Early Detection of Rare Diseases using Natural Language Processing

Eglė Kondrataitė, Gražina Korvel

Vilnius University, Institute of Data Science and Digital Technologies, Akademijos str. 4 Vilnius
egle.kondrataite@mif.stud.vu.lt, grazina.korvel@mif.vu.lt

Early and accurate detection of rare diseases is an important aspect of reducing disease progression and improving the quality of life of the affected people. It is estimated that there are about 36 million people in the EU who suffer from more than 5000 rare diseases [4]. The majority of these conditions are of genetic origin and usually appear in childhood. They can often lead to disability, chronic illness, or even premature death [5]. It is difficult to accurately identify a rare disease because the symptoms can be similar to those of more widespread illnesses. Patients that are affected by such conditions constantly face delayed diagnosis, which can lead to psychological and economic challenges for them and their families [6]. Late diagnosis is associated with reduced quality of life and increased mortality rates. Therefore, early diagnosis help doctors to closely monitor the progression of the disease and avoid rapid negative changes in the patient's health.

Early detection of health risks is one of the key foundations of modern healthcare. It allows doctors to carry out necessary tests and prescribe early treatment to control the disease. However, the task of accurately and quickly diagnosing rare diseases is very challenging for general practitioners who may lack knowledge of these conditions [2]. In addition, rare diseases are under-represented in the International Classification of Diseases, version 10 (ICD-10), which is widely used for disease identification [1]. Therefore, there is a great need for new technologies to identify rare diseases.

Natural language processing (NLP) methods are rapidly gaining popularity in the medical field as electronic health records (EHRs) are increasingly implemented. These records are a rich source of data consisting of structured and unstructured information. The structured data includes the patient's medical history, diagnoses, medications, medical and surgical procedures, and allergies. The unstructured data consists of physicians' free-text notes that can include important observations about patient's

DOI: https://doi.org/10.15388/LMITT.2024.9

health. According to [3] NLP methods can be very useful for processing unstructured data in patient health records. Specifically, they have been used to uncover meaningful insights about Dravet syndrome from narrative medical reports in electronic health records [3]. In addition, NLP methods have also been used to analyze free-text clinical notes to detect depression in patients diagnosed with breast and colorectal cancer [2]. However, the application of such methods faces certain challenges. These can include data quality, medical terms in different languages, or the complexity of medical terminology in general. Some conditions may have different synonyms and abbreviations to describe them. For example, "obsessive-compulsive disorder", "anancastic neurosis", and "OCD" are the same disease. In addition, symptoms can present similar challenges where patients' complaints can be described by medical terms and also by short phrases [2]. However, the main difficulty is data annotation. In order to apply machine learning (ML) algorithms, the data must be labelled. However, annotating rare diseases in clinical notes requires expertise in specific fields, hence significant cost and time from clinical experts [1]. Another scientific challenge is that most rare diseases have a limited number of cases and may present with unusual symptoms. Therefore, the development of NLP models that can handle unique linguistic features and terminology of rare diseases is necessary.

Keywords: Early Detection, Rare Diseases, Natural Language Processing, Electronic Health Records

References

[1] Hang Dong, Victor Suarez-Paniagua, Huayu Zhang, Minhong Wang, Emma Whitfield, and Honghan Wu. Rare disease identification from clinical notes with ontologies and weak supervision. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 2294–2298. IEEE, 2021.

[2] Angela Leis, David Casadevall, Joan Albanell, Margarita Posso, Francesc Macia, Xavier Castells, Juan Manuel Ramirez-Anguita, Jordi Martinez Roldan, Laura I. Furlong, Ferran Sanz, et al. Exploring the association of cancer and depression in electronic health records: Combining encoded diagnosis and mining free-text clinical notes. JMIR Cancer, 8(3): e39003, 2022.

[3] Tommaso Lo Barco, Nicolas Garcelon, Antoine Neuraz, and Rima Nabbout. Natural history of rare diseases using natural language processing of narrative unstructured electronic health records: the example of dravet syndrome. Epilepsia, 65(2):350–361, 2024.

[4] Julio Lopez-Bastida, Juan Oliva-Moreno, Renata Linertova, and Pedro Serrano-Aguilar. Social/economic costs and health-related quality of life in patients with rare diseases in europe. The European Journal of Health Economics, 17(Suppl 1):1–5, 2016.

[5] Shruti Marwaha, Joshua W. Knowles, and Euan A. Ashley. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome Medicine, 14(1):23, 2022.
[6] Yvonne Zurynski, Marie Deverell, Troy Dalkeith, Sandra Johnson, John Christodoulou, Helen Leonard, Elizabeth J. Elliott, and APSU Rare Diseases Impacts on Families Study group. Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. Orphanet Journal of Rare Diseases, 12:1–9, 2017.