

Article

Digital-Reported Outcome from Medical Notes of Schizophrenia and Bipolar Patients Using Hierarchical BERT

Rezaul K. Khandker ¹, Md Rakibul Islam Prince ², Farid Chekani ¹, Paul Richard Dexter ^{3,4}, Malaz A. Boustani ^{3,4} and Zina Ben Miled ^{2,4,*}

¹ Merck & Co., Inc., 126 E Lincoln Ave, Rahway, NJ 07065, USA

² Department of Electrical and Computer Engineering, Indiana University—Purdue University Indianapolis, 723 W. Michigan St., Indianapolis, IN 46202, USA

³ Indiana University School of Medicine, 340 W 10th St., Indianapolis, IN 46202, USA; mboustan@iu.edu (M.A.B.)

⁴ Regenstrief Institute, Inc., 1101 W. 10th Street, Indianapolis, IN 46202, USA

* Correspondence: zmiled@iupui.edu

Abstract: Patient-reported (PRO) and clinician-reported (CRO) outcomes are assessment instruments that are completed by patients and trained healthcare professionals, respectively. A PRO is a report of the direct experience of the patient with a given disease condition. A CRO is an assessment of the condition of the patient by the healthcare provider. PROs may not be accessible to all patients, especially those suffering from severe disease conditions. CROs are time-consuming and therefore administered infrequently. In the present study, we introduce a new form of assessment, the digital-reported outcome (DRO), which is automatically derived from the medical notes of the patient. DROs have a low overhead and can be generated at each patient's visit to complement other outcome-assessment instruments and enhance clinical decision support by identifying at-risk patients. In this study, a DRO is developed to evaluate the functional impairment in the daily activities of two cohorts of patients suffering from bipolar disorder and schizophrenia. The input of the DRO is a single medical note from the electronic medical record of the patient. This note is submitted to a hierarchical bidirectional encoder representations from transformers (BERT) model. First, a sentence-level embedding is produced for each sentence in the note using a token-level attention mechanism. Second, an embedding for the entire note is constructed using a sentence-level attention mechanism. Third, the final embedding is classified using a feed-forward neural network. The model is trained to classify patients into moderate or severe functioning impairment levels according to the general assessment of functioning (GAF) scale, a CRO instrument for the assessment of the impact of mental illness on the daily activities of the patient. The DRO is validated using medical notes that were labeled by multiple healthcare providers from different healthcare institutions. The results indicate that a general DRO is able to classify patients from the two cohorts according to the two functioning impairment levels (severe versus moderate) prior to the onset of disease with an AUC of 76%. Disease-specific DROs are only applicable after the onset of the disease and produced AUCs of nearly 85%. The methodology introduced in the present paper is practical and can support the automated monitoring of the severity of the functioning impairment of bipolar and schizophrenia patients. Extending the proposed DRO to other psychiatric conditions and types of impairments is the subject of ongoing research work.



Citation: Khandker, R.K.; Prince, M.R.I.; Chekani, F.; Dexter, P.R.; Boustani, M.A.; Ben Miled, Z. Digital-Reported Outcome from Medical Notes of Schizophrenia and Bipolar Patients Using Hierarchical BERT. *Information* **2023**, *14*, 471. <https://doi.org/10.3390/info14090471>

Academic Editor: Evaggelos Karvounis

Received: 2 June 2023

Revised: 1 August 2023

Accepted: 17 August 2023

Published: 22 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: mental health; impairment; machine learning; BERT; hierarchical attention; schizophrenia; bipolar disorder

1. Introduction

Schizophrenia and bipolar disorder affect millions of people worldwide [1]. Years of evidence-based medicine have increased our understanding of risk factors and optimal

treatments for mental diseases. However, these diseases have a high burden due to their clinical heterogeneity, variances in severity, and progression paths [2,3]. This presents a unique opportunity for the use of machine learning (ML) in mental health to aid in the diagnosis, treatment, and monitoring of patients at risk. For instance, an ML model was shown to enhance suicide prevention by using data from multiple sources such as IoT devices and social media to identify patterns in the data associated with suicidal behavior [4]. In other studies, natural language processing (NLP) techniques were used to understand psychopathology [5] and to detect mood changes from social media data [6].

