Chapter 5

Applying Natural Language Processing to Patient Messages to Identify Depression Concerns in Cancer Patients

Marieke M. van Buchem, Anne A.H. de Hond, Claudio Fanconi, Vaibhavi Shah, Max Schuessler,
Ilse M.J. Kant, Ewout W. Steyerberg,
Tina Hernandez-Boussard

Journal of the American Medical Informatics Association,
17 July 2024

5.1 Abstract

Objective

This study aims to explore and develop tools for early identification of depression concerns among cancer patients by leveraging the novel data source of messages sent through a secure patient portal.

Materials and Methods

We developed classifiers based on logistic regression (LR), support vector machines (SVMs), and 2 Bidirectional Encoder Representations from Transformers (BERT) models (original and Reddit-pretrained) on 6600 patient messages from a cancer center (2009-2022), annotated by a panel of healthcare professionals. Performance was compared using AUROC scores, and model fairness and explainability were examined. We also examined correlations between model predictions and depression diagnosis and treatment.

Results

BERT and RedditBERT attained AUROC scores of 0.88 and 0.86, respectively, compared to 0.79 for LR and 0.83 for SVM. BERT showed bigger differences in performance across sex, race, and ethnicity than RedditBERT. Patients who sent messages classified as concerning had a higher chance of receiving a depression diagnosis, a prescription for antidepressants, or a referral to the psycho-oncologist. Explanations from BERT and RedditBERT differed, with no clear preference from annotators.

Discussion

We show the potential of BERT and RedditBERT in identifying depression concerns in messages from cancer patients. Performance disparities across demographic groups highlight the need for careful consideration of potential biases. Further research is needed to address biases, evaluate real-world impacts, and ensure responsible integration into clinical settings.

Conclusion

This work represents a significant methodological advancement in the early identification of depression concerns among cancer patients. Our work contributes to a route to reduce clinical burden while enhancing overall patient care, leveraging BERT-based models.

5.2 Background and significance

Depression is common in cancer patients and negatively associated with treatment outcomes, prognosis, and quality of life[1–5]. Despite its prevalence in cancer patients (20% in the United States[6]), depression often remains underdiagnosed. This leads to delayed intervention, poorer treatment adherence, and potential exacerbation of the patient's overall health status[1–3,7–9]. Early identification of depression symptoms may facilitate timely mental health support.

A majority of tools for depression screening in cancer patients utilize structured surveys. Many of these tools perform well in the identification of cancer patients with depression. However, most clinicians do not use structured depression scales during routine clinical care, as they perceive them as too long[10]. To address this, machine learning (ML) approaches using clinical data have also been explored[11]. For example, Cho et al. trained an ML model to predict depression using nationwide clinical check-up data, attaining an AUC of 0.84[12]. While some ML tools demonstrate promising performance, they primarily rely on clinician-generated data, which may not fully capture the patient's perspective or experiences. This limitation highlights the need for alternative data sources that can provide a more comprehensive view of the patient's mental health status.

Patient-generated health data, such as messages sent through secure patient portals, present a unique yet underutilized source of information for identifying signs of depression. These messages, often exchanged between patients and healthcare providers, can provide insights into the patient's mental health status, potentially enabling early detection and treatment of depression symptoms. However, the increasingly high number of patient messages also leads to challenges for healthcare professionals. Previous studies suggest that the high volume of clinical communications can lead to professional exhaustion and burnout[13–15]. Applying natural language processing (NLP) to patient-generated health data might offer a solution by monitoring all incoming messages for potential signs of depression.

Most previous work on applying NLP to identify mental health issues in patient-generated health data is focused on social media data[16–25]. Social media is a valuable source, as it provides extensive documentation of people's personal thoughts, experiences, and ideas. Especially Reddit is an interesting medium as it is fully anonymous, enabling honest conversations between users[26,27]. However, the downside of detecting mental health issues through these kinds of media is that it is difficult to provide support to the individual users. A recent study described the development of an NLP model to predict suicide-related events from patient portal messages, showing promising results comparable to commonly used assessment tools[28]. Therefore, it might be possible to assist healthcare professionals in identifying cancer patients that potentially suffer from depression.

Objective

This study aims to develop a proof-of-concept model using NLP to analyze incoming patient messages and identify those messages indicative of depression concerns. We compare classical machine learning approaches with neural networks-based Bidirectional Encoder Representations from Transformers (BERT) models[29] and investigate whether domain-adaptive pretraining on Reddit data improves the performance of the model.

