Clinical characteristics and prognostic factors for Crohn’s disease relapses using natural language processing and machine learning – a pilot study

F. Gomollón García; J. Pérez Gisbert; I. Guerra Marina; R. Plaza Santos; R. Pajares Villarroya; L. Moreno Almazán; M.C. López Martín; M. Domínguez Antonaya; M.I. Vera Mendoza; J. Aparicio; V. Martínez; I. Tagarro García; A. Fernández Nistal; S. Lumbreras Sancho; C. Maté Ruiz; C. Montoto

Abstract-

**Background **
The impact of relapses on disease burden in Crohn’s disease (CD) warrants searching for predictive factors to anticipate relapses. This requires analysis of large datasets, including elusive free-text annotations from electronic health records. This study aims to describe clinical characteristics and treatment with biologics of CD patients and generate a data-driven predictive model for relapse using natural language processing (NLP) and machine learning (ML).
**Methods **
We performed a multicenter, retrospective study using a previously validated corpus of CD patient data from eight hospitals of the Spanish National Healthcare Network from 1 January 2014 to 31 December 2018 using NLP. Predictive models were created with ML algorithms, namely, logistic regression, decision trees, and random forests.
**Results **
CD phenotype, analyzed in 5938 CD patients, was predominantly inflammatory, and tobacco smoking appeared as a risk factor, confirming previous clinical studies. We also documented treatments, treatment switches, and time to discontinuation in biologics-treated CD patients. We found correlations between CD and patient family history of gastrointestinal neoplasms. Our predictive model ranked 25 000 variables for their potential as risk factors for CD relapse. Of highest relative importance were past relapses and patients’ age, as well as leukocyte, hemoglobin, and fibrinogen levels.
**Conclusion **
Through NLP, we identified variables such as smoking as a risk factor and described treatment patterns with biologics in CD patients. CD relapse prediction highlighted the importance of patients’ age and some biochemistry values, though it proved highly challenging and merits the assessment of risk factors for relapse in a clinical setting.

Index Terms- artificial intelligence, big data, electronic health records, inflammatory bowel disease, natural language processing

Citation:
Gomollón, F.; P. Gisbert, J.; Guerra, I.; Plaza, R.; Pajares Villarroya, R.; Moreno Almazán, L.; López Martín, M.C.; Domínguez Antonaya, M.; Vera Mendoza, M.I.; Aparicio, J.; Martínez, V.; Tagarro, I.; Fernández-Nistal, A.; Lumbreras, S.; Maté, C.; Montoto, C. "Clinical characteristics and prognostic factors for Crohn's disease relapses using natural language processing and machine learning – a pilot study", European Journal of Gastroenterology & Hepatology, , .