JAMIA Open, 2024, **7(3)**, ooae082
https://doi.org/10.1093/jamiaopen/ooae082
Research and Applications

Research and Applications

Using real-world electronic health record data to predict the development of 12 cancer-related symptoms in the context of multimorbidity

Anindita Bandyopadhyay, MS[1], Alaa Albashayreh, PhD, MSHI, RN[2], Nahid Zeinali, MS[3], Weiguo Fan, PhD[1], Stephanie Gilbertson-White, PhD, APRN-BC, FAAN*[,2]

[1]Department of Business Analytics, University of Iowa, Iowa City, IA 52242, United States, [2]College of Nursing, University of Iowa, Iowa City, IA 52242, United States, [3]Department of Informatics, University of Iowa, Iowa City, IA 52242, United States

*Corresponding author: Stephanie Gilbertson-White, PhD, APRN-BC, FAAN, University of Iowa College of Nursing, 50 Newton Road, Iowa City, IA 52242, United States (stephanie-gilbertson-white@uiowa.edu)

Abstract

Objective: This study uses electronic health record (EHR) data to predict 12 common cancer symptoms, assessing the efficacy of machine learning (ML) models in identifying symptom influencers.

Materials and Methods: We analyzed EHR data of 8156 adults diagnosed with cancer who underwent cancer treatment from 2017 to 2020. Structured and unstructured EHR data were sourced from the Enterprise Data Warehouse for Research at the University of Iowa Hospital and Clinics. Several predictive models, including logistic regression, random forest (RF), and XGBoost, were employed to forecast symptom development. The performances of the models were evaluated by F1-score and area under the curve (AUC) on the testing set. The SHapley Additive exPlanations framework was used to interpret these models and identify the predictive risk factors associated with fatigue as an exemplar.

Results: The RF model exhibited superior performance with a macro average AUC of 0.755 and an F1-score of 0.729 in predicting a range of cancer-related symptoms. For instance, the RF model achieved an AUC of 0.954 and an F1-score of 0.914 for pain prediction. Key predictive factors identified included clinical history, cancer characteristics, treatment modalities, and patient demographics depending on the symptom. For example, the odds ratio (OR) for fatigue was significantly influenced by allergy (OR = 2.3, 95% CI: 1.8-2.9) and colitis (OR = 1.9, 95% CI: 1.5-2.4).

Discussion: Our research emphasizes the critical integration of multimorbidity and patient characteristics in modeling cancer symptoms, revealing the considerable influence of chronic conditions beyond cancer itself.

Conclusion: We highlight the potential of ML for predicting cancer symptoms, suggesting a pathway for integrating such models into clinical systems to enhance personalized care and symptom management.

Lay Summary

This research explores electronic health records and machine learning for predicting common symptoms in cancer patients, particularly those experiencing more than one chronic condition. The main objective was to predict factors causing 12 cancer-related symptoms such as pain, fatigue, and anxiety through data from over 8000 patients. The study found that symptoms were driven primarily by chronic conditions, rather than stage or treatment of cancer. These findings suggest that new clinical tools could use such an assessment to provide patients with real-time alerts and very personalized symptom management plans both within and outside of the clinic. This approach potentially could personalize and, in doing so, effect better care for the patient with cancer, thus improving quality of life through the anticipation and management of symptoms.

Key words: clinical notes; electronic health records; natural language processing; machine learning; symptoms.

Background and significance

People diagnosed with cancer often experience physical and emotional symptoms associated with both the disease itself and its treatments, such as pain, fatigue, anxiety, depression, nausea, and vomiting.[1,2] Poorly managed symptoms can lead to decreased quality of life, increased health services utilization, and delays in or early cessation of treatments.[3,4] Proactive symptom management support is critical to comprehensive cancer care.[1,5]

The ability to predict which symptoms will manifest in which patients and at what stage in the disease trajectory is crucial.[6,7] Accurate prediction of symptom development allows for more personalized symptom management care, potentially improving patient outcomes, and optimizing resource allocation.[6–9] Despite this potential, predicting symptoms remains challenging due to the multifactorial nature of the symptom experience, which involves a complex interplay among disease-related, treatment-related, cancer-related, and patient-related factors.[10,11] Specifically, the contribution of multimorbidity (ie, having a diagnosis of 2 more chronic conditions) in the development of cancer symptoms is not well understood.[12]

Received: June 7, 2024; **Revised:** August 9, 2024; **Editorial Decision:** August 12, 2024; **Accepted:** September 5, 2024

Researchers widely regard symptom self-reporting as the gold standard in symptom management research. However, the literature underrepresents the symptom experience of individuals who are too ill, unwilling, or unable to participate and do not meet the inclusion criteria for individual studies.[13] Electronic health records (EHRs) can provide valuable real-world data from large and diverse clinical populations, including those excluded from traditional research, offering a more comprehensive view of symptom development.[14,15] EHR data presents challenges, mainly due to the typical documentation of symptoms in unstructured clinical narratives. However, natural language processing (NLP) techniques[16] have enabled the extraction of symptom information at a large scale.[15,17] NLP can transform unstructured narratives into structured data for various analytic approaches.[18] Machine learning (ML) is a powerful analytical approach that can handle large, complex, and variable EHR data.[19,20] ML algorithms can learn from the patterns in the EHR data, including the NLP-extracted symptoms, and make predictions about future symptom occurrence,[21,22] therefore providing researchers, clinicians, and patients with more precision about developing cancer symptoms.

ML approaches in health science research have grown extensively in recent years due to their capability to learn from complex and high-dimensional data.[21,23] A recent systematic review of ML to predict cancer symptoms demonstrates the rapid growth of literature in this area over the past 5 years.[24] Due to growing access to standardized and curated EHR data, researchers have increasingly employed ML techniques to build predictive models for various clinical tasks such as diagnosis, prognosis, and treatment response predictions.[8,19,21,23] Moreover, studies have demonstrated ML as a potential tool for predicting various symptoms in patients with cancer.[25] These studies have typically considered a range of predictor variables, including patient demographics, cancer characteristics (eg, cancer primary site, stage), treatment-related factors (eg, chemotherapy, radiation, surgery), and comorbidities.[26,27] While these studies have demonstrated promising predictive performance, they often lack clinical interpretability.[28,29] The so-called "black box" nature of many ML models makes it difficult for clinicians to understand the potential physiologic or behavioral mechanisms underlying the predictions made by these models.[30] The opacity of black-box models limits ML models' acceptance and practical applicability in clinical settings.[31] However, the field of ML has witnessed significant advancements, leading to the emergence of methods that offer plausible explanations for model predictions.[32,33] These advancements align with recent actions under the current US administration's comprehensive strategy for responsible AI innovation, aiming to ensure the safe, secure, and trustworthy development and application of AI technologies across various sectors, including healthcare.