5.3 Methods

Data

The dataset consisted of patient-initiated messages that were sent through a secure patient portal by patients from a comprehensive cancer cohort, containing all patients who visited the Stanford Cancer Center from 2009 until 2022. The secure patient portal allows patients to send messages to their care team. We only included English, patient-initiated messaging threads and excluded all standard communication, e.g. reminders for appointments and invitations for patient satisfaction questionnaires. We did not exclude patients with a prior depression diagnosis, as this has previously been found to be the most predictive factor for developing depression[30,31]. We aimed for a final sample size of at least 5000 manually annotated messages, based on similar studies leveraging BERT for binary text classification on manually annotated data[32–34]. For the final annotation sample, we randomly sampled 50% of messages from the dataset. To decrease the imbalance in the sample towards non-concerning messages, we selected the other 50% of the annotation sample to contain potentially alarming or concerning content. To this end, an experienced social worker and data scientist created two lists of words that signaled concern or alarm respectively (see Appendix A). The selected sample included all the messages containing alarming words, supplemented with randomly sampled messages containing concerning words. This distribution was chosen to maximize the number of potentially concerning messages within the annotation sample. Additionally, special attention was given to discerning depression from anxiety. Our word lists included terms related to both depression and anxiety to ensure these messages were manually annotated, enhancing the model's capability to distinguish messages expressing anxiety from potential depression.

Ethical Considerations

This study was approved by the Stanford institutional review board (#47644). Informed consent was waived for this retrospective study for access to personally identifiable health information as it would not be reasonable, feasible, or practical. The data are housed in the Stanford Nero Computing Platform, which is a highly secure, fully integrated internal research data platform meeting all security standards for high risk and protected health information data. The security is managed and monitored, and the platform is updated and adapted to meet regulatory changes.

Annotation process

The definition of when a message was 'concerning for depression' was created through multi-stakeholder input (see Appendix B). We organized brainstorm sessions with oncologists, data scientists, medical students, and a social worker, during which we iteratively worked towards a definition and annotation guideline that everyone agreed on (see Appendix B). When consensus was reached, the annotation was performed by seven healthcare professionals. We created a diverse group of annotators, in terms of clinical experience, age, gender, race, and ethnicity. The final set of 6,600 annotated messages was used as the reference standard. Of 6,600 messages, a set of 100 messages was annotated by all annotators. We computed the inter annotator agreement on this sample using Krippendorf's Alpha. We used majority vote to define the reference standard for these 100 messages. See Figure 1 for an overview of the annotation process.

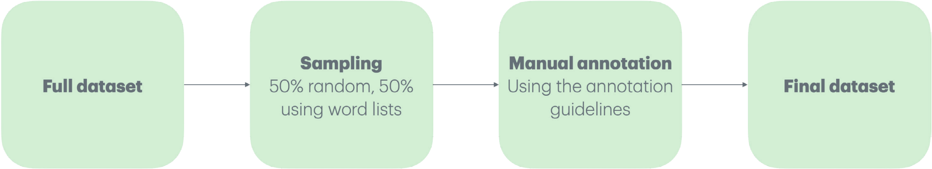

Full dataset
Sampling
50% random, 50% using word lists
Manual annotation
Using the annotation guidelines
Final dataset

Figure 1: overview of the sampling and annotation pipeline.

Models

For this study, we aimed to compare the performance of two baseline machine learning models (logistic regression [LR] and support vector machines [SVM]) to two bidirectional encoder representations from transformers (BERT) models: a naïve BERT base model and a BERT base model that has undergone continued domain-adaptive pretraining on Reddit data. The method of continued domain-adaptive pretraining is computationally less expensive than a full pretraining task, while it has shown to improve the performance on specific tasks[35–40]. We chose the BERT base in combination with continued pretraining on Reddit data, as opposed to ClinicalBERT[41] or BioBERT[36], because Reddit data includes a wide range of discussions, including those related to personal experiences and mental health, which are closer to the type of conversational and informal language found in patient portal messages. Patient portal messages often reflect patients' everyday language and concerns, which may not be captured in more formal medical records or biomedical literature. This similarity in language and context can help the model better understand and interpret patient messages. We specifically used the Depression subreddit to include the type of language that people use to talk about depression.

For the LR and SVM models, we first preprocessed the data by changing all letters to lowercase, removing stop words, and stemming the words. We used the term frequency – inverse document frequency (TF-IDF) to extract features from the data. The final dataset was split into 70-15-15 train-validation-test sets. To find the best hyperparameters for the TF-IDF vectorizer, the LR and the SVM, we performed a grid search using the train and validation set. We then fit the LR and the SVM model with the best hyperparameters and performed a bootstrap with 1000 samples using the test set to determine the performance of the models.

