# Electronic Health Records with Unstructured Text to Predict Outcomes With Machine Learning: A Therapeutic Area Fingerprint


UNIVERSITAT DE BARCELONA

AUTHORS

*Manuel Cossio1,2 & Ramiro Gilardino1*

HE-XPERTS CONSULTING

AFFILIATIONS

1. *HE-Xperts Consulting LLC, Miami, FL, USA.*
2. *Universitat de Barcelona, Barcelona, Spain*



*Keywords: Electronic Health Records, Machine Learning, Health Outcomes, Digital Healthcare*


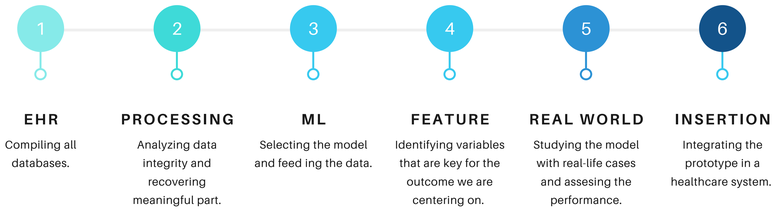


***Figure 1. Life cycle scheme of ML prototypes for EHR automatic analysis.*** *Globes indicate succesive steps.*

## *01. Introduction*

Electronic health records (EHR) are digital repositories that contain information about patients' medical history including symptoms, clinical examination findings, test results, procedures and prescriptions (1). All the stored variables are usually related to an outcome of each patient (Figure 1). These outcomes can be used as labels to train machine learning (ML) algorithms and thus build automatic classifiers for particular clinical conditions (2). Clinical information can be contained in variables with a range of values or in the form of clinical notes (free text). The clinical notes need to be processed with natural language processing (NLP) techniques to extract data points to constitute imputs for ML algorithms (1, 2).

## *02. Objective*

The aim of this work is to analyze the state of ML research applied to EHR focused the frequency of use of structured and unstructured data (free text) in the form of clinical notes and its processing with NLP.

## *References*


1. Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PloS one, 13(8), e0202344.
2. Lv, H., Yang, X., Wang, B., Wang, S., Du, X., Tan, Q., ... & Xia, Y. (2021). machine learning–driven models to predict prognostic outcomes in patients hospitalized with heart failure using electronic health records: retrospective study. Journal of medical Internet research, 23(4), e24996.


## *03. Methods*

Our systematic review was conducted searching for articles in English from inception and up to September 8, 2021. The databases analyzed were Scopus and Google Scholar and they were screened for titles, abstracts and keywords containing the words 'machine learning' AND 'electronic health records'. The search for articles was circumscribed to only full-text articles. After the identification of the articles in each database, both pools were unified, removing the duplicates.

## *04. Results*

Of the studies analyzed (n = 117), they belonged to the following medical specialties (Fig. 1) in order of frequency: cardiovascular (n = 27), psychiatry (n = 19), oncology (n = 14), diabetes (n = 13), neurology (n = 9), infectology (n = 8), nephrology (n = 4), rheumatology and gastroenterology (each n = 3), emergentology, hepatology, gynecology, metabolism and ophthalmology (each n = 2), and finally surgery, pneumonology, traumatology and dermatology (each n = 1).

Cardiovascular, psyquiatry and oncology presented the highest proportions of EHR with structured and unstructured data. It can also be seen that psychiatry presented the highest proportion of not specified data type and infectology presented the dominant proportion of structured data only (Fig. 2).

With respect to NLP techniques (Fig. 3), the first most frequent was cTakes (cardiovascular, dermatology and psychiatry) and the second most frequent was MetaMap (cardiovascular, neurology and psychiatry).

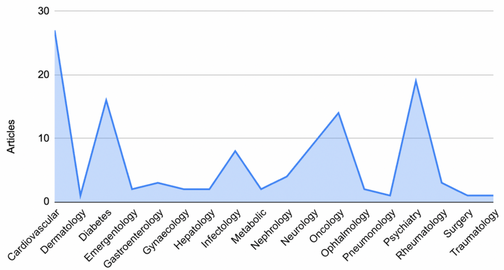


***Figure 2. Number of articles of ML and EHR cataloged in this research work.*** *Peaks represent maximum valius per specialty and valleys represent minimum values per specialty.*

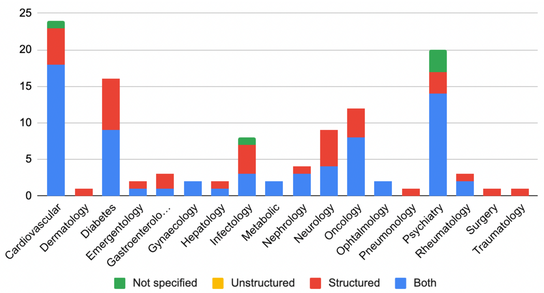


***Figure 3. Proportion of structured and unstructured data in EHR per specialty.*** *'Not specified' reflects absence of details in each respective article and 'both' represents cases with unstructured and structured data in the same database.*

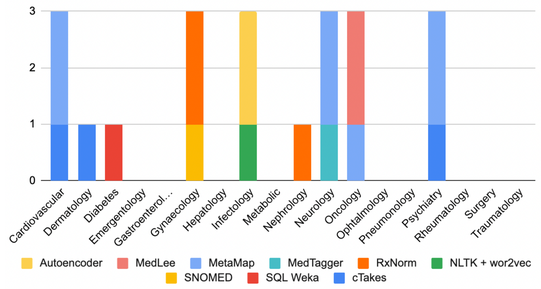


***Figure 4. Frequency of use of different NLP techniques per specialty.*** *Bars extending [0,1] represent first most frequent technique and bars extending [1,3] represent second most frequent technique.*

## *05. Conclusion*

After a meticulous analysis of the data, it can be concluded that the different medical specialties have different proportion of structured and unstructured data and accordingly, employ different techniques of NLP. It was also seen that certain techniques are regularly adopted in several specialties to process unstructured data.

Finally, gaps of oportunity to continue the research in ML and NLP can be visualized in each specialty. This finding could be key to promote the automatization of EHR analysis to quantify the value of interventions and determine burden of disease.