Final Project

Enoch Kang
ekang79@gatech.edu

Abstract—Social determinants of health (SDoH) carry extensive medical significance. However, due to their multifactorial and qualitative nature, most electronic health records (EHR) systems today do not offer many functions with respect to evaluation of SDoH. This often leads to an undeserved neglect of SDoH in medical decision making. Using a generic dataset of medical records, this project trained a prototype classification model that classifies a patient's SDoH level into Low, Medium, or High based on marital status, employment status, yearly family income, and education level. Then the model is incorporated into a SMART on FHIR application that a medical provider can use to collect or supply information regarding a given patient's SDoH to classify his or her level of SDoH. This allows for a more intuitive documentation and systematic evaluation of SDoH than what most EHR systems offer today Also, this facilitates a more ready involvement of SDoH in medical decision making.

1 BACKGROUND

Many studies highlight the importance of socioeconomic factors on health. Jemal et al. concluded that almost half of all deaths among working-aged adults in the U.S. were accounted for by potentially avoidable factors associated with lower educational status (Jemal, et al. 2008). Meta-analysis of Galea and colleagues demonstrated that comparable deaths occur due to low education, racial segregation, and low social support as to myocardial infarction, cerebrovascular disease, and lung cancer, respectively (Galea, et al. 2011). More recent studies and reviews only reverify the importance of social factors on health (Braveman & Gottlieb 2014). However, such importance is yet to be reflected by most medical systems.

2 PROBLEM STATEMENT

Most information regarding significant socioeconomic factors are often sporadically recorded into different domains of electronic health records (EHRs) while systematic evaluation of an individual's social determinants of health (SDoH) is rarely possible. This hinders incorporation of SDoH into medical decision making.

3 SOLUTION

The goal of this project was to develop an application that enables the medical providers to incorporate patients' SDoH into medical decision making more readily.

3.1 Collecting the patient's SDoH information

Through literature review, relevant SDoH's were determined as follows: educational level, income level, employment status, and marital status (References 5-7). Initially, I intended to use natural language processing models like ClarityNLP to mine the patient's SDoH information that is sporadically present across his or her medical records. However, inspection of all medical records to find four specific pieces of information seemed unnecessarily inefficient, especially when all the factors have corresponding LOINC codes. As a result, I decided to collect the information by retrieving pieces of information with specific locators like LOINC codes.

3.2 Training a SDoH classification model

I used a publicly available Synthea sample dataset to train a classification model. An initially unforeseen obstacle was that essentially no labelled datasets exist regarding SDoH. Because I considered classification through machine learning algorithm as an essential component of this prototype application, I decided to use pseudo-labels instead of removing the classification model altogether from the project. I performed clustering with the sample dataset to divide the group into three clusters. Upon careful examination of the resulting clusters, I assigned High, Medium, and Low to appropriate clusters. Then, using the clusters as labels, I trained a decision tree classification model that classifies the level of SDoH given a patient's educational level, yearly family income, employment status, and marital status.

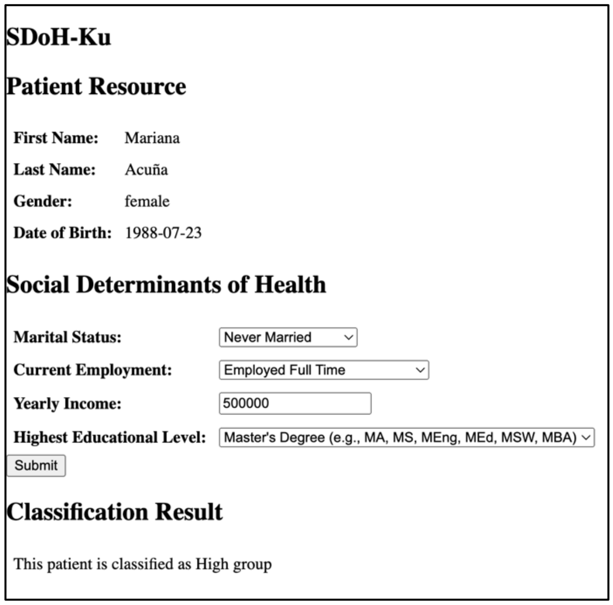

Figure 1 — Sample result screen

3.3 Developing an application that incorporates all the parts

A SMART on FHIR application was developed as initially intended. An initially unforeseen obstacle was that I needed to figure out how to connect the different components of the application when they are developed in different languages. While most of the application itself is built in Javascript, the classification model was developed in Python. Flask was used to set up a Python backend as a solution. A successful application run looks like Figure 1.

4 OUTCOME AND FURTHER WORK

A completely functional application was developed. When application launches, the medical provider logs in and selects a patient. Available information are automatically retrieved and displayed on screen while the provider could edit the retrieved information or type in any missing information. When the medical provider clicks *Submit*, the application classifies the patient's SDoH level and displays it on screen.

The project was generally successful in that an application that retrieves and classifies a patient's SDoH was fully developed. A medical provider could use the results to incorporate SDoH into making medical decisions. However, there is room for some improvements or further work as well. For example, a non-generic dataset could be used. This project used sample synthetic patient records provided by Synthea. The use of a real, non-synthetic dataset could improve the accuracy of the classification model. Also, while this project used clustering results as pseudo-labels, classification based on this is somewhat arbitrary. The use of a dataset with ground truth labels would be ideal. Alternatively, further research on the actual differences in medical outcomes among these clusters may strengthen the justification for the use of such clusters. Lastly, much more careful ethical considerations are necessary before further developing this application because SDoH involve sensitive personal information like marital status and income. Building a machine learning classification model based on such sensitive information inevitably brings ethical controversies. While this project is relatively free from such controversies because it is only a prototype that involves synthetic data, if further improvement on this project is to be pursued in the future, much effort would be necessary to ethically build and run this application.

5 REFERENCES

1. Braveman, P., & Gottlieb, L. (2014). The social determinants of health: it's time to consider the causes of the causes. Public health reports, 129(1_suppl2), 19- 31.
2. Chen, M., Tan, X., & Padman, R. (2020). Social determinants of health in elec- tronic health records and their impact on analysis and risk prediction: a sys- tematic review. Journal of the American Medical Informatics Association, 27(11), 1764-1773.
3. Galea, S., Tracy, M., Hoggatt, K. J., DiMaggio, C., & Karpati, A. (2011). Esti- mated deaths attributable to social factors in the United States. American journal of public health, 101(8), 1456-1465.
4. Jemal, A., Thun, M. J., Ward, E. E., Henley, S. J., Cokkinides, V. E., & Murray, T. E. (2008). Mortality from leading causes by education and race in the United States, 2001. American journal of preventive medicine, 34(1), 1- 8.Joyner, D. A. (2017). Scaling Expert Feedback: Two Case Studies. In *Proceed- ings of the Fourth Annual ACM Conference on Learning at Scale*. Cambridge, Massachusetts.
5. Lantz, Paula M., et al. "Socioeconomic factors, health behaviors, and mortality: results from a nationally representative prospective study of US adults." Jama 279.21 (1998): 1703-1708.
6. "Social Determinants of Health." *World Health Organization*, 30 May 2019, www.who.int/health-topics/social-determinants-of-health#tab=tab_1.
7. "Social Determinants of Health - Healthy People 2030 | health.gov." *U.S. Department of Health And Human Services*, health.gov/healthypeople/priority-areas/social-determinants-health. Accessed 29 Oct. 2023.