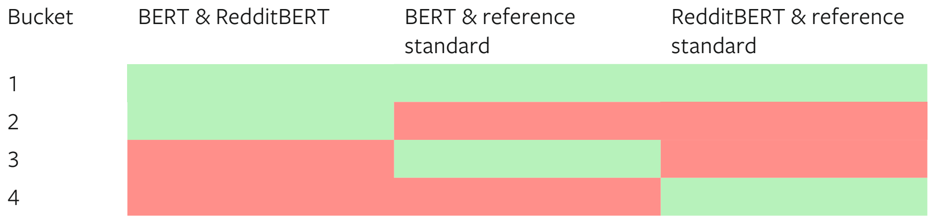
For our domain-adaptive pretraining task, we chose a specific Reddit community ('subreddit') called 'r/Depression'. This is a large subreddit with more than 900,000 members that has been in use since 2009 and, therefore, provides extensive data on a large time span. It is the biggest subreddit focused on depression, with millions of posts. From this subreddit, we scraped 1,000,000 posts. We then continued pretraining the BERT base model for 20 epochs [35]. The pretraining was performed using four GPUs on the Google Cloud Platform. The final model is referred to as RedditBERT. Both BERT base and RedditBERT were finetuned on the binary classification task of identifying concerning messages, using the annotated sample of patient messages.

Associations between model predictions and patient characteristics

We conducted a comparative analysis of patient characteristics and clinical outcomes between patients who sent messages deemed concerning by RedditBERT, and those who did not. Included outcomes were a depression diagnosis, prescriptions for depression medication, and mental health referrals (Appendix C). Differences in categorical variables were assessed using a chi-square test. For continuous variables, an independent samples T-Test was performed. In cases where continuous variables encompassed more than two groups, a one-way Analysis of Variance (ANOVA) was performed.

Explainability

We used Local Interpretable Model-Agnostic Explanations (LIME) to compare BERT and RedditBERT explanations on a patient message level[42]. LIME provides the local importance of each word to the model's classification of a specific patient message. Our sample included 32 texts categorized into four distinct buckets based on the outputs of the two BERT models (BERT and RedditBERT) and their alignment with the reference standard (human annotation) (see Figure 2). Subsequently, our seven annotators were asked to compare the predictions generated by the two models and the corresponding LIME explanations. Through a structured survey, annotators were asked to indicate which prediction they agreed with and which explanation they preferred (Appendix D). Additionally, the survey provided an opportunity for the annotators to explain their decisions.



Bucket	BERT & RedditBERT	BERT & reference standard	RedditBERT & reference standard
1	🟩	🟩	🟩
2	🟩	🟥	🟥
3	🟥	🟩	🟥
4	🟥	🟥	🟩

Figure 2: description of the four buckets used for evaluation of the explainability. An empty cell indicates agreement, a diagonal line indicates disagreement.

5.4 Results

Patient characteristics

The total data set included 6,600 messages from 3,312 unique patients. The cohort consisted of more females (60%), an average age of 61 years old and a majority of White and Asian, privately insured patients (see Table 1 for more characteristics). Our final test set consisted of 907 messages (14% of the total labeled set) from 760 unique patients. 90 messages (10%) were annotated as concerning for depression.

Table 1. Patient characteristics cohort.

	N=3,312
Demographics	
Female sex, N(%)	2,002 (60)
Age, mean (std)	61.3 (13.7)
English speaking, N(%)	3,080 (93)
Race	
Asian (%)	725 (22)
Black (%)	65 (2)
White (%)	2,133 (64)
Other (%)	364 (11)
Ethnicity	
Hispanic/Latino (%)	204 (6)
Non-hispanic/non-latino (%)	3,060 (92)
Other (%)	47 (1)
Depression diagnosis, N(%)	1,116 (34)
Insurance (%)	
Private	1,980 (60)
Medicare	550 (17)
Medicaid	260 (8)
Other	522 (16)

Inter-annotator agreement

The inter annotator agreement (IAA) calculated over all seven annotators was 0.38 according to Krippendorf's Alpha, which can be considered moderate. We observed a large variation in IAA between different sets of annotators, ranging from 0.32 to 0.52, depending on which annotator was removed from the set.

Table 2. Performance metrics of four models classifying patient messages as concerning for depression.

