in improving existing scores and clinical guidelines. To do so information derived from EHRs and ECGs will be annotated and employed. Advancing in the explainability of model's decisions which is key in medical decission making.
- With respect to HIV, we expect that the application of the basic NLP tools and the more advanced tools developed in the project will allow better analysis of the unstructured text of clinical reports and notes in a way that will improve early diagnosis.

5. Progress achieved

Since the start of the project, we have made progress on all objectives.

- *NLP resources:* We have generated both data collections and fundamental tools for the development of the project. Among them we have compiled an initial corpus in Spanish for the detection of suicide attempts [3]. We have also developed tools for the detection of fake news in medicine [4], which allow us to filter information after information retrieval processes. Another important line for the generation of timelines to analyze the evolution of health problems has been the development of advanced models for the identification and normalization of dates in medical texts [5]. Regarding critical general resources, we have built domain specific medical LMs and LLM [6].
- *Advances in early detection of health risks:* We have also made progress in the early detection of risks in different areas, such as mental health [3], HIV screening [7], ICD predicition [8], and early detection of cardiovascular diseases [9].
- *Advances in medical argumentation:* We have established a methodology and built a bechmark to evaluate LLM's generated medical arguments [10, 11]. We have also developed tools for finding the argument supporting a correct medical hypothesis [12]. And finally we have explored cross-lingual Transfer and few shot techniques for Argument Mining [13].

We have also participated in several related competitions, and our proposals have achieved relevant rankings [14, 15, 16, 17, 18].

6. Acknowledgements

EDHER-MED, with subprojects ENIGMA (PID2022-136522OB-C21) and EDHIA (PID2022-136522OB-C22) are funded by the Spanish Ministry of Science and Innovation. In addition, this work has been partially supported by the Spanish Ministry of Science and Innovation within the OBSER-MENH Project (MCIN/AEI/10.13039 and NextGenerationEU"/PRTR) under Grant TED2021-130398B-C21.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] L. T. Bayliss, S. Christensen, A. Lamont-Mills, C. du Plessis, Suicide capability within the ideation-to-action framework: A systematic scoping review, PloS one 17 (2022). doi:10.1371/journal.pone.0276070.

[2] J. Liu, J. S. Barrett, E. T. Leonardi, L. Lee, S. Roychoudhury, Y. Chen, P. Trifillis, Natural history and real-world data in rare diseases: applications, limitations, and future perspectives, The Journal of Clinical Pharmacology 62 (2022) S38–S55.

[3] J. Fernandez-Hernandez, L. Araujo, J. Martinez-Romo, Generation of social network user profiles and their relationship with suicidal behaviour, Procesamiento del lenguaje natural 72 (2024) 87–98.

[4] J. R. Martinez-Rico, L. Araujo, J. Martinez-Romo, Building a framework for fake news detection in the health domain, Plos one 19 (2024). doi:10.1371/journal.pone.0305362.

[5] A. Sánchez de Castro, L. Araujo, J. Martinez-Romo, Generative LLMs for multilingual temporal expression normalization, in: ECAI 2024, IOS Press, 2024, pp. 3789–3796.

[6] I. García-Ferrero, R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, A. Zaninello, MedMT5: An open-source multilingual text-to-text LLM for the medical domain, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11165–11177. URL: https://aclanthology.org/2024.lrec-main.974/.

[7] R. Morales-Sánchez, S. Montalvo, A. Riaño, R. Martínez, M. Velasco, Early diagnosis of HIV cases by means of text mining and machine learning models on clinical notes, Computers in Biology and Medicine 179 (2024) 108830.

[8] N. Lebeña, A. Pérez, A. Casillas, Quantifying decision support level of explainable automatic classification of diagnoses in spanish medical records, Computers in Biology and Medicine 182 (2024) 109127.

[9] A. G. Olea, A. G. Domingo-Aldama, M. M. Prado, K. G. Galletebeitia, A. A. Salazar, M. M. Rada, I. G. Díaz, A. Costa, I. Cano, F. Díaz, et al., Rendimiento de las expresiones regulares en el análisis de informes de alta presentes en la historia clínica electrónica exprimiendo los datos secundarios, Revista Española de Cardiología 77 (2024) 33–34.

[10] I. De la Iglesia, I. Goenaga, J. Ramirez-Romero, J. M. Villa-Gonzalez, J. Goikoetxea, A. Barrena, Ranking over scoring: Towards reliable and robust automated evaluation of LLM-Generated medical explanatory arguments, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 9456–9471.

[11] I. Alonso, M. Oronoz, R. Agerri, MedExpQA: Multilingual benchmarking of large language models for medical question answering, Artificial intelligence in medicine 155 (2024) 102938.

[12] I. Goenaga, A. Atutxa, K. Gojenola, M. Oronoz, R. Agerri, Explanatory argument extraction of correct answers in resident medical exams, Artificial Intelligence in Medicine 157 (2024) 102985.

[13] A. Yeginbergen, M. Oronoz, R. Agerri, Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques, arXiv preprint arXiv:2407.03748 (2024).

[14] X. Larrayoz, N. Lebeña, A. Casillas, A. Pérez, Eating disorders detection by means of deep learning., in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.
[15] J. Martinez-Romo, J. Huesca-Barril, L. Araujo, E. d. L. C. Marin, UNED-UNIOVI at EmoSPeech-IberLEF2024: Emotion identification in spanish by combining multimodal textual analysis and machine learning methods, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), 2024.
[16] X. Larrayoz, A. Casillas, M. Oronoz, A. Pérez, Mental disorder detection in spanish: hands on skewed class distribution to leverage training, in: IberLEF (Working Notes). CEUR Workshop Proceedings, 2024.
[17] M. Sierra-Callau, M. Á. Rodríguez-García, S. Montalvo-Herranz, R. Martínez-Unanue, UNED_MRES Team at MentalRiskES2024: Exploring hybrid approaches to detect mental disorder risks in social media, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), 2024.
[18] H. Fabregat, D. Deniz, A. Duque, L. Araujo, J. Martinez-Romo, NLP-UNED at eRisk 2024: approximate nearest neighbors with encoding refinement for early detecting signs of anorexia, in: Working Notes of CLEF, 2024, pp. 9–12.