The aim of the present study is to develop a low-maintenance, low-overhead digital instrument (DRO) for the assessment of the severity of functioning impairment in patients with mental illness. Two different types of instruments are currently being used for this purpose: patient-reported outcomes (PRO) and clinician-reported outcomes (CRO). Example PROs include the patient health questionnaire (PHQ) [7] and the young mania rating scale (YMRS) [8]. The PHQ went through several revisions in order to reduce the time necessary for its completion. This PRO was initially a revision of another instrument called the primary care evaluation of mental disorders. Subsequently, it was further reduced into PHQ-9, a PRO focusing on depression [9]. YMRS is a PRO that consists of 11 questions, thereby making it easy to administer [8].

While PROs are completed by the patients or their caregivers, CROs are administered by healthcare providers and are time-consuming. The positive and negative syndrome scale (PANSS) [10] is a CRO that consists of 30 questions covering three components: a positive, a negative, and a general component. The general assessment of functioning (GAF) [11] is an assessment of the psychological, social, and occupational functioning of the patient. This CRO is a revised version of the global assessment scale [12], which was initially introduced in 1970. Healthcare providers assign a severity score to the patient ranging from 0 to 100, where higher scores represent superior functioning in daily activities. The GAF score is primarily derived from information in the patient's medical record or information gathered during the patient's encounter with the healthcare provider.

The GAF was widely used. It was included in the Diagnostic and Statistical Manual of Mental Disorders (DSM) version IV. It was then replaced by the WHO Disability Assessment Schedule 2.0 (WHODAS 2.0) [13]. This latter CRO consists of 36 questions addressing cognition, mobility, self-care, getting along, life activities, and participation [14]. While the WHODAS 2.0 is more recent, the present study used medical notes that were annotated according to the GAF score. Because this CRO was administered over an extended period of time, it provides a large number of data that were annotated by multiple healthcare providers, which are needed for the development and validation of the proposed DRO. Ideally, the DRO must learn from the available data to assess the functional impairment of the patients with a level of accuracy comparable to that produced by health providers.

2. Related Work

The purpose of the proposed DRO is to classify each medical note according to the functional impairment level of the patient suffering from bipolar disorder or schizophrenia. Medical notes consist of text data, and their processing relies on advanced NLP techniques. Specifically, large language models have emerged as an enabler for various applications that require the processing of text data from various sources. Language models leverage the concept of transfer learning. They are pre-trained on a large corpus of text data and subsequently fine-tuned for various downstream applications as needed. These models rely on a self-supervised training methodology [15] coupled with a transformer architecture [16] and were shown to facilitate the extraction of efficient contextualized features from text data. Self-supervised training eliminates the need for labeled data, thereby allowing language models to be trained with large corpora. Language models start by tokenizing the text following a pre-established vocabulary. The size of the vocabulary is a trade-off between fewer out-of-vocabulary words and higher model complexity. The encoder–decoder architecture and self-attention mechanism of the transformer architecture enable the capture

of complex dependencies from text data. This mechanism was initially introduced for machine translation in [17]. The encoder–decoder transformer architecture was extended by allowing the encoder to identify important input keywords for each target keyword produced by the decoder (i.e., self-attention). Vaswani et al. [16] subsequently proposed the first model based entirely on attention by replacing the recurrent layers common in previous encoder–decoder architectures with multi-headed self-attention.

The present paper investigates three language models: BERT [18], BERT mini [19], and clinical BERT [20]. As mentioned above, self-attention in language models attempts to learn the dependencies between word pairs. The attention range and the size of the input text have a direct impact on the computational complexity of the model. For instance, the BERT [18] and Longformer [21] language models can process at most 512 and 4096 tokens, respectively, at a time. Moreover, BERT was trained on a general English corpus from sources such as Wikipedia. This language model is large and produces embeddings with dimension 768. BERT mini is a distilled version of BERT which generates embeddings of size 256. A reduced embedding size has the benefit of requiring a smaller dataset for fine-tuning the model to the downstream application. Clinical BERT is a BERT language model that was fine-tuned with emergency medical notes, making it more suitable for medical applications.

The capabilities of the language models have been demonstrated for various applications, including question-answering [22] and text summarization [23]. These example applications focus on general corpora from Wikipedia and news abstracts, respectively. In the medical domain, language models were used for multi-task information extraction from medical notes [24]. They were also used for feature extraction from Twitter posts to identify patients at risk of developing depression [25]. In [26], the BERT language model was used to detect incoherence among sentences transcribed from interviews of schizophrenia patients. The language models BERT and clinical BERT were also successfully used to extract phenotypes related to major depressive disorder (MDD) from medical notes [27].