Metric, mean [95% CI] based on 1000 bootstraps	Log Reg *Threshold: 0.2**	SVM *Threshold: 0.5**	BERT *Threshold: 0.5**	RedditBERT *Threshold: 0.5**
AUROC	0.79 [0.74-0.83]	0.83 [0.78-0.87]	0.86 [0.82-0.90]	0.88 [0.85-0.91]
Precision	0.32 [0.25-0.39]	0.36 [0.28-0.44]	0.37 [0.30-0.44]	0.33 [0.26-0.39]
Recall	0.51 [0.40-0.61]	0.60 [0.49-0.70]	0.68 [0.59-0.78]	0.74 [0.66-0.84]
F1-score	0.39 [0.31-0.47]	0.45 [0.37-0.52]	0.48 [0.40-0.55]	0.46 [0.39-0.53]

** = threshold chosen that led to the highest F1 score*

Table 3. Predictive performance per subgroup.

	BERT AUC [95% CI]	Recall [95% CI]	RedditBERT AUC [95% CI]	Recall [95% CI]
Overall	0.86 [0.82-0.90]	0.69 [0.59-0.78]	0.88 [0.85-0.91]	0.74 [0.65-0.83]
Sex				
Female (*n=476*)	0.85 [0.79-0.91]	0.73 [0.61-0.84]	0.88 [0.84-0.92]	0.73 [0.61-0.84]
Male (*n=284*) 0.89 [0.83-0.94] 0.62 [0.43-0.78] 0.90 [0.84-0.94] 0.76 [0.62-0.90]				
Race				
Asian (*n=136*)	0.87 [0.74-0.98]	0.75 [0.53-0.95]	0.91 [0.83-0.97]	0.63 [0.37-0.86]
Black (*n=16*)	0.82 [N/A]	0.33 [N/A]	0.75 [N/A]	0.33 [N/A]
White (*n=519*)	0.86 [0.81-0.90]	0.67 [0.54-0.79]	0.88 [0.85-0.92]	0.79 [0.68-0.89]
Other (*n=81*)	0.83 [0.64-0.98]	0.74 [0.44-1.00]	0.90 [0.79-0.98]	0.75 [0.44-1.00]
Ethnicity				
Hispanic/Latino (*n=44*)	0.80 [0.54-0.99]	0.66 [0.33-1.00]	0.88 [0.73-0.99]	0.78 [0.44-1.00]
Non-Hispanic/non-Latino (*n=709*)	0.87 [0.83-0.91]	0.69 [0.58-0.79]	0.89 [0.86-0.92]	0.75 [0.66-0.84]
Other (*n=7*)	0.95 [N/A]	0.67 [N/A]	0.95 [N/A]	0.33 [N/A]

Model performance

The TF-IDF parameters and hyperparameters of the logistic regression (LR) and support vector machine (SVM) can be found in Appendix E. The LR model had a mean area under the ROC curve (AUROC) of 0.79 (95% confidence interval (CI): 0.74-0.83) while the SVM attained an AUROC of 0.83 (95% CI: 0.78-0.87).

Both BERT and RedditBERT were trained and validated for 5 epochs on 5,693 labeled messages. See Appendix E for hyperparameters. Both models outperformed the LR and SVM and RedditBERT slightly outperformed BERT, with a mean AUROC of 0.88 (95% CI: 0.85-0.91) versus 0.86 (95% CI: 0.82-0.90), respectively. A threshold of 0.5 led to the highest F1 score for the BERT models and the SVM. For the LR, a threshold of 0.2 led to the highest F1 score (Table 2). In total, RedditBERT labeled 200 messages as concerning (22%). When comparing the predictive performance per subgroup, BERT showed bigger differences in performance across sex, race, and ethnicity than RedditBERT. For both models there was a decreased performance for Black patients (Table 3).

Associations between model predictions and patient characteristics

There was a significant difference in race in the classification of patients' messages. Messages of White patients were more often classified as concerning, while messages of Asian patients were less often classified as concerning (see Appendix F). Furthermore, patients on Medicaid or Medicare also sent more messages classified as concerning. Patients who sent messages classified as concerning by RedditBERT had a higher chance of receiving a depression diagnosis, a prescription for antidepressants, or a mental health referral within the next 3, 6, and 12 months after sending the concerning message. Patients sending a concerning message were also more likely to already have a depression diagnosis, a prescription for antidepressants, or a mental health referral (see Appendix F).

Explainability

The explanation of which words contributed to the prediction per message differed for BERT and RedditBERT, with RedditBERT highlighting more words than BERT (see Appendix G). Annotators preferred BERT's explanation to RedditBERT's explanations for 14 out of 26 texts (54%). Annotators often opted for RedditBERT's explanation when it highlighted words or sentences that BERT missed. On the other hand, annotators sometimes preferred BERT's explanation because they found RedditBERT highlighted words that did not make sense in the eyes of the annotators. Furthermore, several annotators mentioned that the words highlighted as 'not concerning' did not always seem to make sense to them (Table 4).