One notable method is SHapley Additive exPlanations (SHAP), which has garnered substantial attention.[32,34] SHAP interprets the model's output in terms of its input variables (ie, predictor variables), making it possible to understand each variable's contribution to the final prediction model.[32] SHAP has been widely used for analysis of acute myocardial infarction to nasopharyngeal cancer survival and the risk assessment of lymph node metastasis in papillary thyroid carcinoma cases.[35–37]

Interpretable ML methods hold promise for predicting symptom development in patients with cancer, particularly in the context of multimorbidity. However, the use of SHAP for predicting symptoms in patients with cancer and multimorbidity stays unexplored. This study seeks to address this gap by creating ML models to predict symptoms in patients with cancer and multimorbidity. It utilizes SHAP to elucidate the models' predictions and identify the primary predictors of symptom development.

Objective

This research focuses on developing interpretable ML algorithms tailored for individual patients. Our goal is to predict 12 prevalent symptoms in patients with cancer accounting for the role of other multimorbid diagnoses. The resulting algorithms will improve clinical decision-making tools and inform intelligent recommendation systems for symptom management in patient-centric technologies such as mobile applications.

Materials and methods

Study design and population

This study is a population-based retrospective analysis utilizing EHR sourced from the Enterprise Data Warehouse for Research (EDW4R),[38] the central repository for the University of Iowa Hospitals and Clinics (UIHC). Our cohort included adult patients diagnosed with cancer and at least one other chronic condition who underwent treatment at these facilities from 2017 to 2020. Eligibility criteria for the study required participants to be at least 18 years old and have accessible EHR data at the time of data extraction.

Prediction outcomes

In this study, we focused on predicting 12 symptoms in cancer patients: anxiety, appetite loss, constipation, depressed mood, disturbed sleep, fatigue, impaired memory, nausea/vomiting, pain, pruritus, shortness of breath, and swelling. These symptoms were identified from an earlier analysis that involved 572 626 EHR notes post-cancer diagnosis of these patients using NLP.[39] For this task, we employed NimbleMiner,[40] a sophisticated ML-NLP tool, ensuring the accuracy and reliability of symptom extraction. The tool's performance was confirmed through comparison with 1112 manually annotated EHR notes, yielding a high inter-annotator reliability score (0.924). The precision (0.878), recall (0.876), and F1-score (0.877) of NimbleMiner in identifying these symptoms show its effectiveness. The presence or absence of symptoms, denoted as 1 or 0, respectively, were used as the prediction outcomes of our study.

Study variables

Data collection spanned various EHR domains. Sociodemographic information encompassed *age*, *biological sex* (female or male), *race* (White or non-White), and *marital status* (married or unmarried), with race categorized simply due to the White patient population. Cancer characteristics were extracted from the North American Association of Central Cancer Registries (NAACCR) table in the EW4R and included primary site (digestive organs, breast, urinary, respiratory, and other less frequent sites), cancer stage (*in situ*/localized, regional/distant, and unstaged), and treatment

types (*surgery*, *chemotherapy*, *radiotherapy*, and *hormonal therapy)*. Patients had pre-existing chronic conditions before cancer diagnosis, ensuring symptom data was collected post-diagnosis.

We transformed 918 ICD codes into 60 chronic conditions using the Calderón-Larrañaga classification system, which comprehensively represents patient health complexity and defines chronic conditions as long-term health issues persisting for a year or more.[41,42] We chose this system for its broad overview of multimorbidity, crucial for studying cancer-related symptom development at the person-level rather than just mortality risk.[43] Multimorbidity levels were categorized as low (0-8), moderate (8-13), high (13-19), and very high (19-44), treated as nominal variable with the category "low" as reference. Cancer stage (reference: unstaged) and age group (reference: 0-20 years) variables were also treated this way, while gender, race, marital status, primary site, and treatment types were treated as discrete variables.

ML model development and evaluation

Data preprocessing in the model development phase involved removing records with missing data for the target variables, as there were no missing values for the predictor variables. We also created dummy variables for the categorical study variables. Generating multimorbidity scores and categorizing age into different life stages were done as part of variable engineering. The training of ML models involved using 3 different algorithms—logistic regression (LR), random forest (RF), and XGBoost (XGB)—LR for its interpretability, and RF and XGB for their robustness and superior handling of non-linear relationships.

To address the class imbalance in the symptoms, the Synthetic Minority Over-sampling Technique (SMOTE)[44] was used. Hyperparameter tuning was implemented to optimize model performance. The models, trained on a balanced dataset, underwent 5-fold cross-validation for performance assessment. Model efficacy was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and F1-score, ensuring reliable and accurate symptom prediction in patients with cancer with multimorbidity.

Statistical analysis

To obtain descriptive tables and conduct analytics, we analyzed a total sample of 8156 patients, as detailed in Table 1. This involved calculating the mean and standard deviation of age along with the distribution of *age groups*, *gender*, *race*, *marital status*, *cancer primary site*, *stage*, *treatment modality*, and *multimorbidity levels*. For Table 2, we examined 569 374 EHR notes of the 8156 patients, classifying them by note type, encounter type, and author type. This analysis provided a comprehensive overview of the sample characteristics and the nature of the EHR data used in our study.

Model interpretability

To make our ML models understandable, we used SHAP.[32,34] SHAP values provide a measure of the impact of each patient characteristic on a model's prediction, indicating which variables most significantly influence outcomes. It can also dissect individual predictions to reveal the contribution of each variable to the specific predicted outcome for a single patient. This clarity of variable importance aids clinicians in interpreting model predictions for personalized patient care.

Table 1. Sample characteristics.