Despite the above-mentioned successful applications, only a small fraction (around 6% according to [28]) of research studies use medical notes. Medical notes pose a major challenge to pre-trained language models: the input token limit can be restrictive [29], especially if the attention range among related tokens is scattered throughout the long text of the medical note [30]. To overcome this constraint, some studies either truncate the input text or rely on topic modeling in order to focus text processing to a reduced set of selected keywords [31,32]. These strategies are able to address the input limit constraint of language models. However, they can overlook crucial information or are unable to leverage relevant context embodied in dispersed keywords in the medical notes. Therefore, a strategy that can accommodate the entire medical note is needed.

In the present study, we propose to adapt the hierarchical attention network (HAN) initially proposed in [33] for the development of the proposed DRO. HAN is a language model architecture that can accommodate long input text by progressively building the embedding representation of the text. HAN starts with the GloVe [34] embeddings for the tokens in the vocabulary. The size of the GloVe vocabulary is small, and the embeddings are non-contextual. Therefore, in the present study, we replace the GloVe embeddings with BERT embeddings, a language model that considers bi-directional context and benefits from a larger vocabulary. This is especially beneficial for the semantic disambiguation of homonyms and negations prevalent in medical notes. Moreover, as mentioned above, a larger vocabulary has more coverage, with fewer out-of-vocabulary words. The hierarchical architecture was also explored in a recent study [35] where the target application consisted of automatically assigning the proper international disease codes (ICD) [36] to medical notes. In this case, a single attention mechanism was applied at the sentence level and the sentence embeddings were then combined using average pooling. In contrast, the hierarchical mechanism used in the proposed DRO promotes a dual-level attention mechanism at both the token and sentence levels.

The methodology and application introduced in the present study extend the use of machine learning models to further improve the quality of care for patients suffering from bipolar disorder and schizophrenia. These models apply the transformer-based architecture to the entire medical note by using a hierarchical attention mechanism that can overcome the input size limit of traditional language models.

3. Methods

The aim of the present study is to develop a DRO that can reliably identify bipolar and schizophrenia patients with severe functioning impairment in their daily activity from their medical notes. The DRO consists of two stages. First, an embedding of the medical note using the hierarchical attention mechanism is produced. Second, this embedding is submitted to a neural network that classifies the functioning impairment of the patient into two classes: moderate or severe. The study cohort, the architecture of the classifier, training, and evaluation methodologies, is described next.

3.1. Study Cohort

The patient data used in the present study were obtained from the Indiana Network for Patient Care (INPC) over the period from 1 January 2005 to 31 December 2019. INPC is an operational community-wide electronic health network that currently includes data from 19 hospitals in seven healthcare systems, the Marion County Health Department, and various physician practices.

Two cohorts of psychiatric patients are considered:

- **Schizophrenia:** This cohort includes patients with at least two clinical visits with schizophrenia disease codes and anti-psychotic medication use for at least 3 continuous months.
- **Bipolar:** The bipolar cohort consists of patients with at least two clinical visits with bipolar type I or mixed bipolar disease codes and anti-psychotic medication use for at least 3 continuous months.

It is possible for patients to be assigned to the schizophrenia and bipolar cohorts if they satisfy both selection criteria.

Patients were included in the study if they met the selection criteria for one of the cohorts. The index date for the patients in both cohorts is defined as the first date of diagnosis. Additionally, patients had to have at least one interaction with INPC every year during the study period and be 18 years or older on the index date. This criterion made sure that patients included in the study are adults and active users of INPC. Patients are excluded if they belong to a protected group (prisoners, patients living in nursing homes, etc.) or have schizophrenia or bipolar diagnoses during the incubation period (i.e., from 1 January 2005 to 31 December 2005). The latter exclusion criterion was enforced to ensure that each patient had at least one year of medical data prior to the index date.

3.2. Data Annotation and Preprocessing

Annotating sufficient medical notes to enable the training and validation of the proposed DRO is time-consuming and costly. To overcome this challenge, the present study relies on medical notes that were scored according to the GAF scale by multiple healthcare providers from several institutions affiliated with INPC. These medical notes are semi-structured and follow the DSM-IV format. The structured section consists of multiple axes where Axis I includes information on clinical disorders such as major psychiatric disorders and Axis V includes the GAF score for the specific visit. This section is followed by an unstructured free-text section used by healthcare providers to document the status of the patient. The structured section is used to determine the label (i.e., impairment level) of each data sample, and the unstructured section is the input of the proposed DRO.