Table 4. Annotators' reasons for choosing BERT or RedditBERT's explanation.

Reasons for choosing RedditBERT	Reasons for choosing BERT
"Difficult. I like the explanations of [RedditBERT] a bit more, because it seems to pick out more complete sentences like ' am extremely tired' and 'have not ... able to sleep more'."	"I prefer [BERT] because the blue [non-concerning] words in [RedditBERT], do not make sense to me. Why should testosterone be marked as non-concerning."
"This is the best use case of this model. A clear cry for help. I like the explanations of [RedditBERT] better because it picks out more complete sentences 'I'm pretty depressed' and has a stronger reaction on the 'psychologist'."	"[RedditBERT] highlights a lot of text that I do not think relevant in either direction."
"I think in general it is good [RedditBERT] picks up on prescription names."	"Again, there is a lot of text highlighted in both that does not really make sense to me. [BERT] highlights less text."
"I like how [RedditBERT] picks up on the 'love to talk to somebody'. "	"I do not agree with the extra highlighted words really in [RedditBERT], as the only indication of concern is the 'depressed'."

5.5 Discussion

In this study, we demonstrate a proof-of-concept for leveraging patient-generated health data for the early identification of depression concerns in cancer patients. By employing natural language processing (NLP) techniques, specifically Bidirectional Encoder Representations from Transformers (BERT) and domain-adaptive pretraining using Reddit data (RedditBERT), we highlight the potential of artificial intelligence in enhancing mental health surveillance. However, the performance disparities observed across patient subgroups, notably concerning race and ethnicity, necessitate a careful consideration of the ethical implications and potential biases introduced by these models.

The good discriminatory ability across all models showed the potential of patient messages as a valuable source for depression risk stratification. Our results are comparable to one other study that used patient portal messages to identify a mental health event, namely suicide [28]. For this study, the authors reported an AUROC of 0.71. Both findings underline the potential of using patient messages as a unique data source, which provides a current snapshot of how a patient is feeling and directly represents the patient's voice. This untapped data source has the opportunity to improve personalized, proactive identification of mental health issues. However, more research on this topic is needed, as these are the only studies describing the application of NLP on this data source.

There was no significant difference between the naïve base BERT and the domain-adaptive pretrained RedditBERT model. This finding contradicts previous studies in which domain-specific models like BioBERT (pre-trained on biomedical texts) and ClinicalBERT (pretrained on clinical texts), and continuously pretrained BERT models outperformed base BERT [35,36,41,43]. However, within the mental health domain, a recent study also found that depression classification did not improve significantly with continued pretraining[44]. More research is needed to assess the value of social media data for continued pretraining in the mental health domain.

We found a notable difference in how words were weighed in the explanations provided for BERT and RedditBERT, but there was no difference on average in the quality of the explanations. Explanations may help generate trust in deep neural network models, like BERT, which are inherently uninterpretable[45]. Yet, post-hoc explainability methods like LIME are difficult to validate, and their effect on clinical decision making is still unknown. More research is needed on the added value of explainability methods in increasing trust versus the potential to harm trust[46]. Alternatively, neural networks with a more inherent interpretability mechanism could lead to better explanations[47].

The subgroup analysis showed slight differences in performance between sex, race, and ethnicity. Compared to BERT, RedditBERT performed more consistently between subgroups and had a slightly better recall for male patients and White patients, which could be due to Reddit being predominantly used by males[48]. Furthermore, both models performed worse on Black patients, which can be explained by the low number of Black patients within our sample. This finding highlights the importance of addressing potential biases and ethical considerations associated with deploying AI models in healthcare, emphasizing the need for equitable and unbiased implementations. The National Institute of Health's All of Us Research Program is a great example of an initiative aiming to collect data from a diverse group of participants across the US[49]. For future research, we recommend training models on such a diverse dataset to decrease differences in subgroup performance.

A depression diagnosis or prescription of depression medication occurred more often after a concerning message was sent compared to after a non-concerning message was sent. This suggests that our models were able to identify messages that were truly indicative of depression concerns. These may be patients that could benefit from additional mental healthcare outreach. Important to note, however, is that some of these patients already received a depression diagnosis or treatment. This highlights the classification capabilities of the model, although the model might not perform well for prediction. This assumption is underlined by a recent study, where we show that using the output of our model does not improve the performance of a prediction model for depression[31].