	Total sample ($N = 8156$)
Age, mean (standard deviation)	60.5 (14.4)
0-20	28 (0.3%)
20-39	750 (9.1%)
40-59	2718 (33.1%)
60-79	4078 (49.7%)
80+	629 (7.7%)
Gender	
Female	4379 (53.4%)
Male	3777 (46.6%)
Race	
White	7648 (93.2%)
Non-White	555 (6.8%)
Marital status	
Married	4812 (58.7%)
Unmarried	3391 (41.3%)
Cancer primary site	
Digestive organs	1108 (13.5%)
Breast	680 (8.3%)
Urinary	670 (8.2%)
Respiratory	559 (6.8%)
Other sites	5186 (63.2%)
Stage	
In-situ and localized	2188 (26.7%)
Regional and distant	1946 (23.7%)
Un-staged	4069 (49.6%)
Treatment modality	
Surgery	6004 (73.6%)
Chemotherapy	2752 (33.7%)
Radiotherapy	2140 (26.2%)
Hormonal therapy	1177 (14.4%)
Number of chronic conditions	
Low (0-8)	2420 (29.7%)
Moderate (8-13)	2039 (25.0%)
High (13-19)	1785 (21.9%)
Very high (19-44)	1912 (23.4%)

The sample size in this table is based on records with structured data.

Table 2. Electronic health record note characteristics.

	Total sample ($N = 569\,374$)
Note type	
Clinic notes	159 747 (28.1%)
Telephone note	146 580 (25.7%)
Progress note	103 790 (18.2%)
Other	159 257 (28.0%)
Encounter type	
Hospital	247 322 (43.4%)
Clinic	135 561 (23.8%)
Telephone	116 094 (20.4%)
Other	70 397 (12.4%)
Author type	
Physician	244 387 (42.9%)
Nurse	178 840 (31.4%)
Other	146 147 (25.7%)

Percentages may not sum up to 100% due to rounding.

Results

Our study utilized EHR data from 8156 patients receiving at least 2 encounters for cancer treatment at the UIHC between January 2008 and December 2018. The average patient age was approximately 60 years, with a majority being female and White (Table 1). These patients presented various cancer

primary sites, including those of the *digestive organs*, *breast*, *urinary*, and *respiratory systems*, encompassing both *regional/distant* and *in situ/localized* stages of cancer. Treatment modalities ranged from *surgery* and *chemotherapy* to *radiotherapy* and *hormonal therapy*.

The dataset of 569 374 EHR notes from 8156 patients included clinic (28.1%), telephone (25.7%), and progress notes (18.2%), from hospital settings (43.4%). Authored by physicians (42.9%), nurses (31.4%), and other staff (25.7%), the records reflect the collaborative effort in patient care documentation (Table 2).

The prevalence of symptoms among patients is compared in the training ($n = 6524$) and testing ($n = 1632$) groups, as illustrated in Table 3. This comparison highlights the occurrence rates of various symptoms like anxiety, appetite loss, and pain, providing a clear view of their distribution across both datasets.

In our study, as evidenced by the comparative AUC and F-1 score analysis (Table 4) and ROC curve analysis (Figure 1), the RF model consistently demonstrates superior performance in predicting a variety of cancer-related symptoms, showing a slight advantage over the LR and XGB models with its marginally higher AUC and F1-score (highest macro average AUC and F1-score as well). The predictive models demonstrated varying levels of effectiveness for different symptoms. For constipation, the model achieved an AUC of 0.754, indicating good predictive performance, crucial for early intervention in cancer patients with multimorbidity to improve comfort and prevent severe complications like bowel obstruction. Fatigue, with an AUC of 0.781, was effectively predicted, allowing for proactive management through personalized interventions such as nutritional support and exercise programs. The pain prediction model also performed well (AUC = 0.779), emphasizing the importance of prompt pain management to enhance overall well-being and treatment adherence. Anxiety prediction (AUC = 0.762) was effective, highlighting the model's utility in identifying patients needing early psychological support to improve mental health and treatment adherence. Depression, with an AUC of 0.738, underscored the necessity for routine screening and early intervention to mitigate its adverse effects on treatment outcomes and quality of life.

Table 3. Symptom prevalence in training and test patient populations.

	Patients with symptoms (Train) ($n = 6524$)	Patients with symptoms (Test) ($n = 1632$)
Anxiety	5599 (85.8%)	1429 (87.6%)
Appetite loss	2307 (35.4%)	555 (34.0%)
Constipation	3809 (58.4%)	943 (57.8%)
Depressed mood	3376 (51.7%)	869 (53.2%)
Disturbed sleep	1724 (26.4%)	405 (24.8%)
Fatigue	4794 (73.5%)	1222 (74.9%)
Impaired memory	1240 (19.0%)	348 (21.3%)
Nausea	5073 (77.8%)	1254 (76.8%)
Pain	6281 (96.3%)	1570 (96.2%)
Pruritus	2571 (39.4%)	677 (41.5%)
Shortness of breath	5298 (81.2%)	1351 (82.8%)
Swelling	5769 (88.4%)	1454 (89.1%)

Table 4. Performance results of 3 machine learning models (bold values indicate the best performance for each symptom).

	Logistic regression		Random forest		XGBoost	
Symptoms	AUC	F1 score	AUC	F1 score	AUC	F1 score
Anxiety	0.754	0.854	0.762	**0.888**[a]	**0.764**	0.857
Appetite loss	0.760	0.615	**0.769**	0.610	**0.769**	**0.627**
Constipation	**0.748**	0.727	0.744	**0.733**	0.736	0.718
Depressed mood	**0.745**	**0.672**	0.738	0.669	0.741	0.668
Disturbed sleep	0.711	0.480	0.731	0.482	**0.732**	**0.499**
Fatigue	0.786	0.810	0.781	**0.827**	**0.794**	0.820
Impaired memory	0.662	0.380	**0.721**	0.373	0.690	**0.424**
Nausea	**0.749**	0.802	**0.749**	**0.835**	0.744	0.797
Pain	0.765	0.914	**0.779**	**0.954**	0.765	0.929
Pruritus	0.730	**0.620**	0.729	0.598	**0.734**	0.589
Shortness of breath	**0.770**	0.838	0.769	**0.873**	**0.770**	0.836
Swelling	0.780	0.878	**0.783**	**0.912**	0.790	0.888
Macro average	0.747	0.716	**0.755**	**0.729**	0.753	0.720

[a] Note: Bold values indicate the best performance for each symptom across the three models (based on both AUC and F1 score).

