Incidence, Prevalence, and Clinical Characteristics of Patients With Chronic Lymphocytic Leukemia (CLL) in Spain: Natural Language Processing (NLP) Analysis of Electronic Health Records (EHRs)

Jose Angel Hernandez, MD[1], Nasim Bahar, MSc, MPH[2], Cristina Bas, MD[3], Alberto Prieto, PhD[2], Macarena Ortiz, MD[4], Eva Castillo, MD[5], Antonio Gutierrez, MD[6]

[1]University Hospital Infanta Leonor, Madrid, Spain; [2]BeiGene International, GmbH, Basel, Switzerland; [3]BeiGene ESP, SL, Madrid, Spain; [4]Regional University Hospital Malaga, Malaga, Spain; [5]Savana Research, Madrid, Spain; [6]Lymphoma Unit, Department of Hematology, Son Espases University Hospital/IdISBa, Palma, Spain

OBJECTIVES: Unlike registries or claim-based real-world data, NLP analysis of EHRs analyzes diverse datasets, thereby reducing selection bias to provide accurate patient data. This study used free-text data extracted with NLP from EHRs to determine the incidence, prevalence, and primary clinical characteristics of CLL in Spanish patients to reduce knowledge gaps and enhance disease management.

METHODS: This was a multicenter, retrospective study in adult patients with CLL from 1 Jan 2016 to 31 Dec 2021 in 3 Spanish hospitals. EHRs were evaluated using EHRead technology, a data-driven system based on NLP and machine learning, according to clinical terminology (SNOMED CT). Incidence and prevalence were estimated, and descriptive statistics were determined for 205 variables.

RESULTS: A total of 697 patients with CLL were included in the study, out of a population of 2,069,341 patients with a total of 88,872,628 EHRs. The overall age-standardized (2013 European population) incidence and prevalence for the study period were 3.38 (95% CI, 2.55-4.22) and 49.81 (95% CI, 47.65-51.98) cases per 100,000 person-years, respectively. The mean age was 72.1 (SD, 12.6) years and 299 patients (42.9%) were female. Among patients with available information, most were smokers or ex-smokers (76.9%) and 38.2% were alcohol drinkers. The most common clinical alterations observed were lymphocytosis (65.9%), lymphadenopathy (35.4%), anemia (26.3%), infections (18.4%), thrombocytopenia (14.2%), and splenomegaly (13.9%). Eastern Cooperative Oncology Group performance status (ECOG PS) was reported in 13.5% of patients, most frequently ECOG PS 0 (43.6%). Cytogenetic alterations were reported in 27.5% of patients.

CONCLUSIONS: This study presents comprehensive information on patients with CLL in Spain, obtained through NLP analysis of EHRs. These findings confirm the clinical characteristics described in the literature and reinforce the role of artificial intelligence and NLP as reliable methods for studying disease incidence and prevalence.