Winter 12-10-2022

Identifying Features for the Prediction of Housing Instability in Patient Populations

Joshua R. Vest

Ofir Ben-Assuli

Identifying Features for the Prediction of Housing Instability in Patient Populations

Completed Research Paper (Extended Abstract)

Joshua R Vest
Indiana University & Regenstrief Institute
joshvest@iu.edu

Ofir Ben-Assuli
Ono Academic College
ofir.benassuli@gmail.com

Abstract

Computable phenotypes are representations of patient characteristics or conditions that can be obtained from electronic health records (EHRs) and other data sources. Computable phenotypes may be an approach to address the challenge healthcare organizations face in measuring patients' social factors for intervention and service delivery. Housing instability is a strong determinant of overall health, utilization, and cost. Instable housing increases stress and depression, financial concerns, disruptions to social support, and exposure to health risks. This analysis sought to identify relevant features for use in the development of a preliminary housing instability computable phenotype using both rule-based and supervised machine-learning techniques. This preliminary study's sample size was not sufficient to develop a phenotype and overall prediction of housing instability had mediocre performance. However, the machine learning approaches appear stronger than a rule-based algorithm and do provide guidance for future development.

Keywords

Computable phenotyping; feature engineering; expert panel; healthcare

Introduction

Computable phenotypes are representations of patient characteristics or conditions that can be obtained from electronic health records (EHRs) and other data sources by combining a defined set of features and logical expressions (Verchinina et al. 2018). Computable phenotypes may be the product of human-developed rule-based algorithms or classification via machine learning techniques (Shivade et al. 2014). Computable phenotypes have been developed for clinical conditions such as asthma, diabetes, COVID-19 infection, and others. Once developed, computable phenotypes serve as a solution to the challenges of imprecision definitions about human conditions and are useful for research activities such as decision support, cohort identification, and risk stratification analyses.

Social factors encompass the host of patients' nonclinical, economic, contextual, and psychosocial characteristics and are important drivers of morbidity, mortality, unnecessary utilization, disparities, and increased costs (Pruitt et al. 2018). Social factors may be amenable to representation by computable phenotypes. For example, EHRs contain numerous data elements not directly related to patient health or clinical care but may be crucial to predict housing instability. For example, demographics, insurance information, billing histories, appointment status, emergency contacts, and language preferences are reflective of characteristics of social and economic wellbeing. Likewise, EHRs contain address information, which can be linked to geospatial repositories to gain a broader understanding of patients' environments.

The objective of this paper was to identify relevant features for use in the development of a preliminary computable phenotype for one social factor, housing instability, using both rule-based and supervised machine-learning techniques.

Methods

The study sample included 165 adult primary care and emergency department (ED) patients who sought care at an Indianapolis, IN USA safety-net hospital system during August and September of 2020. Patients were eligible for inclusion if they: were 18 years old or older; did not require an interpreter; were able to complete a self-administered survey unassisted; and were not marked as positive for COVID-19 symptoms in the EHR scheduling system. Housing instability was collected via patient surveys and all patients received an incentive for participation. Survey responses were linked via patient identifiers to the hospital system's EHR records. Lastly, survey responses were linked by address to aggregate measures from the US Census Bureau. Prevalence estimates social factors, including housing instability, among the study population have been published previously (Vest et al. 2021).

Housing instability was defined as housing disruptions or related problems, from frequent moves or difficulty paying rent to being evicted or being homeless (Burgard et al. 2012). We measured housing instability using the 10-item Housing Instability Index (Rollins et al. 2012). The dichotomous items sum to an overall risk score (Cronbach's alpha = 0.80). We classified those in the top quartile (score of 3 or greater) as housing instable (Rollins et al. 2012).

From the three data sources, we extracted and created features representative of *demographics and contact information*, *encounter history*, c*linical history*, *geospatial characteristics*, and *financial information*. These features were the product of an expert panel on healthcare data for social factor measurement (Vest et al. 2022) and a previously developed and validated NLP algorithm (Allen et al. 2021).

For the rule-based algorithm, we adopted a hierarchical approach to identify patients with housing instability. Given that positive cases of social factors are more likely to be documented than negatives within EHRs, we focused on those features an expert panel perceived as the most distinguishing indicators of a positive case of housing instability. First, we prioritized documented instance of homelessness, which is a the most acute manifestation of housing instability. Second, we focused positive cases identified through direct provider-patient interaction through screening or as documented in clinical notes. Next, we looked at features that were consistent with existing definitions housing instability. Lastly, because housing instability includes the ability to pay rent or mortgage, we also included indication of financial stress.

