Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230274

Predicting Depression Risk in Patients with Cancer Using Multimodal Data

Anne DE HOND[a,b,1], Marieke VAN BUCHEM[a,b], Claudio FANCONI[b,c], Mohana ROY[b], Douglas BLAYNEY[b], Ilse KANT[d], Ewout STEYERBERG[a] and Tina HERNANDEZ-BOUSSARD[b]
[a] *Leiden University Medical Center, Leiden, The Netherlands*
[b] *Stanford University, Stanford, CA, USA*
[c] *ETH Zürich, Zürich, Switzerland*
[d] *University Medical Center Utrecht, Utrecht, The Netherlands*

Abstract. When patients with cancer develop depression, it is often left untreated. We developed a prediction model for depression risk within the first month after starting cancer treatment using machine learning and Natural Language Processing (NLP) models. The LASSO logistic regression model based on structured data performed well, whereas the NLP model based on only clinician notes did poorly. After further validation, prediction models for depression risk could lead to earlier identification and treatment of vulnerable patients, ultimately improving cancer care and treatment adherence.

Keywords. Natural Language Processing, machine learning, oncology, depression

1. Introduction

Patients with cancer starting invasive treatment programs often develop depression that physicians struggle to recognize at an early stage [1,2]. We developed a prediction model for early identification of patients at risk for depression within the first month of chemo- or radiotherapy treatment to assist physicians and healthcare workers.

2. Methods

We included adult patients receiving cancer treatment from a comprehensive cancer center in the United States (2008-2022). Exclusion criteria were patients younger than 18 years, no clinician notes within the two weeks prior to treatment or a depression diagnosis within the year prior to treatment. Depression was defined as a depression diagnosis via ICD-9 and ICD-10 codes. Depression risk was predicted within one month after the start of cancer treatment. We included several structured data features from the patient's Electronic Health Record (EHR), like gender, age, cancer stage, history of depression, and patient email classification scores. We also included unstructured text data from clinician notes. Several machine learning (ML) models (e.g., LASSO logistic regression, random forest, gradient boosting decision trees) were compared to predict

[1] Corresponding Author: Anne de Hond, E-mail: a.a.h.de_hond@lumc.nl.

depression risk on combinations of the structured data. Three (multimodal) Natural Language Processing models (DistilBERT [3]) were developed on different combinations of the structured data and unstructured data. We split data randomly for all models in the same 2/3 train and 1/3 test set. Model performance was measured via the area under the receiver operating characteristic curve (AUC) and calibration plots (slope and intercept). To identify potential fairness issues for specific demographic groups, calibration slope and intercept were also compared across gender and race/ethnicity.

3. Results

A total of 437 (3%) of 16,159 patients received a depression diagnosis within one month after the start of cancer treatment. The best performing ML model (LASSO logistic regression based on structured data) had an AUC of 0.74 (95% CI 0.71-0.78), whereas the model based solely on clinician notes performed poorly (0.50 AUC, 95% CI 0.49-0.52). The model underestimated risks for female and Non-Hispanic Black patients and overestimated for male and Non-Hispanic Asian patients.

4. Discussion

The best performing model (LASSO logistic regression on structured data) had reasonable AUC and calibration. We found discrepant model calibration across race/ethnicity and sex. The miscalibration could result in a disproportionate amount of missed patients needing additional mental health resources in specific groups. A next step could be to apply bias mitigation techniques for in- or post-processing during model development, like threshold selection and recalibration within groups.

5. Conclusion

We developed a robust model to predict depression risk among patients with cancer and demonstrated the importance of structured data to predict depression risk. Future studies might improve the prediction of depression risk in patients with cancer by refining the outcome label and supplementing predictors with patient-reported outcome measures and social determinants of health.

References

[1] Pitman A, Suleman S, Hyde N, Hodgkiss A. Depression and anxiety in patients with cancer. BMJ. 2018;361:k1415. doi: 10.1136/bmj.k1415.

[2] Dreismann L, Goretzki A, Ginger V, Zimmermann T. What if… I Asked Cancer Patients About Psychological Distress? Barriers in Psycho-Oncological Screening From the Perspective of Nurses-A Qualitative Analysis. Front Psychiatry. 2021;12:786691. Epub 2022/02/15. doi: 10.3389/fpsyt.2021.786691. PubMed PMID: 35153856; PubMed Central PMCID: PMCPMC8825354.

[3] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108. 2019.