Symptom prediction care example: fatigue

We have selected fatigue as the exemplary symptom for the SHAP analysis due to its prevalence and significant impact on the lives of patients with cancer with multimorbidity. The SHAP analysis for all 12 symptoms is detailed in the appendix (Supplementary Appendix A and B).

The SHAP summary plot for fatigue (Figure 2), featuring the top 20 predictors, identifies *allergy*, *colitis and related diseases*, and *peripheral neuropathy* as the foremost predictors of fatigue according to the best RF model. The distribution of SHAP values for most variables shows that the presence of the variable (red dots) spreads widely to the right of the zero line, and the absence of the variable (blue dots) clusters narrowly to the left. This pattern indicates that most variables have a significant and variable influence on increasing the model's prediction when present and a minor, stable influence on decreasing the prediction when absent. The model also underscores the importance of *cancer stages*, particularly noting that *regional/distant* have a greater predictive value for fatigue than *in-situ/localized*. Additionally, *multimorbidity* (ie, the total number of chronic conditions) appears as a critical predictor. The *multimorbidity* category of *very high* (ie, 19-44) predicts fatigue more significantly than *high* (ie, 13-19). None of the cancer treatment modalities is identified as a top predictor of fatigue.

To demonstrate how these results can be used to model predictions at the person level, SHAP force plots were generated for the 12 target symptoms. We calculated the total number of symptoms per patient and subsequently calculated the quartiles for these symptom totals. Two patients were randomly selected from different quartiles for comparison: 1 from the third quartile, exhibiting 10 or more symptoms, and another from the first quartile, with 5 or fewer symptoms. The first patient, representing the high symptom count category, is a married White male aged between 40 and 59, diagnosed with cancer of the male urinary system and possessing a very high multimorbidity index. The second patient, from the low symptom count category, shares a similar demographic profile but has cancer in the mesothelial system and a low multimorbidity count. Despite having different numbers of symptoms, both patients are at the *in-situ*/localized stage of cancer. This makes it easier to study how the number of symptoms affects patients at a similar stage in their illness.[45]

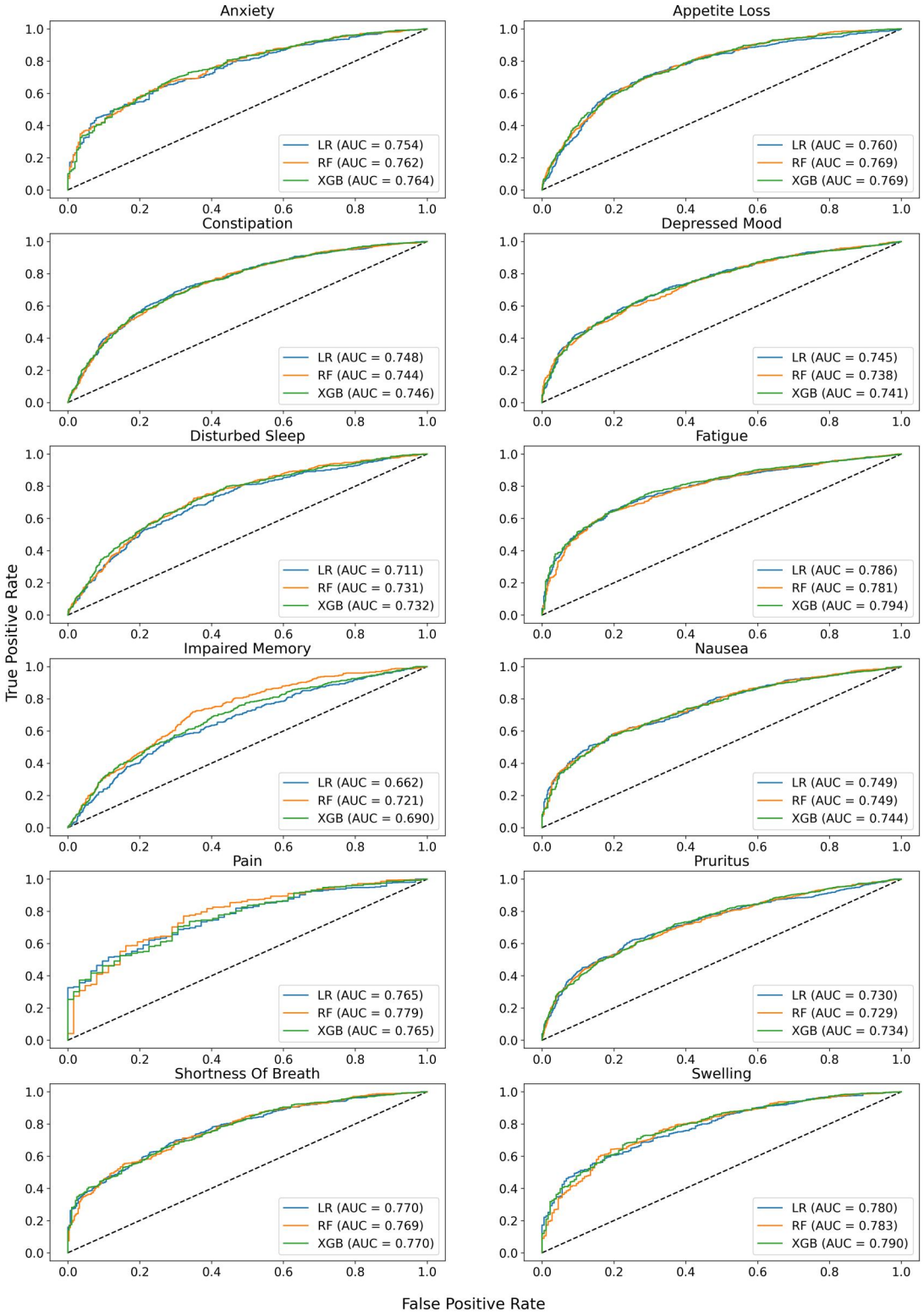

Figure 1. Comparative ROC curves for symptom prediction models. Each panel is a symptom and shows ROC curves for models including logistic regression (LR), random forest (RF), and XGBoost (XGB), with the area under the curve (AUC) metric provided for each.

These patients were randomly selected from their respective symptom prevalence groups.

