2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Prediction of Poor Mental Health Following Breast Cancer Diagnosis Using Random Forests [1]

Eugenia Mylona, Konstantina Kourou, Georgios Manikis , Haridimos Kondylakis, Kostas Marias, Evangelos Karademas, Paula Poikonen-Saksela, Ketti Mazzocco, Chiara Marzorati, Ruth Pat-Horenczyk, Ilan Roziner, Berta Sousa, Albino Oliveira-Maia, Panagiotis Simos and Dimitrios I. Fotiadis, *Fellow Member IEEE*

***Abstract*— Breast cancer diagnosis has been associated with poor mental health, with significant impairment of quality of life. In order to ensure support for successful adaptation to this illness, it is of paramount importance to identify the most prominent factors affecting well-being that allow for accurate prediction of mental health status across time. Here we exploit a rich set of clinical, psychological, socio-demographic and lifestyle data from a large multicentre study of patients recently diagnosed with breast cancer, in order to classify patients based on their mental health status and further identify potential predictors of such status. For this purpose, a supervised learning pipeline using cross-sectional data was implemented for the formulation of a classification scheme of mental health status 6 months after diagnosis. Model performance in terms of AUC ranged from 0.81± 0.04 to 0.90± 0.03. Several psychological variables, including initial levels of anxiety and depression, emerged as highly predictive of short-term mental health status of women diagnosed with breast cancer.**

I. Introduction

Breast cancer (BC) has become the most commonly diagnosed cancer, surpassing lung cancer, and it accounts for 25% of all cancer cases among women [1]. With numbers of patients expected to rise further due to the increasing trends in both incidence and survival, BC emerges as a major public health problem and socio-economic challenge [2,3].

Diagnosis of BC may have a tremendous impact on self-image and sexuality, among other psychosocial factors, and treatment-related secondary effects on physical appearance and reproductive potential [4]. Thus, patients with BC often experience intense psychological reactions, such as denial, anger, helplessness, hopelessness and fear, that in many cases may condition psychiatric morbidity and suicidal thoughts [5]. In fact, among BC survivors, the prevalence of mental disorders, such as anxiety and depression, may be as high as 48% in the year after diagnosis [6]. Early identification of patients displaying such symptoms is crucial in order to anticipate treatment, prevent further deterioration of mental health and improve quality of life. It is also of paramount importance to understand which factors discriminate BC patients who experience mental health stability from those who do not [7]. Identifying potentially beneficial or detrimental factors can have important clinical implications and, eventually, guide interventions to support patients in their psychological recovery from the experience of cancer.

Despite the growing need for improving our understanding and capacity to predict mental health in women diagnosed with BC, research and high-quality evidence on mental well-being remain scarce. This is mainly due to the extreme complexity of the phenomenon, that needs to be modelled, encompassing clinical, physiological, molecular, lifestyle and psychological components. Although, researchers have sought to explain the complex process of psychological adaptation to BC, in the majority, they have intentionally restricted their scope to a limited set of predictors as a compromise to the use of conventional statistical methods [8,9].

In the era of Big Data, Machine Learning (ML) algorithms have emerged as an appealing alternative to conventional statistical approaches for improving prediction accuracy, thanks to their ability to efficiently handle a large amount of heterogeneous data and complex interactions[10,11]. Random Forests (RF), in particular, is efficient in handling highly non-linear data and a large number of features, agile in terms of noise in data, and simpler to tune. In psycho-oncology, ML algorithms could be applied for developing clinically adoptable approaches to identify patients at risk, improve risk prediction of poor mental health and discriminate patients based on their mental health status.

The goal of this study was (i) to develop a robust predictive model of mental health among patients with BC, using a large pool of heterogeneous data, namely, clinical, psychosocial, socio-demographic and lifestyle data and (ii) to identify the

[*]This work is funded by the European Commission: Project BOUNCE. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777167.