The GAF impairment scale consists of ten levels where the lower and higher levels describe patients with poor and good functioning, respectively [37]. One important advantage of the GAF scale is that a score is assigned to each medical note. Most other instruments

(e.g., PHQ, PANSS, and YMRS) are the result of a survey questionnaire. Therefore, there is no direct relation between the medical notes in the patient record and the assigned score. Despite the benefits of using an already annotated large dataset for the development of the DRO, there are two issues that needed to be addressed. First, when each medical note is considered independently, the assigned GAF score can be subjective. This issue is mitigated by the large number of medical notes in each cohort and the diverse pool of healthcare providers that authored and scored these notes. Second, the distribution of the medical notes across the 10 levels of the GAF score is not uniform. Therefore, the original GAF scale was mapped to a modified scale with only two levels:

- The severe impairment level corresponds to a GAF score of less than 50. As per the GAF scale, scores in this range are assigned when the patient is in danger of severely hurting himself or others; exhibits delusions; experiences hallucinations; or shows major impairment in several areas, such as work, school, family relations, judgment, thinking, or mood.
- The moderate impairment level corresponds to a GAF score above 51. This GAF range is characterized by the absence or presence of moderate symptoms such as occasional panic attacks, depressed mood, mild insomnia, and difficulty concentrating.

Multi-class classifiers require training datasets proportional in size to the number of classes with sufficient samples in each class. The revised binary functioning impairment scale for daily activities (i.e., severe versus moderate) overcomes this issue while still identifying the most vulnerable patients from the bipolar and schizophrenia cohorts. Similar mappings of the GAF scale were previously used to reduce inter-rater variability and to mitigate the non-uniform distribution of the GAF scores [12].

As mentioned above, the GAF score, which is reported in the structured section of the medical note, is used to obtain the appropriate functioning impairment label needed to train the proposed DRO. The unstructured section is representative of medical notes that are entered by healthcare providers during an encounter with the patient. This free-text section is preprocessed and used as input to the DRO. During preprocessing, XML tags are removed, the text is forced to lowercase, and non-alphanumeric characters are replaced with spaces. Moreover, all the hex, Unicode symbols, and numbers are removed while preserving sentence delimiters. Sentence delimiters are used to split the text into sentences. Sentences with more than 30 words are split into sentences with at most 25 words, thus ensuring that the number of tokens associated with each sentence is within the imposed limits of pre-trained language models. Finally, common stop-words are removed using the NLTK standard English stop-word collection [38], and the mention of drug names is replaced by their respective anatomical therapeutic chemical (ATC) [39,40] group name.

3.3. Hierarchical Attention Model

The proposed DRO follows the HAN architecture [33] and consists of three components: (1) the token encoder [41], (2) the sentence encoder, and (3) the prediction layer (Figure 1). The token encoder generates the sentence-level embedding using the BERT, BERT mini, or clinical BERT bi-directional gated recurrent unit (GRU) token encoder. The original HAN architecture used GloVe embeddings. In the present study, BERT embeddings are used because they are context-sensitive.

Each medical note consists of n sentences, and each sentence $i \in [1, n]$ is a sequence of L^i tokens T_j^i where $j \in [1, L^i]$. The forward GRU reads the sentence i starting from token T_1^i to $T_{L^i}^i$, and the backward GRU reads the sentence starting from token $T_{L^i}^i$ down to T_1^i . An annotation h_j^i for the token T_j^i is obtained by concatenating the hidden layers of the forward and backward GRU encoders. An attention mechanism is then used to identify the tokens that are relevant to the impairment class. The aggregate representation of these tokens constitutes the overall sentence embedding s_i as shown in Equation (1).

$$s_i = \sum_j \alpha_j^i h_j^i \tag{1}$$

where $\alpha_j^i = Attention(h_j^i, u_j) = Softmax(Tanh(W_j h_j^i + b_j)^T u_j)$, W_j is a weight matrix, b_j is a bias vector, and u_j is a context vector. To generate the sentence embedding s_i , the annotation h_j^i of the token T_j^i is first processed by a linear layer with a *Tanh* activation function. This layer is followed by a *Softmax* layer after multiplication by the context vector u_j . The output of the *Softmax* layer, α_j^i , represents the normalized attention weights.