In the force plots for Patient 1 and Patient 2 (Figure 3A and B, respectively), we see contrasting predictions for fatigue based on their individual demographic and health status. Patient 1 has a high prediction of 0.87 for fatigue, well above the base value of 0.5. This is influenced by conditions like *allergy*, *anemia*, *neurotic stress-related and somatoform diseases*, *very high multimorbidity*, and *other metabolic diseases*, each pushing the prediction upwards. The absence of *peripheral neuropathy* slightly lowers the prediction for patient 1 and does not significantly alter the overall high

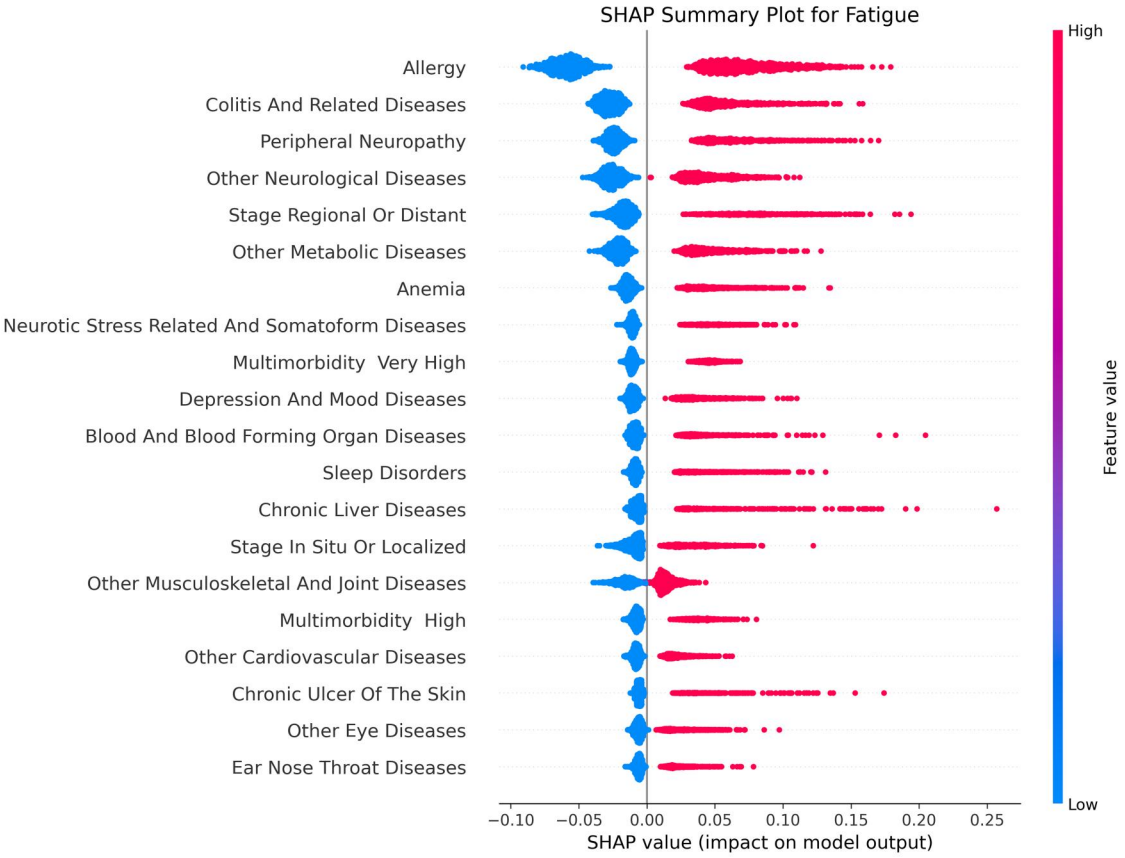

Figure 2. SHAP value summary plot for fatigue prediction using random forest classifier. The *y*-axis represents the variables in the decreasing order of their importance and the *x*-axis shows the SHAP values indicating their impact. Positive SHAP values push predictions towards the positive class, while negative values push towards the negative class. Dot colors are variable value levels—red for high and blue for low. The spread and density of the dots across the plot reflect the variability and frequency of each condition's influence on prediction of fatigue.

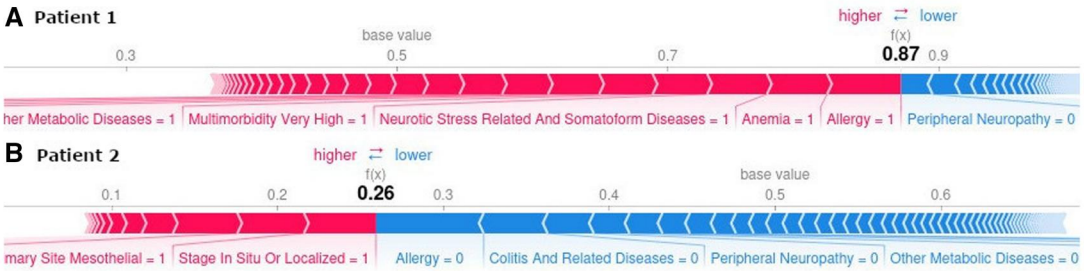

Figure 3. SHAP force plots for patients 1 and 2. Individual impact of variables on prediction of fatigue for patient 1 (A) and patient 2 (B), with color-coded arrows indicating the direction and magnitude of each variable's influence, all converging to shift the prediction from a base value.

likelihood of experiencing fatigue. In contrast, Patient 2's prediction of 0.26 indicates a lower possibility of fatigue, with the absence of *allergy*, *colitis and related diseases*, *peripheral neuropathy*, and *other metabolic diseases* driving the prediction down. The presence of *primary site mesothelial* and *stage in situ/localized* has a minor upward effect but not enough to shift the overall prediction towards a higher likelihood of fatigue.

Discussion

In this study, we successfully built ML models to predict 12 common cancer symptoms and to create person-level algorithms that predict the likelihood of symptom development at the individual level. The recognition of variability in the *primary site*, *stage*, and *treatment of cancer* and pre-existing *multimorbidity* underscores the need for greater precision in predicting which patients will develop which cancer symptoms. This study challenges the current paradigm in cancer symptom management, showing that multimorbidity and patient characteristics influence symptoms more than cancer diagnosis or treatments. Clinicians should consider multimorbidity in their assessments and develop personalized management plans for proactive monitoring and early intervention. Clinical informatics tools can be developed to enhance clinicians' ability to include such a wide array of factors. EHR systems that both capture multimorbidity data and provide predictive tools to identify at-risk patients are needed to implement these goals.