E. Mylona, K. Kourou and D. I. Fotiadis are with the Department of Biomedical Research, FORTH-IMBB, Ioannina, Greece and the Unit of Medical Technology and Intelligent Information Systems, Materials Science and Engineering Department, University of Ioannina, Ioannina, GR 45110, Greece (corresponding author phone: +301651009006; fax: +302651008889; e-mail: fotiadis@uoi.gr).

G. Manikis, H. Kondylakis, K. Marias and E. Karademas are with the Computational Biomedicine Laboratory, FORTH-ICS, Heraklion, Greece.

P. Poikonen-Saksela is with the Helsinki University Hospital Comprehensive Cancer Center and Helsinki University, Finland.

K. Mazzocco is with the Department of Oncology and Hemato-oncology, University of Milan and the Applied Research Division for Cognitive and Psychological Science, European Institute of Oncology IRCCS, Milan, Italy.

C. Marzorati is with the Applied Research Division for Cognitive and Psychological Science, European Institute of Oncology IRCCS, Milan, Italy.

R. Pat-Horenczyk is with the School of Social Work and Social Welfare,The Hebrew University of Jerusalem, Israel.

I. Roziner is with Department of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel

B. Sousa is with the Breast Unit, Champalimaud Clinical Centre/ Champalimaud Foundation, Champalimaud Research, Lisboa, Portugal.

A. Oliveira-Maia is with the Champalimaud Research and Clinical Centre, Champalimaud Centre for the Unknown, Lisboa, Portugal.

P. Simos is with the Computational Biomedicine Laboratory, FORTH-ICS and the School of Medicine, University of Crete, Heraklion, Greece.

most important factors that are related to and distinguish between BC patients who experience mental stability and those who do not. Cross-sectional ML techniques based on RF were used to formulate a classification scheme of mental health status, 6 months after diagnosis.

II. Materials and Methods

A. Study population

The study population consisted of 731 female BC patients from a large multicentre prospective study at five clinical centers: the European Institute of Oncology (IEO) in Italy (n=205), the Rabin and Shaare Zedek Medical Centers (HUJI) in Israel (n=151), the Helsinki University Hospital (HUS) in Finland (n=236) and the Champalimaud Clinical Centre (CHAMP) in Portugal (n=139). The study was developed in accordance with the principles stated in the Declaration of Helsinki and was approved by the European Institute of Oncology Ethical Committee at the IEO (Approval No R868/18-IEO916) and the ethical committees of each participating hospital.

All participants have provided signed informed consent. They had histologically confirmed invasive early or locally advanced operable BC with tumor stage I, II or III and they received some type of medical therapy. The study was addressed to patients aged 40 to 70 years. Analysis was conducted on data from 690 patients (mean age = 55 years), from whom complete data on all variables were available.

Participants were assessed at three time points: baseline, 3, and 6 months thereafter. At baseline (M0), occurring within 3-4 weeks from diagnosis, only non-cancer-specific measures were administered. Cancer-specific measures were assessed at Month 3 (M3) and Month 6 (M6), when patients had meaningful experience with the illness. The number of patients with complete outcome records at M3 and M6 were 604 and 549, respectively. All data were integrated and homogenized using an ontology [12] and a data infrastructure built to enable the uninterrupted data collection, integration, cleaning and homogenization [13].

B. Outcome description

The outcome measure for this study was self-reported severity of mental health symptoms, 6 months after BC diagnosis, assessed using the Hospital Anxiety and Depression scale (HADS) [14]. It is a 14-item scale, with seven items relating to anxiety and seven relating to depression symptoms. HADS is a widely used self-report scale that has clinically validated cutoffs for poor mental health (indicated by clinically significant symptoms of anxiety and depression). Accordingly, HADS scores were binarized with a cutoff value of 14 points. Participants who scored higher than the threshold value were assigned to the poor mental health group.