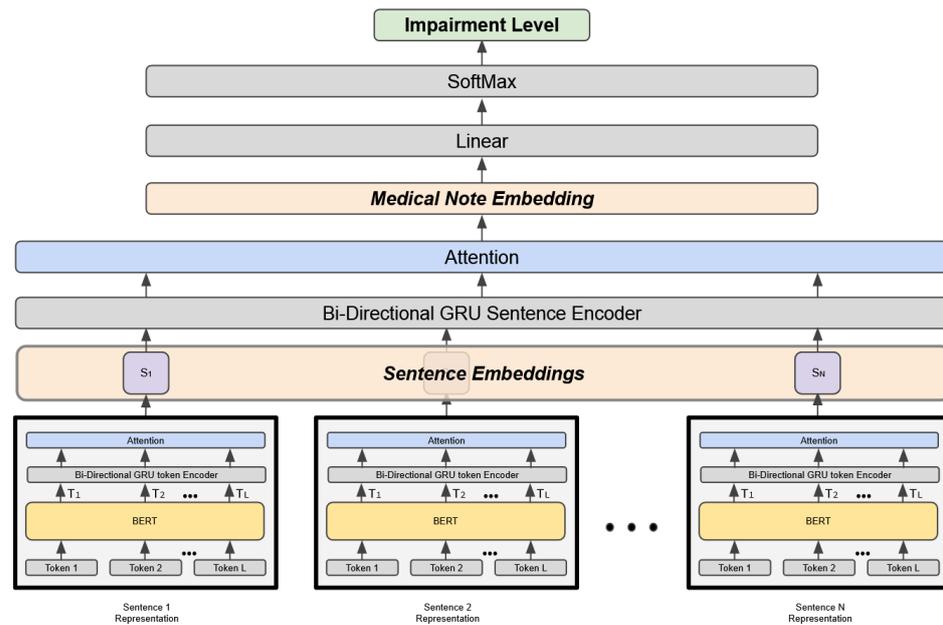


Figure 1. Hierarchical attention architecture of the DRO model.

The second component of HAN shown in Figure 1 is the sentence encoder. This component follows the same approach as the token encoder to produce the embedding of the entire medical note. However, instead of starting from the token embeddings, it uses the sentence embeddings produced by the token encoder. The third component is the prediction layer. This layer uses the output of the sentence encoder to classify the medical note according to the appropriate functioning impairment level.

The above architecture is used to develop two types of DROs. The first type, general DRO (GDRO), is trained using the medical notes from the two cohorts prior to the index date. Three language models are evaluated for the token-level embeddings: BERT, BERT mini, and clinical BERT. The second type is a disease-specific DRO. To produce each disease-specific DRO, GDRO is fine-tuned with medical notes from each of the cohorts after the index date. The two disease-specific DROs are labeled SDRO and BDRO for the schizophrenia and bipolar disease conditions, respectively.

4. Results

The number of patients in the bipolar and schizophrenia cohorts is 1746 and 1767, respectively. These patients can have one or more medical note over the study period. Table 1 shows the distribution of these notes across the two impairment levels before and after the index date. For the two cohorts, the percentage of notes assigned to the severe functioning impairment level is higher than that of those assigned to the moderate functioning impairment level. Moreover, for both the bipolar and schizophrenia cohorts, there is a significant increase ($\approx 20\%$) in the percentage of notes assigned to the severe impairment level and consequently a significant decrease in the percentage of notes assigned to the

moderate impairment level during the post-index period compared to the pre-index period. This trend indicates a deterioration in the impairment level of the patients post-diagnosis.

Table 1. Distribution of the medical notes across the two impairment levels before and after the patient's index date.

	Pre-Index Period		Post-Index Period	
	Severe	Moderate	Severe	Moderate
Bipolar	2324 (45%)	2803 (55%)	3408 (69%)	1548 (31%)
Schizophrenia	3387 (56%)	2612 (44%)	5363 (78%)	1487 (22%)

The number of samples from each cohort used to train and test the DRO models is shown in Table 2. The notes available for the development of GDRO are from the pre-index period, whereas the notes available for the fine-tuning of the disease-specific DROs (i.e., BDRO and SDRO) are from the post-index period. To compare the different DROs and adjust for class imbalance, an equal number of notes was randomly selected from the two impairment levels followed by a 67/33 training/testing dataset split from each impairment level. This two-third/one-third split ratio was selected to ensure sufficient testing data from each impairment level. For GDRO, 2324 notes are selected from each impairment level and cohort over the pre-index period. This number is dictated by the severe impairment level class of the bipolar cohort, which has the lowest number of samples over the pre-index period (Table 1). A total of 9296 samples with an equal number of samples from the two impairment levels and the two cohorts are collected for the training and testing of GDRO (Table 2). Similarly, 1487 notes were selected from the two impairment levels for the fine-tuning and evaluation of each disease-specific DRO. This number was dictated by the moderate impairment level of the schizophrenia cohort during the post-index period.