Limitations

One limitation is the moderate inter-annotator agreement. This can be attributed to the diversity among the annotators and the inherent subjective interpretation of what qualifies as a 'concerning' message in patient emails. This is highlighted by the large variation in IAA, depending on which annotators are included. Although we believe the IAA could be improved by excluding some annotators, it also mirrors the real-world where different healthcare professionals may interpret patient communications differently. Relevant literature describing similar use cases, such as annotating Twitter data for mental health symptoms, report similar moderate inter-annotator agreements[50,51]. Despite the moderate agreement, the study's rigorous approach in involving multiple annotators and the alignment with existing literature provide valuable insights into the complexities of labeling subjective content. Taking this into account, we conclude that the use of patient messages combined with labels from our diverse, clinical group of annotators greatly improved the method's potential to be applicable in healthcare practice.

Furthermore, our current approach to upsampling concerning messages may lead to a bias in the training data towards messages that are more easily identifiable as concerning. As the current study is a proof-of-concept, we chose this method to keep the manual annotation feasible while still ensuring that there were enough concerning messages to train the model effectively. However, future work should explore more sophisticated sampling techniques to better represent the full spectrum of patient messages and minimize potential biases.

Another limitation of this study is the focus on a single institution, which may limit the generalizability of our results to other settings. Especially the high number of privately insured patients is not representative for the general population. Nevertheless, this cohort included a diverse population in terms of race and ethnicity, with a substantial percentage of Hispanic and Asian patients. This study can thus be seen as a proof-of-concept and sets the stage for future investigations into the ways different ethnic and cultural groups express mental health concerns in their communications. By recognizing and addressing these differences, subsequent studies can delve deeper into tailoring interventions that resonate effectively across diverse patient populations.

Lastly, the study's framework might not capture patients who do not use emails for communication or are hesitant to reach out, thereby potentially missing a subset of the population in need.

Future implications

Given our exploration in the use of advanced NLP models for the identification of depression concerns in cancer patients, the broader implications for healthcare are significant. The advantage of BERT and RedditBERT over traditional methods underscores the potential of integrating more sophisticated language models into clinical practice. With the ongoing advancements in NLP, especially in the field of large language models (LLM's), there is the potential to further refine these models, making them even more relevant and effective in a clinical context. Future work should focus on comparing several newer language models to determine if they could provide improved performance in identifying depression. Recent studies have shown that it is also possible to use LLMs to create chatbots for counseling, offering another promising avenue for providing mental health support[52]. However, while the promise of these advanced NLP models in healthcare is evident, it's crucial to approach their integration with caution. Before such models can be responsibly incorporated into clinical settings, additional research is required to address potential biases as were demonstrated in the current study and evaluate the real-world impact on physician-patient interaction and clinical outcomes[53–55].

Furthermore, our study significantly contributes to the literature by emphasizing the underutilized potential of patient-generated health data, specifically messages sent through a secure patient portal. This novel approach taps into valuable information exchanged between patients and healthcare providers, offering insights into the mental health state of a patient and enabling early detection of depression concerns. This data is systematically collected, as opposed to, for example, patient reported outcomes (PROs). The collection of PROs is often burdensome to patients and healthcare providers, may not capture all patients' concerns, and rely on patients' memory to report symptoms that have occurred prior to the patient's visit.18 Thus, patient messages should be seen as a valuable additional data source for clinical research and surveillance.

Following our proof-of-concept study, we propose several next steps. First of all, to ensure broader applicability of such a tool, the training dataset should be extended with data representative of the general population. Secondly, it is important to conduct a temporal validation to assess the model's performance over time. Lastly, other types of explainability methods should be tested to determine if some provide a better understanding of the model's behavior than the current method. These steps will help refine the model further and enhance its applicability and trustworthiness in a clinical setting.

Conclusion

In conclusion, this work represents a significant methodological advancement in the early identification of depression concerns among cancer patients, addressing a critical gap in patient care. Our work contributes to a route to reduce clinical burden while enhancing overall patient care, leveraging BERT-based models. Further research is needed to address biases, evaluate real-world impacts, and ensure responsible integration into clinical settings. As the study highlights, the interpretability of these models is paramount for clinician trust and responsible implementation in healthcare settings, particularly for vulnerable patient populations.

Author contributions

Marieke M. van Buchem, Anne A.H. de Hond, Ewout W. Steyerberg, Ilse M.J. Kant, and Tina Hernandez-Boussard were responsible for the conceptualization and design of the study. Marieke M. van Buchem and Anne A.H. de Hond performed the data extraction. Marieke M. van Buchem performed the data analysis. Max Schuessler and Vaibhavi Shah provided clinical advice. Marieke M. van Buchem drafted the original manuscript. All authors had full access to all the data, critically analyzed, reviewed, contributed, and approved the final manuscript.

