LITERATURE REVIEW

Paper 1: <https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0040-1702001>

# **Medical Information Extraction in the Age of Deep Learning**

Udo Hahn *,* Michel Oleynik

Objectives: This review aims to examine recent advancements in medical Information Extraction (IE) as documented in the literature from the last three years. Our primary focus is on the significant shift in methodology from traditional Machine Learning (ML) techniques to the use of Deep Neural Networks (DNNs). We will explore the applications of this new approach, with a specific emphasis on two fundamental IE tasks: named entity recognition and relation extraction, focusing on diseases and drugs (or medications) and their associations.

Methods: We conducted a thorough search for relevant publications published between 2017 and early 2020 within three major scientific domains: medicine and medical informatics, natural language processing, and neural networks and artificial intelligence.

Results: Over the past decade, the field of Natural Language Processing (NLP) has undergone a substantial transformation in methodology, transitioning from symbolic representations to distributed ones based on Deep Learning (DL) principles. This shift has also been gradually adopted within the medical NLP community, as evidenced in this survey of medical IE. The review showcases compelling experimental evidence that DL-based approaches consistently outperform non-DL methods by a considerable margin. However, the challenges of working with small and restricted datasets and the unique linguistic characteristics of medical sublanguages require adaptive learning strategies to address these inherent issues.

Conclusions: The shift from (feature-engineered) ML to DNNs brings about a fundamental change in the methodological framework for medical NLP. This transition is not limited to medical IE but is expected to have a profound impact on various areas of medical informatics, both NLP and non-NLP related.

Is this conversation helpful so far?

Paper 2: <https://www.mdpi.com/2077-0383/11/11/2967>

Song, G.; Chung, S.J.; Seo, J.Y.; Yang, S.Y.; Jin, E.H.; Chung, G.E.; Shim, S.R.; Sa, S.; Hong, M.S.; Kim, K.H.; et al. Natural Language Processing for Information Extraction of Gastric Diseases and Its Application in Large-Scale Clinical Research. *J. Clin. Med.* **2022**, *11*, 2967. <https://doi.org/10.3390/jcm11112967>

A natural language processing (NLP) pipeline was created to extract clinical data from unstructured esophagogastroduodenoscopy (EGD) reports, focusing on 10 gastric diseases. Validation with 1000 reports showed high accuracy, with sensitivity, positive predictive value (PPV), and F1 score exceeding 0.96 for gastritis and 0.97 for ulcers and neoplastic diseases. Applying the pipeline to 248,966 reports spanning a decade revealed patient demographics, disease extent, and locations. This study demonstrates the NLP pipeline's potential for automated extraction of gastric disease information from EGD reports, enabling large-scale clinical research to enhance our understanding of gastric diseases.

Paper 3:

Title: Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML.

The abstract emphasises the significance of precise prognostic staging in cancer prediction and therapy decisions. It emphasises the difficulty of unstructured medical records as well as the requirement for a standardised clinical decision stage technique. The literature supports the need for such a procedure, particularly in cases of breast cancer. This work used data from 465 patients in India to successfully extract critical prognostic features from various medical records using natural language processing, machine learning, and rule-based approaches. The study predicts prognostic phases with excellent accuracy in both rural and urban settings, demonstrating the possibility

for improved breast cancer prediction. A generic staging system like this can considerably improve patient care and treatment decisions.

Title: Clinical information extraction applications: A literature review

Because of the increased integration of electronic health records (EHRs), opportunities for automated healthcare systems and clinical research have emerged. Information Extraction (IE), which pulls clinical data from textual records, is a critical component enabling secondary EHR utilisation. This review of the literature looks at recent research on clinical IE applications. A thorough search yielded 263 articles published between January 2009 and September 2016. These publications are evaluated based on their publication sources, data origins, clinical IE tools, methodology, and applications in disease, drug research, and workflow advancements. Despite the numerous uses of clinical IE, there is a significant gap between EHR-based clinical studies and clinical IE research, showing that there is opportunity for development in bridging this gap and developing healthcare informatics.

Title: Information Extraction of Medical Materials: An Overview of the Track of Medical Materials MedOCR

This conference paper examines the critical function of Information Extraction (IE) in the medical and insurance industries when dealing with electronic medical record contents. It emphasises the ability of artificial intelligence to extract and use important information from these records, lowering labour costs and increasing efficiency. Currently, manual data extraction is the most used way. The report, however, highlights the growing interest in using Optical Character Recognition (OCR) and Natural Language Processing (NLP) technology to automate the procedure. To that end, the authors present the Medical OCR dataset (MedOCR) and host an evaluation competition as part of the eighth China Health Information Processing Conference (CHIP2022). The competition drew 18 teams, with OCR-based systems producing excellent results with a concentration on Acc as the assessment parameter. The report emphasises the importance of information extraction in the medical sector and serves as a useful resource for future research and development in this field.

Title: Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach

This work addresses the difficult job of obtaining critical information from infectious illness cases in medical literature, which is critical for public health research. The study uses natural language processing (NLP) to mine clinical and social determinant data from published cases. The proposed framework combines data preparation, natural language processing for named entity recognition, and evaluation components, with an emphasis on COVID-19 case reports. When compared to benchmark approaches, the results show enhanced performance in named entity recognition and relation extraction. The work emphasises the utility of using transfer learning for future research in this domain and the potential for adapting this method to other infectious diseases.

Title: Using natural language processing to automatically classify written self-reported narratives by patients with migraine or cluster headache.

This work investigates the use of natural language processing (NLP) to self-reported narratives of headache problem patients, with the goal of automatically classifying and extracting relevant information from clinical descriptions. The study identifies different word choices in narratives of migraine and cluster headache patients using lexical, semantic, and theme analysis, which aligns with expert knowledge of these conditions. In identifying headache attack descriptions, machine learning (ML) approaches such as logistic regression and support vector machines perform well. The work emphasises NLP's ability to detect key language elements in clinical narratives and the promise for future breakthroughs using larger datasets and neural NLP methods.

Title: Precision information extraction for rare disease epidemiology at scale.

The recent call by the United Nations to address the issues faced by the 300 million people globally who live with rare diseases emphasises the crucial need for comprehensive epidemiological data. Existing methods for finding, extracting, and curating epidemiologic information (EI) for rare diseases are time-consuming and error-prone, limiting our understanding of these conditions. In response, this paper provides a fresh approach: the creation of EpiPipeline4RD, a deep learning-based pipeline for the extraction of rare disease epidemiology information. The study produces good precision, recall, and F1 scores using a curated corpus for Named Entity Recognition, illustrating the efficacy of an automated curation paradigm. The EpiPipeline4RD project has the potential to greatly increase public health support and research in rare illnesses.