Table 2. Number of severe and moderate medical notes in the training and testing datasets of each DRO model.

Model	Training		Testing	
	Severe	Moderate	Severe	Moderate
GDRO	3100	3100	1548	1548
BDRO/SDRO	991	991	496	496

4.1. Impairment Classifiers

The DRO models are trained using the Adam optimizer [42] over seven epochs with a batch size of 16 and a learning rate of 5×10^{-5} . The values of the hyper-parameters are established using a three-fold nested cross-validation. Moreover, three language models for token-level embeddings are evaluated: BERT, BERT mini, and clinical BERT.

The mean and standard deviation of the AUC, accuracy, sensitivity, and specificity of each DRO are included in Table 3. These performance metrics were obtained using a three-fold cross-validation. The *Softmax* function in the prediction layer (Figure 1) produces a probability score indicating the likelihood that a medical note belongs to either the severe or moderate impairment level. The AUC is derived from this probability and is a measure of the area under the receiver operating characteristic curve (ROC). This curve is a plot of the sensitivity of the classifier against (1—specificity) and represents the ability of the classifier to discriminate between the two impairment levels as the discrimination threshold is varied. The ROC curves for all DRO models developed using the BERT, BERT mini, and clinical BERT embeddings are shown in Figure 2. The accuracy, sensitivity, and specificity metrics are calculated based on the binary assignment derived by using a specific threshold. The values of the metrics shown in Table 3 are obtained using a threshold of 0.5. Varying the threshold can increase either the sensitivity or the specificity and decrease the other.

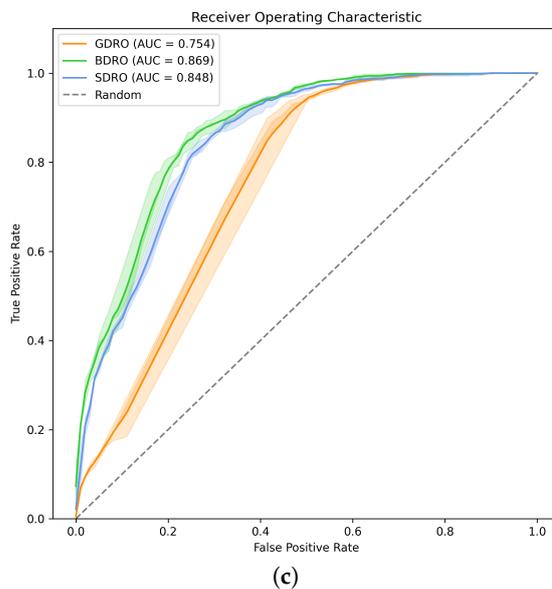
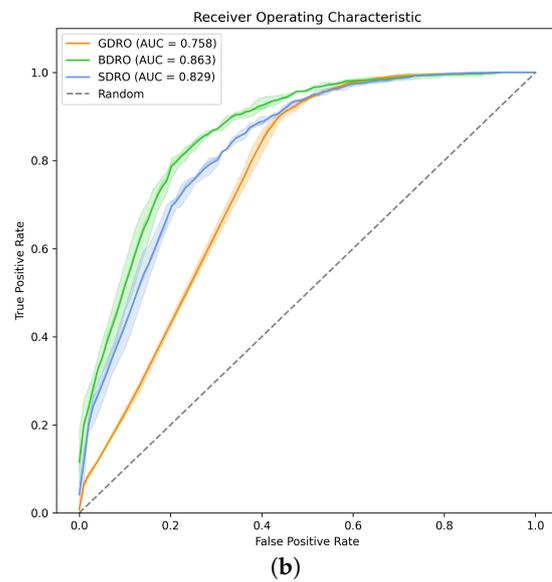
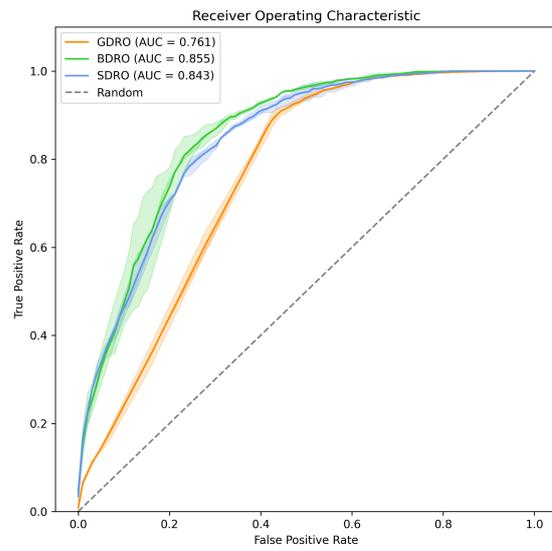