Funding

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR003142 and the National Library Of Medicine of the National Institutes of Health under Award Number R01LM013362. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.v.B. and A.d.H. received a travel grant from the Cath-

arine van Tussenbroek Fund and the Prins Bernhard Cultuur Fund to support this research.

Conflicts of interest

The authors have no competing interests to disclose.

Data availability

The data underlying this article cannot be shared due to the sensitivity of the content and the privacy of individuals that participated in the study.

Supplementary information

References

1. Linden W, Vodermaier A, MacKenzie R, *et al*. Anxiety and depression after cancer diagnosis: Prevalence rates by cancer type, gender, and age. *J Affect Disord*. 2012;141:343–51.
2. SMITH HR. Depression in cancer patients: Pathogenesis, implications and treatment (Review). *Oncol Lett*. 2015;9:1509–14.
3. Pitman A, Suleman S, Hyde N, *et al*. Depression and anxiety in patients with cancer. *BMJ*. 2018;361:k1415.
4. Colleoni M, Mandala M, Peruzzotti G, *et al*. Depression and degree of acceptance of adjuvant cytotoxic drugs. *Lancet*. 2000;356:1326–7.
5. Grassi L, Indelli M, Marzola M, *et al*. Depressive symptoms and quality of life in home-care-assisted cancer patients. *J Pain Symptom Manag*. 1996;12:300–7.
6. HHS SA and MHSA (SAMHSA). Substance Abuse and Mental Health Services Administration; mental health and substance abuse emergency response criteria. Interim final rule. *Fed Regist*. 2001;66:51873–80.
7. Walker J, Hansen CH, Martin P, *et al*. Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: a cross-sectional analysis of routinely collected clinical data. *Lancet Psychiatry*. 2014;1:343–50.
8. Caruso R, Breitbart W. Mental health care in oncology. Contemporary perspective on the psychosocial burden of cancer and evidence-based interventions. *Epidemiology Psychiatr Sci*. 2020;29:e86.
9. Mitchell AJ, Chan M, Bhatti H, *et al*. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. *Lancet Oncol*. 2011;12:160–74.
10. Mitchell AJ, Meader N, Davies E, *et al*. Meta-analysis of screening and case finding tools for depression in cancer: Evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. *J Affect Disord*. 2012;140:149–60.
11. Iyortsuun NK, Kim S-H, Jhon M, *et al*. A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare*. 2023;11:285.
12. Cho S-E, Geem ZW, Na K-S. Prediction of depression among medical check-ups of 433,190 patients: A nationwide population-based study. *Psychiatry Res*. 2020;293:113474.
13. Tai-Seale M, Dillon EC, Yang Y, *et al*. Physicians' Well-Being Linked To In-Basket Messages Generated By Algorithms In Electronic Health Records. *Heal Aff*. 2019;38:1073–8.
14. Adler-Milstein J, Zhao W, Willard-Grace R, *et al*. Electronic health records and burnout: Time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. *J Am Med Inform Assoc*. 2020;27:531–8.
15. Lieu TA, Altschuler A, Weiner JZ, *et al*. Primary Care Physicians' Experiences With and Strategies for Managing Electronic Messages. *JAMA Netw Open*. 2019;2:e1918287.
16. Arachchige IAN, Sandanapitchai P, Weerasinghe R. Investigating Machine Learning & Natural Language Processing Techniques Applied for Predicting Depression Disorder from Online Support Forums: A Systematic Literature Review. *Information*. 2021;12:444.

17. Tejaswini V, Babu KS, Sahoo B. Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model. *ACM Trans Asian Low-Resour Lang Inf Process*. 2022;23(1):1-20.

18. Katchapakirin K, Wongpatikaseree K, Yomaboot P, *et al.* Facebook Social Media for Depression Detection in the Thai Community. *2018 15th Int Jt Conf Comput Sci Softw Eng (JCSSE)*. IEEE; 2018;00:1–6.

19. Asad NA, Pranto MdAM, Afreen S, *et al.* Depression Detection by Analyzing Social Media Posts of User. *2019 IEEE Int Conf Signal Process, Inf, Commun Syst (SPICSCON)*. IEEE; 2019;00:13–7.

20. Kabir MK, Islam M, Kabir ANB, *et al.* Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques. *JMIR Form Res*. 2022;6:e36118.

21. Dessai S, Usgaonkar SS. Depression Detection on Social Media Using Text Mining. *2022 3rd Int Conf Emerg Technol (INCET)*. IEEE; 2022;00:1–4.