A key strength of our approach is using free-text notes for symptom data, which are not readily available in structured

EHR data. These notes, authored primarily by physicians and nurses, cover a wide range of visit types (in-patient progress notes, outpatient clinic notes, and telephone calls). The range of visit types, capturing diverse interactions where symptoms may be documented. We excluded admission, discharge, and nurses' flowsheet notes, focusing instead on the chronic symptom experience of patients with cancer, which is largely out-patient. Including inpatient notes would skew the data toward acute symptoms linked to hospitalization. Future research should consider health system factors, like care setting, in symptom development.

Implementing interpretable ML algorithms, in this study, a highly demonstrated a targeted approach to symptom prediction among patients with cancer with multimorbidity. Our method, through intentional selection of individual patient characteristics is a move toward precision clinical decision-support for health care providers. This precision is particularly relevant to the 12 symptoms selected for this study. Anxiety, appetite loss, constipation, depressed mood, disturbed sleep, fatigue, impaired memory, nausea/vomiting, pain, pruritus, shortness of breath, and swelling are highly distressing and common across cancer primary sites and stages, and they have significant effect on quality of life.[46–48] Identifying patients who are likely to experience these symptoms provides an opportunity for personalized management strategies that optimize patient care and improve overall well-being. For instance, constipation has an AUC of 0.754, indicating good predictive performance. This means the model can reliably identify cancer patients with multimorbidity who are at risk for constipation. Early identification and intervention can significantly improve comfort and quality of life and prevent severe complications like bowel obstruction. Healthcare providers can proactively manage constipation through dietary adjustments, hydration, and medications, thus reducing the burden of this symptom and enhancing overall patient care.

RF demonstrated a notable performance in predicting a range of symptoms in patients with cancer (Table 4, Figure 1). The superior performance of RF can be attributed to RF's ability to handle variable interactions without extensive variable engineering.[49] Additionally, RF's robustness against overfitting[50] and its capacity to manage missing data and variability common in real-world[51] make it particularly adept for EHR derived datasets. Even in cases where symptom predictions may not be inherently complex, RF maintained comparable performance with other models, indicating its versatility.

The analysis of SHAP summary plots reveals the complex and varied set of risk factors affecting symptom development in this sample, with specific chronic conditions, total multimorbidity, and demographic factors being key (Supplementary Appendix A). Review of the factors present across all 12 symptoms, *allergies*, *mental health disorders*, and *total multimorbidity* are significant predictors, with their presence or absence markedly affecting symptom likelihoods. The protective and risk association of having a diagnosis of *allergy* in the development of all the symptoms is consistent with established research describing the complex interplay between immune response and the development of both physical and emotional symptoms.[52,53] This finding is supported by previous published research describing how multiple symptoms significantly affect patient-reported outcomes, highlighting the importance of other health conditions on the symptom experience of patients with cancer.[46,54]

The impact of conditions like *peripheral neuropathy* and *colitis* in patients with cancer points to a complicated interaction where the effects of conditions not related to cancer play a significant role. The interactions between cancer and other chronic medical conditions make sense anecdotally but it is particularly hard to predict and manage in clinical care resulting in providers basing their management of established clinical guidelines.[45] Our approach of using real-world data and a wide range of potential predictors builds on previous research demonstrating that cancer related symptoms appear and vary depending on the *stage* and *location* of the cancer and the *treatment* methods used.[55–57] Future research is needed to include the potential effect of medications for the management of chronic conditions as well as the presence of these symptoms prior to the diagnosis of cancer. For example, does a diagnosis of chronic pain and use of analgesics for that condition impact the development of pain during cancer treatment?

The force plot analysis for patients 1 and 2 across various symptoms offers a compelling comparison, highlighting the significant impact of chronic conditions, demographic factors, and specific diseases on symptom likelihood (Supplementary Appendix B). These person-level predictions demonstrate the potential clinical utility of this research. Not only can we identify the most salient factors associated with symptom develop across the sample, but we can also derive person-level predictions that reflect the complex relationship among factors that provide either protection or increased risk for developing symptoms. Patient 1's symptom predictions are markedly influenced by the presence of conditions such as *depression and mood diseases*, *neurotic stress-related and somatoform diseases*, *allergies*, and others, consistently elevating the likelihood of experiencing a range of symptoms from anxiety to swelling. The absence of certain conditions occasionally mitigates these predictions, albeit modestly. In contrast, patient 2's analyses reveal a nuanced interplay where the absence of *allergies* and specific diseases results in lower predictive values for symptoms, with some conditions and demographic factors providing mixed effects. The contrast between these 2 patients underscores the individualized nature of symptom development, emphasizing the need for personalized healthcare approaches that account for number and range factors that influence symptom development.

Including multiple chronic conditions in the models revealed their significant role in cancer symptom development, surpassing the impact of *cancer diagnosis*, *stage*, and *treatment*. These results advance cancer symptom science by helping our understanding of individual variability. They also pave the way for clinical prediction models integrated into EHRs, offering real-time alerts and symptom management recommendations to healthcare providers. Similarly, these algorithms can power eHealth interventions, such as mobile apps, to provide patient symptom management directly.

Limitations

Even though the demographics of our sample were representative of Iowa patients with cancer, using data from one institution alone detracts from generalizing our findings. Furthermore, it does not allow drawing any causal inference because of its retrospective design. In this proof-of-concept

study, symptoms were treated as dichotomous outcome variables. Next steps will provide more nuance in the predictions including the temporal nature of symptoms waxing and waning over time. Lastly, this study is also limited by selection bias and general limitations common to EHR data.

Although this study predicted individual symptoms, future studies could predict co-occurring symptoms for a comprehensive outlook. This work applies common, uncomplicated predictive models. Future research is needed that employs more sophisticated methods to determine if stronger results can be achieved. In addition, it is important for integration of self-reported symptoms in EHR data with self-reported symptom records to understand the experience of symptoms comprehensively.