Figure 2. AUCROC curves for all the proposed DROs. (a) DROs with BERT embeddings, (b) DROs with BERT mini embeddings, and (c) DROs with clinical BERT embeddings.

Table 3. Three-fold cross-validation means (standard deviation) of the AUC, accuracy, sensitivity, and specificity of the DRO models with the BERT, BERT mini, and clinical BERT embeddings.

	Model	AUC	Accuracy	Sensitivity	Specificity
BERT	GDRO	76.09 (0.44)	72.97 (0.16)	55.48 (1.61)	90.42 (1.84)
	BDRO	85.57 (1.33)	77.76 (1.06)	78.55 (3.52)	77.00 (4.61)
	SDRO	84.36 (0.33)	75.38 (0.42)	60.37 (2.67)	90.50 (1.83)
BERT mini	GDRO	75.81 (0.35)	72.73 (0.55)	56.60 (3.41)	88.83 (2.51)
	BDRO	86.29 (0.63)	79.42 (0.64)	77.48 (2.70)	81.36 (1.42)
	SDRO	82.89 (0.55)	74.83 (0.54)	73.38 (4.94)	76.31 (4.44)
clinical BERT	GDRO	75.28 (1.75)	72.80 (1.15)	53.47 (4.51)	92.08 (2.35)
	BDRO	86.92 (0.76)	79.66 (0.45)	75.59 (2.38)	83.77 (2.48)
	SDRO	84.84 (0.14)	78.00 (0.43)	72.73 (1.07)	83.31 (1.77)

The general, pre-index GDRO has an AUC greater than 75% with all three embeddings (Figure 2). BDRO and SDRO, the disease-specific DROs, have higher AUCs, with the AUC of BDRO exceeding 85%. For most of the models, the standard deviation of the AUC is less than 1%, indicating that the models are stable. Moreover, the variation in AUC due to the use of different token-level embeddings is less than 2%. According to [43], an AUC in the range of 70% to 80% is considered acceptable for diagnostic applications, and a range of 80% to 90% is considered excellent.

Despite the lack of variability in AUC resulting from different token embeddings, Table 3 shows that there could be significant variances in sensitivity and specificity. For instance, the disease-specific SDRO for schizophrenia shows lower sensitivity and higher specificity with the token embedding BERT compared to BERT mini and clinical BERT. This observation indicates that, in practice, a calibration process is needed to select the appropriate cut-off threshold for each token embedding prior to deploying the DRO in production. This threshold will allow healthcare systems to adjust the true positive and true negative rates of the model to the desired referral rates for patients with severe impairment.

4.2. Attention Mechanism

In order to investigate which sentences the DRO model is attending to in the medical notes, the weights from the attention layers are visualized in Figures 3 and 4 for samples taken from the bipolar severe and moderate impairment levels, respectively. This heatmap shows the combination of the token-level and the sentence-level attentions $\sqrt{\alpha_j^i \alpha^i}$ after normalization. The terms α_j^i and α^i represent the token-level (Equation (1)) and the sentence-level attentions, respectively. The combination of the two attentions scales-up with important tokens in important sentences.