C. Predictor variables

Considering different sets of predictors, three models were developed for classifying BC patients based on their mental health level at M6. A large heterogeneous set of continuous and discrete variables with acceptable number of non-missing values (≤ 30%) were considered, including:

- clinical variables (e.g. medical history, BMI, cancer stage, menopausal status, treatment type)
- socio-demographic variables (e.g. age, education, income, marital status, income, employment),
- lifestyle variables (e.g. exercise, diet, smoking)
- psychosocial variables from validated questionnaires (e.g. social support, dispositional optimism, sense of coherence, flexibility in coping, mindfulness, positive and negative affect, quality of life etc.)

In total, 42 variables, collected at baseline (M0), were included in the first model (Model A). In addition to these variables, baseline mental health data were considered for the second model (Model B; 46 variables). The third model (Model C) included the variables from Model B in addition to mental health variables collected at M3 (55 variables). Details about the models and variables can be found elsewhere [15].

C. Data pre-processing

Data preprocessing included imputation of missing values and oversampling/ undersampling to account for class imbalance in the outcome variable. Multiple imputation through chained equations (MICE), using the R package MICE, was performed to maximize information and minimize bias due to missing data (20 imputations using 50 iterations). Imputation methods included predictive mean matching for numerical variables, logistic regression for binary variables and polytomous logistic regression for ordinal variables.

In total, 457 (83%) patients had HADS scores indicative of adequate mental health and 92 (17%) reported psychological symptoms indicative of poor mental health status at M6, suggesting an important class imbalance. Outcome classes were balanced with ratio 55% adequate/45% poor mental health, in a two-step process where the Synthetic Minority class Oversampling Technique (SMOTE) [16] was combined with Wilson's Edited Nearest Neighbor Rule (ENN) [17]. Firstly, SMOTE oversampling was applied to the minority class (poor mental health group) in order to generate "synthetic" samples by interpolating new points between marginal outliers and inliers. Then, undersampling using ENN was used in order to (1) remove noisy examples from the majority class (adequate mental health group) and (2) clean "synthetic" samples (generated with SMOTE in the previous step) that were created too deep in the majority class.

D. Model building and performance assessment

Fig. 1 shows the analysis workflow summarized in five steps. First, data imputation was applied to impute missing values in the original dataset (step 1). To avoid overfitting and make reliable predictions, a nested cross-validation (CV) scheme was followed with the "outer" split serving for evaluation of model performance on the unseen test-out set while the "inner" split was used for building the models and tune the hyper-parameters (steps 2-5). The imputed dataset, was first split into 5 outer folds stratified by outcome class and clinical center (step 2). In every iteration, SMOTE and ENN were applied to the train-out set to balance the two classes (step 3). The RF classifier with 1000 trees was selected for model training and validation. The balanced train-out set was further split into 5 inner-folds and the inner CV was repeated 5 times for hyper-parameter optimization through grid search (step 4). The F1 score was used as a performance metric and the best model was selected according to the 1 standard error (SE) rule