22. Haque A, Reddi V, Giallanza T. Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction. In*: Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks*. Springer International Publishing; 2021:436-447.

23. Ren L, Lin H, Xu B, *et al.* Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation. *JMIR Med Inform*. 2021;9:e28754.

24. Podina IR, Bucur A-M, Todea D, *et al.* Mental health at different stages of cancer survival: a natural language processing study of Reddit posts. *Front Psychol*. 2023;14:1150227.

25. Chen Z, Yang R, Fu S, *et al.* Detecting Reddit Users with Depression Using a Hybrid Neural Network. In*: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE; 2023:193-199.

26. Choudhury MD, De S. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *ICWSM*. 2014;8:71–80.

27. Ammari T, Schoenebeck S, Romero D. Self-declared Throwaway Accounts on Reddit: How Platform Affordances and Shared Norms enable Parenting Disclosure and Support. *Proc ACM Hum-Comput Interact*. 2019;3:1–30.

28. Bhandarkar AR, Arya N, Lin KK, *et al.* Building a Natural Language Processing Artificial Intelligence to Predict Suicide-Related Events Based on Patient Portal Message Data. *Mayo Clin Proc: Digit Heal*. 2023;1:510–8.

29. Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics; 2019:4171-4186.

30. Riedl D, Schüßler G. Factors associated with and risk factors for depression in cancer patients – A systematic literature review. *Transl Oncol*. 2022;16:101328.

31. Hond A de, Buchem M van, Fanconi C, *et al.* Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study. *JMIR Med Inform*. 2024;12:e51925.

32. Sousa MG de, Sakiyama K, Rodrigues L de S, *et al*. BERT for Stock Market Sentiment Analysis. In*: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence* (*ICTAI*). IEEE; 2019:1597-1601.
33. Du J, Xiang Y, Sankaranarayanapillai M, *et al*. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. *J Am Med Inform Assoc*. 2021;28:1393–400.
34. Zhou S, Wang N, Wang L, *et al*. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *J Am Med Inform Assoc*. 2022;29:1208–16.
35. Lamproudis A, Henriksson A, Dalianis H. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In*: Proceedings of the International Conference on Recent Advances in Natural Language Processing—Deep Learning for Natural Language Processing Methods and Application*. INCOMA, Ltd.; 2021:790-797.
36. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
37. Gururangan S, Marasović A, Swayamdipta S, *et al*. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In*: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:8342-8360.
38. Alsentzer E, Murphy JR, Boag W, *et al*. Publicly Available Clinical BERT Embeddings. In*: Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78.
39. Chakrabarty T, Hidey C, McKeown K. IMHO Fine-Tuning Improves Claim Detection. In*: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*). Association for Computational Linguistics; 2019:558-563.
40. Fanconi C, Buchem M van, Hernandez-Boussard T. Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes. *AMIA Jt Summits Transl Sci Proc*. 2023:138-147.
41. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Arxiv. 2020. Accessed July 12, 2024. https://arxiv.org/abs/1904.05342.
42. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In*: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics; 2016:97-101.
43. Peng B, Chersoni E, Hsu Y-Y, *et al*. Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks. In*: Proceedings of the Third Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics; 2021:37-44.
44. Ji S, Zhang T, Ansari L, *et al*. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In*: Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resources Association; 2022:7184-7190.
45. Amann J, Vetter D, Blomberg SN, *et al*. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Heal*. 2022;1:e0000016.

46. Wysocki O, Davies JK, Vigo M, *et al*. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif Intell*. 2023;316:103839.

47. Fanconi C, Vandenhirtz M, Husmann S, *et al*. This Reads Like That: Deep Learning for Interpretable Natural Language Processing. In*: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2023:14067-14076.

48. Reddit.com. Advertising—Audience—Reddit. Discover what makes Reddit ads unique. Accessed January 17, 2021. https://web.archive.org/web/202101

49. Investigators A of URP. The "All of Us" Research Program. *N Engl J Med*. 2019;381:668–76.

50. Homan C, Johar R, Liu T, *et al*. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In*: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics; 2014:107-117.

51. Mowery D, Bryan C, Conway M. Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data. In*: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics; 2015:89-98.

52. Lai T, Shi Y, Du Z, *et al*. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics*. 2023;4:8–33.

53. Nashwan AJ, Abujaber AA, Choudry H. Embracing the future of physician-patient communication: GPT-4 in gastroenterology. *Gastroenterol Endosc*. 2023;1:132–5.

54. Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med*. 2023;29:1930–40.

55. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*. 2023;90:104512.