We used three supervised machine-learning models to predict housing instability using scikit-learn package (Pedregosa et al. 2011): Random Forest with the Gini criterion, with the Entropy criterion, and XGBoost. We split the dataset into training dataset and testing dataset and performed five-fold cross validation. We used SMOTE (over sampling) to resample the data from imblearn package (Lemaître et al. 2017) to account for the imbalanced data.

Results

The prediction models (with resampling) were better than the rule-based model in terms of AUC (as well as specificity, precision and accuracy). In addition, all the approaches where highly specific, e.g. among those without housing instability a high proportion are correctly identified. Conversely, sensitivity, was very poor for all tests, e.g. each approach was very poor at identifying those with housing instability among those screening positive.

Model	**AUC (95% C.I.)**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
Random Forest - Gini	0.843 (0.793-0.892)	77.1	**77.1**	**77.1**	77.1
Random Forest - Entropy	**0.847** (0.798-0.895)	**78.8**	75.4	76.9	**77.5**
XGBoost	0.727 (0.662-0.792)	73.7	62.7	66.4	68.2
Rule-based	0.564	34.0	78.8	39.0	24.8
Table 4. Prediction models using oversampled (SMOTE) dataset					

Discussion

Development of a computable phenotype would be useful both for cohort identification (for research studies) and within decision support systems to facilitate referrals to services. This preliminary study's sample size was not sufficient to develop a phenotype and overall prediction of housing instability had mediocre performance. However, the machine learning approaches appear stronger than a rule-based algorithm and do provide guidance for future development.

References

Allen, K., Hood, D., Cummings, J., Kasthurirathne, S., Embi, P., and Vest, J. 2021. *Extracting Social Variables from Clinical Documentation to Better Facilitate Response to Patient Need*, presented at the AMIA Annual Fall Symposium, San Diego, CA.

Burgard, S. A., Seefeldt, K. S., and Zelner, S. 2012. "Housing Instability and Health: Findings from the Michigan Recession and Recovery Study," *Social Science & Medicine* (75:12), Part Special Issue: Place, Migration & Health, pp. 2215–2224.

Epic Corporation. 2018. *Domain Questions, Answers, and Risk Classification Logic*.

Frederick, T. J., Chwalek, M., Hughes, J., Karabanow, J., and Kidd, S. 2014. "How Stable Is Stable? Defining and Measuring Housing Stability," *Journal of Community Psychology* (42:8), pp. 964–979.

Lemaître, G., Nogueira, F., and Aridas, C. K. 2017. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research* (18:17), pp. 1–5.

Padgett, D. K. 2020. "Homelessness, Housing Instability and Mental Health: Making the Connections," *BJPsych Bulletin* (44:5), pp. 197–201.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. 2011. "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research* (12:85), pp. 2825–2830.

Pruitt, Z., Emechebe, N., Quast, T., Taylor, P., and Bryant, K. 2018. "Expenditure Reductions Associated with a Social Service Referral Program," *Population Health Management* (21:6), pp. 469–476.

Rollins, C., Glass, N. E., Perrin, N. A., Billhardt, K. A., Clough, A., Barnes, J., Hanson, G. C., and Bloom, T. L. 2012. "Housing Instability Is as Strong a Predictor of Poor Health Outcomes as Level of Danger in an Abusive Relationship: Findings From the SHARE Study," *Journal of Interpersonal Violence* (27:4), pp. 623–643.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M. 2014. "A Review of Approaches to Identifying Patient Phenotype Cohorts Using Electronic Health Records," *Journal of the American Medical Informatics Association* (21:2), pp. 221–230.

Verchinina, L., Ferguson, L., Flynn, A., Wichorek, M., and Markel, D. 2018. "Computable Phenotypes: Standardized Ways to Classify People Using Electronic Health Record Data," *Perspectives in Health Information Management* (Fall), 1=6.

Vest, J. R., Adler-Milstein, J., Gottlieb, L. M., Bian, J., Campion, T. R., Cohen, G. R., Donnelly, N., Harper, J., Huerta, T. R., Kansky, J. P., Kharrazi, H., Khurshid, A., Kooreman, H. E., McDonnell, C., Overhage, J. M., Pantell, M. S., Parisi, W., Shenkman, E. A., Tierney, W. M., Wiehe, S., and Harle, C. A. 2022. "Assessment of Structured Data Elements for Social Risk Factors," *The American Journal of Managed Care* (28:1), pp. e14–e23.

Vest, J. R., Wu, W., and Mendonca, E. A. 2021. "Sensitivity and Specificity of Real-World Social Factor Screening Approaches," *Journal of Medical Systems* (45:12), p. 111.

Weeks, W. B., Cao, S. Y., Lester, C. M., Weinstein, J. N., and Morden, N. E. 2020. "Use of Z-Codes to Record Social Determinants of Health Among Fee-for-Service Medicare Beneficiaries in 2017.," *Journal of General Internal Medicine* (35:3), pp. 952–955.