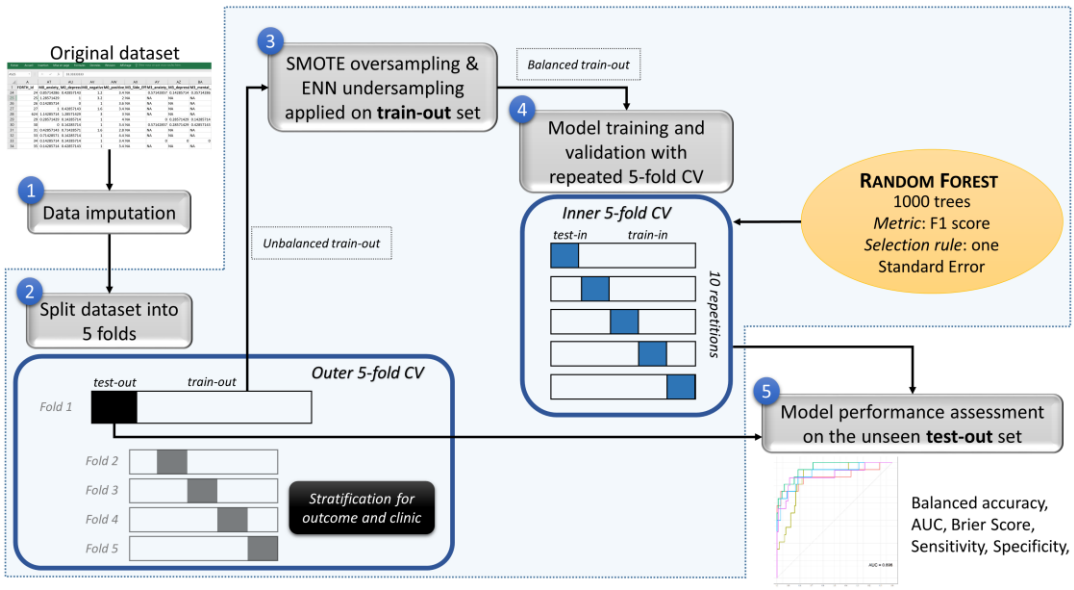

Figure 1. The analysis workflow summarized in five steps.

(the model within 1 SE from the model with maximum F1 score). Feature importance was measured by the mean decrease in impurity (Gini index).

Model performance was assessed on the corresponding test-out set (step 5), using the balanced accuracy, area under the ROC curve (AUC), F1 score, Brier score, Sensitivity, and Specificity. Overall performance was computed as the average of the performances on each test-out set.

III. RESULTS

In Table I, the performance of each model for predicting mental health status at M6 is presented. Overall performance for Model C (including both M0 and M3 mental health data) was higher compared to Model B (with M0 mental health data) and Model A (without M0 or M3 mental health data). For instance, AUC was 0.90 for Model C while for Models A and B AUC was 0.81 and 0.85, respectively. It is noteworthy that sensitivity was considerably higher in Models B and C (up to 0.75) compared to Model A (0.57).

TABLE I. MODEL PERFORMANCE FOR MENTAL HEALTH PREDICTION

Metrics	*Model A*	*Model B*	*Model C*
Balanced Accuracy	0.72 ± 0.07	0.77 ± 0.09	0.81 ± 0.04
ROC AUC	0.81 ± 0.04	0.85 ± 0.04	0.90 ± 0.03
Brier Score	0.13 ± 0.01	0.13 ± 0.01	0.11 ± 0.01
F1 score	0.50 ± 0.10	0.54 ± 0.11	0.61 ± 0.07
Sensitivity	0.57 ± 0.15	0.70 ± 0.18	0.75 ± 0.11
Specificity	0.87 ± 0.03	0.84 ± 0.04	0.86 ± 0.04

Fig. 2 displays variable importance estimated by the RF models for the top 15 predictors. For model A, dispositional optimism and sense of coherence were the two most important coherence and negative affect were significantly more important than other variables. For model C, the most important variable was negative affect (measured at M3) followed by depression and anxiety levels, feeling of helpless-hopelessness and anxious preoccupation.

IV. DISCUSSION AND CONCLUSION

In the present study, we developed predictive models for mental health status of patients with BC using a large set of clinical, psychosocial, socio-demographic and lifestyle variables, providing insights into the most prominent factors affecting mental health in this patient population. Models A and B aimed to answer whether mental health status can be predicted solely from baseline data, either excluding (Model A) or including (Model B) as a covariate the mental health status at baseline. We also investigated whether the performance of the aforementioned models may be improved by including in the analysis psychological data collected three months after the diagnosis (Model C).

Satisfactory prediction performance was reached with all models. The progressive improvement from model A to C indicates that levels of anxiety and depression, both at baseline and 3 months thereafter, have important implications for subsequent mental health. These findings are in line with previous studies [18]. Overall, the vast majority of important predictors were psychological variables with the main protective factors being dispositional optimism, sense of coherence and self-efficacy for coping with cancer while detrimental factors were anxiety, depression, hopelessness and negative affect reported by patients during earlier disease stages. Despite the fact that previous studies have highlighted associations of mental health or quality of life with lifestyle, socio-demographic and clinical factors, in our work this was confirmed only for age and physical activity [19].