There are several various predictors that could have been included in the modeling, such as medications and medical procedures. Medication data, for example, has the potential to serve as an influence on the development of symptoms. Frequently medications are used to manage symptoms and/or causes side-effects/symptoms of their own. However, it is difficult to determine if the patient has filled a prescription and is taking the medication. Furthermore, the purpose of this research was not to determine a specific list of predictors for each symptom, but rather to demonstrate a methodological proof-of-concept that symptom prediction is complex and that at the individual level factors vary widely on what are risk and protective factors.

Conclusion

Our study leverages EHR to delve into the symptomatology of patients with cancer, revealing a complex interplay between numerous factors influencing symptom development. The results of RF models highlight the significance of total multimorbidity, specific chronic conditions, and demographic characteristics in shaping symptom risks. The RF model's performance and robustness make it ideal for future clinical decision-making. The SHAP analysis, focusing on fatigue as an exemplar, illustrates how non-cancer related factors can emerge as primary influencers in the development of symptoms. Furthermore, the individualized force plot analyses for 2 patients show variability in symptom predictions, emphasizing the necessity and opportunity for highly personalized care. By highlighting the nuanced role of non-cancer-specific conditions in symptom development, our findings suggest that a broad and integrated approach to predicting cancer symptoms is possible, thus paving the way for tailored interventions and enhanced patient support systems. Future work should focus on validating these models across with more racially and ethnically diverse samples and in other health care systems. This study marks a pivotal step towards integrating complex clinical prediction models into healthcare systems, potentially transforming patient care through real-time, data-driven insights.

Author contributions

Anindita Bandyopadhyay (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft). Alaa Albashayreh (Conceptualization, Data curation, Methodology, Project administration, Writing—review and editing), Nahid Zeinali (Writing—review and editing), Weiguo Fan (Conceptualization, Supervision, Writing—review and editing), Stephanie Gilbertson-White—Corresponding Author: (Conceptualization, Project administration, Supervision, Writing—review and editing).

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by the Betty Irene Moore Fellowship for Nurse Leaders and Innovators; College of Nursing, University of Iowa; Center for Advancing Multimorbidity Science (CAMS); NINR (National Institute for Nursing Research) grant number P20 1P20NR018081; Holden Comprehensive Cancer Center, University of Iowa, National Cancer Institute (NCI) grant number P30 P30CA086862; Iowa Health Data Resource (IHDR), University of Iowa; and Institute for Clinical and Translational Science, CTSA University of Iowa grant number UL1TR002537.

Conflicts of interest

The authors declare that there are no competing interests.

Data availability

Due to ethical and privacy concerns, the data supporting this study, which includes electronic medical records created by clinicians and nurses containing personally identifiable health information, cannot be shared publicly. However, the data can be made available to qualified researchers upon reasonable request to the corresponding author.

References

1. Henson LA, Maddocks M, Evans C, et al. Palliative care and the management of common distressing symptoms in advanced cancer: pain, breathlessness, nausea and vomiting, and fatigue. *J Clin Oncol*. 2020;38(9):905-914.
2. Llamas-Ramos I, Alvarado-Omenat JJ, Rodrigo-Reguilón M, et al. Quality of life and side effects management in cancer treatment-a cross sectional study. *Int J Environ Res Public Health*. 2023;20(3):1708.
3. Li B, Mah K, Swami N, et al. Symptom assessment in patients with advanced cancer: are the most severe symptoms the most bothersome? *J Palliat Med*. 2019;22(10):1252-1259.
4. Seow H, Tanuseputro P, Barbera L, et al. Development and validation of a prediction model of poor performance status and severe symptoms over time in cancer patients (PROVIEW+). *Palliat Med*. 2021;35(9):1713-1723.
5. Spathis A, Hatcher H, Booth S, et al. Cancer-related fatigue in adolescents and young adults after cancer treatment: persistent and poorly managed. *J Adolesc Young Adult Oncol*. 2017;6(3):489-493.
6. Vickers AJ. Prediction models in cancer care. *CA Cancer J Clin*. 2011;61(5):315-326.
7. Kazem MA. Predictive models in cancer management: a guide for clinicians. *Surgeon*. 2017;15(2):93-97.
8. Kumar Y, Gupta S, Singla R, et al. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Arch Comput Methods Eng*. 2022;29(4):2043-2070.

9. Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. *Mt Sinai J Med*. 2012;79(6):757-768.
10. Ueno T, Ichikawa D, Shimizu Y, et al. Comorbid insomnia among breast cancer survivors and its prediction using machine learning: a nationwide study in Japan. *Jpn J Clin Oncol*. 2022;52(1):39-46.
11. Li M, Zhang J, Zha Y, et al. A prediction model for xerostomia in locoregionally advanced nasopharyngeal carcinoma patients receiving radical radiotherapy. *BMC Oral Health*. 2022;22 (1):239.
12. Ahmad TA, Gopal DP, Chelala C, Dayem Ullah AZ, Taylor SJ. Multimorbidity in people living with and beyond cancer: a scoping review. *Am J Cancer Res*. 2023;13(9):4346-4365.
13. Byju AS, Mayo K. Medical error in the care of the unrepresented: disclosure and apology for a vulnerable patient population. *J Med Ethics*. 2019;45(12):821-823.
14. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care*. 2010;48(6 Suppl):S106-S113.
15. Wong J, Horwitz MM, Zhou L, et al. Using machine learning to identify health outcomes from electronic health record data. *Curr Epidemiol Rep*. 2018;5(4):331-342.
16. Joshi AK. Natural language processing. *Science*. 1991;253 (5025):1242-1249.
17. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc*. 2019;26(4):364-379.
18. Koleck TA, Tatonetti NP, Bakken S, et al. Identifying symptom information in clinical notes using natural language processing. *Nurs Res*. 2021;70(3):173-183.
19. On J, Park H-A, Yoo S. Development of a prediction models for chemotherapy-induced adverse drug reactions: A retrospective observational study using electronic health records. *Eur J Oncol Nurs*. 2022;56:102066.
20. Zhang D, Yin C, Zeng J, et al. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak*. 2020;20(1):280.
21. Rai HM, Yoo J. A comprehensive analysis of recent advancements in cancer detection using machine learning and deep learning models for improved diagnostics. *J Cancer Res Clin Oncol*. 2023;149 (15):14365-14408.
22. Topaz M, Adams V, Wilson P, Woo K, Ryvicker M. Free-text documentation of dementia symptoms in home healthcare: A natural language processing study. *Gerontol Geriatr Med*. 2020;6:2333721420959861.
23. Abdullah Alfayez A, Kunz H, Grace Lai A. Predicting the risk of cancer in adults using supervised machine learning: a scoping review. *BMJ Open*. 2021;11(9):e047755.
24. Zeinali N, Youn N, Albashayreh A, Fan W, Gilbertson White S. Machine learning approaches to predict symptoms in people with cancer: Systematic review. *JMIR Cancer*. 2024;10:e52322.
25. Kurisu K, Inada S, Maeda I, et al.; Phase-R Delirium Study Group. A decision tree prediction model for a short-term outcome of delirium in patients with advanced cancer receiving pharmacological interventions: a secondary analysis of a multicenter and prospective observational study (Phase-R). *Palliat Support Care*. 2022;20 (2):153-158.
26. Xuyi W, Seow H, Sutradhar R. Artificial neural networks for simultaneously predicting the risk of multiple co-occurring symptoms among patients with cancer. *Cancer Med*. 2021;10 (3):989-998.
27. Xu XY, Lu JL, Xu Q, et al. Risk factors and the utility of three different kinds of prediction models for postoperative fatigue after gastrointestinal tumor surgery. *Support Care Cancer*. 2021;29 (1):203-211.
28. Bratko I. *Machine Learning: Between Accuracy and Interpretability. International Centre for Mechanical Sciences*. Springer Vienna; 1997:163-177.
29. Luo Y, Tseng HH, Cui S, et al. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR Open*. 2019;1(1):20190021.
30. Hakkoum H, Abnane I, Idri ALI. Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*. 2022;117:108391.
31. Hayashi Y. Black box nature of deep learning for digital pathology: beyond quantitative to qualitative algorithmic performances. In: Holzinger A, Goebel R, Mengel M, Müller H, eds. *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*. Cham, Switzerland: Springer International Publishing; 2020:95-101. https://doi.org/10.1007/978-3-030-50402-1_6
32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*. Vol. 30. Red Hook, NY: Curran Associates Inc.; 2017.
33. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York, NY: Association for Computing Machinery; 2016:1135-1144. https://doi.org/10.1145/2939672.2939778
34. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67.
35. Ibrahim L, Mesinovic M, Yang K-W, Eid MA. Explainable prediction of acute myocardial infarction using machine learning and shapley values. *IEEE Access*. 2020;8:210410-210417.
36. Alabi RO, Elmusrati M, Leivo I, et al. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Sci Rep*. 2023;13(1):8984.
37. Zou Y, Shi Y, Sun F, et al. Extreme gradient boosting model to assess risk of Central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual prediction using shapley additive explanations. *Comput Methods Programs Biomed*. 2022;225:107038.
38. Davis HA, Santillan DA, Ortman CE, et al. The Iowa health data resource (IHDR): an innovative framework for transforming the clinical health data ecosystem. *J Am Med Inform Assoc*. 2024;313:720-726.
39. Albashayreh A, Bandyopadhyay A, Zeinali N, Zhang MIN, Fan W, Gilbertson White S. Natural language processing accurately differentiates cancer symptom information in electronic health record narratives. *JCO Clin Cancer Inform*. 2024;8:e2300235.
40. Topaz M, Murga L, Bar-Bachar O, et al. NimbleMiner: an open-source nursing-sensitive natural language processing system based on word embedding. *Comput Inform Nurs*. 2019;37 (11):583-590.
41. Calderón-Larrañaga A, Vetrano DL, Onder G, et al. Assessing and measuring chronic multimorbidity in the older population: a proposal for its operationalization. *J Gerontol A Biomed Sci Med Sci*. 2017;72(10):1417-1423.
42. Ward BW, Schiller JS, Goodman RA. Multiple chronic conditions among US adults: a 2012 update. *Prev Chronic Dis*. 2014;11:E62.
43. Dekhtyar S, Vetrano DL, Marengoni A, et al. Association between speed of multimorbidity accumulation in old age and life experiences: a cohort study. *Am J Epidemiol*. 2019;188(9):1627-1636.
44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *JAIR*. 2002;16:321-357.
45. Koo MM, Swann R, McPhail S, et al. Presenting symptoms of cancer and stage at diagnosis: evidence from a cross-sectional, population-based study. *Lancet Oncol*. 2020;21(1):73-79.
46. Cleeland CS. Symptom burden: multiple symptoms and their impact as patient-reported outcomes. *J Natl Cancer Inst Monographs*. 2007;37(37):16-21.

47. Kirkova J, Davis MP, Walsh D, et al. Cancer symptom assessment instruments: a systematic review. *J Clin Oncol*. 2006;24(9):1459-1473.
48. Cleeland CS, Sloan JA, Group AO; ASCPRO Organizing Group. Assessing the symptoms of cancer using patient-reported outcomes (ASCPRO): searching for standards. *J Pain Symptom Manage*. 2010;39(6):1077-1085.
49. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
50. Biau G, Scornet E. A random forest guided tour. *Test*. 2016;25(2):197-227.
51. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.
52. Lian J, Yue Y, Yu W, et al. Immunosenescence: a key player in cancer development. *J Hematol Oncol*. 2020;13(1):151.
53. Hiam-Galvez KJ, Allen BM, Spitzer MH. Systemic immunity in cancer. *Nat Rev Cancer*. 2021;21(6):345-359.
54. Fowler H, Belot A, Ellis L, et al. Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer*. 2020;20(1):2-5.
55. Alexiusdottir KK, Möller PH, Snaebjornsson P, et al. Association of symptoms of colon cancer patients with tumor location and TNM tumor stage. *Scand J Gastroenterol*. 2012;47(7):795-801.
56. George M, Smith A, Sabesan S, et al. Physical comorbidities and their relationship with cancer treatment and its outcomes in older adult populations: systematic review. *JMIR Cancer*. 2021;7(4):e26425.
57. Enien MA, Ibrahim N, Makar W, et al. Health-related quality of life: Impact of surgery and treatment modality in breast cancer. *J Cancer Res Ther*. 2018;14(5):957-963.

JAMIA Open, 2024, 7, 1–10
https://doi.org/10.1093/jamiaopen/ooae082
Research and Applications