HealthManagement
Promoting Management and Leadership

Machine Learning Model to Identify Social Needs From Patient Medical Notes

Social needs and social determinants of health (SDOH) are significant factors influencing clinical outcomes, yet they are often under-documented in healthcare systems' electronic health records (EHRs). Currently, the majority of data regarding social needs and SDOH challenges within EHRs are captured as unstructured medical notes rather than structured data. Researchers aimed to develop a scalable machine learning model for identifying 3 major domains of social needs (residential instability, food insecurity, and transportation issues) from the unstructured data in EHRs.

Addressing challenges in documenting social needs and SDOH in EHRs

With the predominance of unstructured data, the difficulties in recognising social needs as part of disease aetiology, and the lack of standardised coding systems, documenting social needs and social determinants of health in electronic health records poses challenges. However, efforts are being made to address these issues, with a focus on assessing, documenting, and intervening to address social needs and SDOH as part of standard care processes. Machine learning (ML) techniques, particularly natural language processing (NLP), offer promise in extracting information from unstructured EHR data. While traditionally challenging due to variability, recent developments in NLP have enabled reliable extraction of social needs information from EHRs. The article presents the application of NLP techniques to identify patients' social needs in an EPIC-based EHR of a healthcare system in Maryland. A scalable, performant, and rule-based model was developed and tested to identify three major domains of social needs: residential instability, food insecurity, and transportation issues. These domains were prioritised based on their prevalence in the patient population and the need for screening and referral initiatives.

Study design and patient selection for assessing social needs in JHHS

The study included patients aged 18 and above who received care at Johns Hopkins Health System (JHHS) from July 2016 to June 2021 and had at least one free-text note in their electronic health records (EHRs). Table S1 defined the social needs assessed in the study and provided examples of their documentation in the EHR. The study protocol was approved by the Institutional Review Board at Johns Hopkins University School of Public Health. A total of 1,879,626 patients with encounters during the specified period were identified, which was narrowed down to 1,317,335 patients with valid Maryland jurisdiction codes. Among these, 5502 patients had an ICD-10 diagnosis code for residential instability or food insecurity. A matched control sample of 5502 patients without these diagnosis codes was also selected based on age, gender, and place of living. Positive labels (1) were assigned to patients with relevant ICD-10 codes for residential instability or food insecurity, while negative labels (0) were assigned to patients without these codes.

Development and validation of a rules-based NLP algorithm for identifying social needs in EHRs

The study presents the development of a novel rules-based Natural Language Processing (NLP) algorithm to identify various social needs domains from free-text notes in Electronic Health Records (EHRs), with the potential for deployment in healthcare systems. The algorithm's performance was validated on a human-labelled dataset and then applied to a population-level dataset to determine the prevalence of selected social needs. The algorithm demonstrated satisfactory performance across different social domains, with the algorithm for residential instability-homelessness performing the best overall. However, food insecurity and transportation issues algorithms tended to produce more false negatives than false positives, suggesting that the obtained population values may indicate the lower bound for social needs.

Refinement and scalability of the NLP algorithm for clinical deployment

The development process involved a manual approach for key phrase development and a rule-based approach for their identification in free-text notes, which introduced potential subjectivity and bias. To address this, a semi-automated approach using ngram, keyword matching, and statistical analysis was applied to refine keywords. The algorithm's performance was comparable to previous studies, demonstrating its potential for deployment as a clinical decision-support tool. The study also discusses the scalability of the developed pipeline for handling large datasets efficiently, using technologies such as Spark and Spark NLP. The pipeline's performance on provider notes for the patient population indicated its suitability for deployment in clinical settings, with significantly reduced processing time compared to previous pipelines. Furthermore, the

study explores the performance of Machine Learning (ML) models, particularly a Naive Bayes (NB) model, in comparison to the rule-based model. The NB model demonstrated better performance overall and showed robustness in generating false negatives, which is crucial for identifying less frequently documented events like social needs. The study suggests the potential integration of rule-based and ML models into an ensemble model to achieve higher performance.

The authors researched a novel model for identifying social needs from free-text notes, with the potential for deployment in healthcare systems. Both their rule-based algorithm and Naive Bayes Machine Learning model exhibited satisfactory performance across selected social domains. The scalability and efficiency of the pipeline design using suitable technologies allow for the effective processing of large datasets and potential deployment without major modifications. These models could be adapted and integrated into an ensemble model for enhanced performance in efficiently identifying social needs in EHR free-text notes, enabling targeted interventions among patients with such needs. Future research should focus on investigating the generalizability of these models using larger and more diverse datasets to ensure effectiveness across different patient populations. External validation using data from various healthcare systems would further enhance the reliability and applicability of the models in diverse settings.

Source: **JAMIA Open**

Image Credit: **iStock**

Published on : Mon, 18 Mar 2024