Nevertheless, these findings may be considered in light of some limitations. The measured outcomes were collected up to only six months after treatment. Further analyses should be conducted on more extensive follow-ups, in order to better analyze patient's mental health along the entire care process. Alternative algorithms for patient classification (such as Gradient Boosting Machines) and for estimating feature importance (such as SHAP values) are forthcoming in future work. Also, HADS scores, as a continuous variable, may be used in a regression context to preserve individual variability

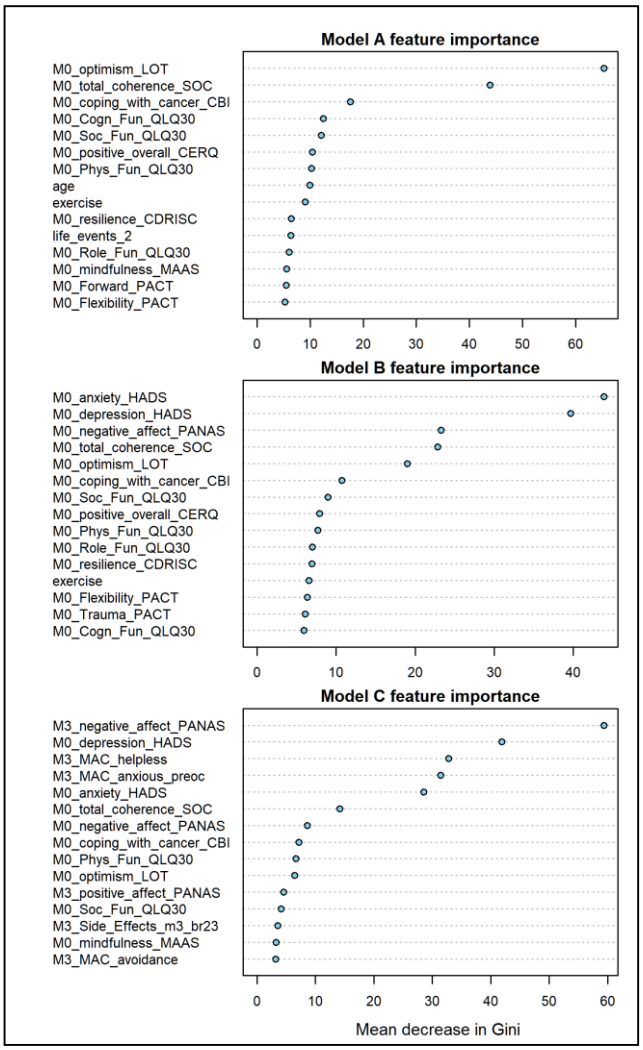

Figure 2. Random forest variable importance for models A, B and C

in symptom severity. Finally, data imputation was performed prior to CV which might have introduced some data leakage to the dataset.

The results of the present work may help clinicians identify patients at high risk of developing mental health disorders in the preliminary phase of the care process. Predicting mental health is critical in making decisions regarding the need of personalized interventions [20]. Early detection of protective and hindering factors related to patients' well-being would help health professionals to tailor preventive psychological programs aimed at enhancing the capacity of BC patients to efficiently adapt to the disease.

To summarize, several psychological variables were found to have a significant impact in the short-term mental health of women diagnosed with BC. Methodological approaches that capture dynamics changes in patient's mental health are still needed in order to unveil the multifactorial process of illness adaptation in the longer term. Within the context of personalized medicine, understanding the role and impact of each factor on patient's trajectory through illness adaptation will enhance the effectiveness of already existing and/or the development of new, more efficient personalized interventions.

REFERENCES

[1] IARC, "Latest Global Cancer Data," 2020. [Online]. Available: https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf.

[2] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, and M. M. Fidler-Benaoudia, "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study," *Lancet Glob. Heal.*, 2020.

[3] H. Kondylakis *et al.*, "Status and Recommendations of Technological and Data-Driven Innovations in Cancer Care: Focus Group Study," *J. Med. Internet Res.*, vol. 22, no. 12, 2020.

[4] M. Ruggeri *et al.*, "Fertility concerns, preservation strategies and quality of life in young women with breast cancer: Baseline results from an ongoing prospective cohort study in selected European Centers," *Breast*, vol. 47, pp. 85–92, Oct. 2019.

[5] J. Walker *et al.*, "Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: A cross-sectional analysis of routinely collected clinical data," *The Lancet Psychiatry*, vol. 1, no. 5, pp. 343–350, Oct. 2014.

[6] H. Carreira, R. Williams, H. Dempsey, S. Stanway, L. Smeeth, and K. Bhaskaran, "Quality of life and mental health in breast cancer survivors compared with non-cancer controls: a study of patient-reported outcomes in the United Kingdom," *J. Cancer Surviv.*, 2020.

[7] H. Kondylakis, P. Simos, E. Karademas, K. Marias, and P. Poikonen-Saksela, "Resilience Indices for Breast Cancer Management," *IEEE Int. Conf. Biomed. Heal. Informatics*, 2021.

[8] M. H. McDonough, C. M. Sabiston, and C. Wrosch, "Predicting changes in posttraumatic growth and subjective well-being among breast cancer survivors: the role of social support and stress," *Psychooncology.*, vol. 23, no. 1, pp. 114–120, Jan. 2014.

[9] R. J. Schlegel, M. A. Manning, L. A. Molix, A. E. Talley, and B. A. Bettencourt, "Predictors of depressive symptoms among breast cancer patients during the first year post diagnosis," *Psychol. Heal.*, vol. 27, no. 3, pp. 277–293, Mar. 2012.

[10] G. C. Manikis *et al.*, "Computational modeling of psychological resilience trajectories during breast cancer treatment," in *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*, Oct. 2019.

[11] K. Kourou *et al.*, "Computational models for predicting resilience levels of women with breast cancer," in *IFMBE Proceedings*, Sep. 2020, vol. 76, pp. 518–525.

[12] H. Kondylakis, E. Alekos, K. Marias, M. Tsiknakis, and N. Papadakis, "Developing the BOUNCE Psychological Ontology," *ISWC*, 2020.

[13] H. Kondylakis *et al.*, "Developing a data infrastructure for enabling breast cancer women to BOUNCE back," *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2019-June, pp. 652–657, 2019.

[14] S. Singer *et al.*, "Hospital anxiety and depression scale cutoff scores for cancer patients in acute care," *Br. J. Cancer*, vol. 100, no. 6, pp. 908–912, Mar. 2009.

[15] K. Kourou *et al.*, "A machine learning-based pipeline for modeling medical, socio-demographic, lifestyle and self-reported psychological traits as predictors of mental health outcomes after breast cancer diagnosis: An initial effort to define resilience effects," *Comput. Biol. Med.*, vol. 131, p. 104266, Apr. 202.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, 2002.

[17] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972.

[18] T. Brandão, M. S. Schulz, and P. M. Matos, "Psychological adjustment after breast cancer: a systematic review of longitudinal studies," *Psycho-Oncology*, vol. 26, no. 7. John Wiley and Sons Ltd, pp. 917–926, 2017.

[19] A. Syrowatka *et al.*, "Predictors of distress in female breast cancer survivors: a systematic review," *Breast Cancer Research and Treatment*, vol. 165, no. 2. Springer New York LLC, pp. 229–245, Sep. 01, 2017.

[20] G. Pravettoni and A. Gorini, "A P5 cancer medicine approach: Why personalized medicine cannot ignore psychology," *J. Eval. Clin. Pract.*, vol. 17, no. 4, pp. 594–596, Aug. 2011.