said feeling safe home . said feeling stressed . said also recently lost job . said started cut . hospital course upon admission psychiatric unit patient reported chronic feelings emptiness chronic suicidal ideations . psychoeducation provided . coping skills reviewed . restarted medications . informed consent obtained regarding medications . denied major adverse effects medications . cognitive behavioral therapy provided . hospitalization progressed reported good improvement mood . became socially interactive . participated groups . said want state hospital . said considering going group home . time discharge reporting positive future oriented thoughts . denying feeling hopeless worthless . treated uti antibiotic . denying feeling hopeless worthless . mental status time discharge alert oriented time place person situation . cooperative . mood stated fair . affect reactive . active suicidal homicidal ideations . delusions . hallucinations . thought process linear organized . insight judgment limited . discharge medications . antianxiety agents . antihistamines daily . fluoroquinolones . anticonvulsants bedtime . hypnotics sedatives sleep disorder agents bedtime . etonogestrel levonorgestrel tablet daily . nasal agents systemic and topical spray nostrils twice day . vistaril . antipsychotics antimanic agents daily dinner . dermatologicals . multivitamin tablet daily . pantoprazole daily breakfast . genitourinary agents miscellaneous pain .

Figure 3. Attention weights produced by BDRO for a section of a medical note in the medical record of a bipolar patient with severe impairment.

patient reporting worsening auditory hallucinations late . also diagnosed elevated lithium level . patient therefore hospitalized address lithium toxicity also treat worsening hallucinations . hospital course patient placed suicide precautions . **outset stay patient reporting worsening auditory hallucinations home . hallucinations seemed come television radio . said talking keeping house like otherwise saying nice things . also endorsed degree paranoia reported believes followed times . patient could explain lithium level elevated . reported husband generally handles medications laid sure perhaps taken extra lithium point . rate lithium level obviously needed addressed . maintained patient outpatient medications exception lithium held first hours . point drew lithium level lithium level . point time restarted patient reported lithium dose twice daily . tolerated well without adverse effects . throughout time simply maintained medications including primary antipsychotic appears trilacon . rather large dose trilacon noon . patient exhibit adverse effects rest medications . patient took part group therapy appropriately . noted respond internal stimuli particular noted observed paranoid . times would slightly irritable early morning beyond affect seemed fairly mobile . overall fairly pleasant . endorse suicidal homicidal thoughts throughout hospitalization . able perform ads overall well . day discharge drew lithium level lithium level came . thus seemed certainly overtaking lithium end lithium level .**

Figure 4. Attention weights produced by BDRO for a section of a medical note in the medical record of a bipolar patient with moderate impairment.

Figure 3 shows that BDRO is attending to patient-reported outcomes (e.g., “chronic feelings emptiness chronic suicidal ideations”) and clinicians reported outcomes (e.g., “mental status time discharge alert oriented time place person situation”). In general, the model assigns high weights to sentences indicative of severe impairment (e.g., “active suicidal homicidal ideations”) and low weights to sentences that show moderate impairment (e.g., “denying feeling hopeless worthless”). The latter example also illustrates the importance of context (i.e., “denying”) for the correct classification of the medical notes.

Compared to the severe impairment (Figure 3), the moderate impairment example (Figure 4) shows fewer sentences with high attention weights. Moreover, most of the sentences attended to reflect moderate functioning impairment (e.g., “patient reporting worsening hallucinations”, and “slightly irritable early morning”).

5. Discussion

Machine learning (ML) methodologies have been effectively utilized to enhance clinical decision support for mental health. Among others, these methodologies were used to estimate treatment outcomes for patients suffering from depression [44]; to identify bipolar patients from a cohort of psychiatric patients [45]; and to identify obsessive-compulsive disorder symptoms from the medical records of patients diagnosed with schizophrenia, schizoaffective disorder, or bipolar disorder [46].

The present paper contributes to these applications with a DRO framework that can be deployed along the electronic health record of a healthcare system in order to continuously monitor the severity of the functioning impairment of bipolar and schizophrenia patients. The results show that the GDRO was able to identify psychiatric patients with severe functioning impairment from the two target cohorts with an AUC of 76% prior to the onset of the disease. After the onset of the disease, bipolar and schizophrenia patients with severe functioning impairment were correctly classified with AUCs of 86% and 84%, respectively, using the respective disease-specific DROs.

The proposed DROs not only show good performance according to accepted standards [43] but also utilize data that are readily available for all patients. Compared to other frameworks that rely on the results of MRI images or gene expression profiles [45], which may not be available for all patients, the proposed DROs use only medical notes routinely collected during the encounters between the patients and their healthcare providers.

The present study demonstrated the potential of the proposed DROs to enhance the healthcare services of patients suffering from bipolar disorders and schizophrenia. However, these DROs suffer from a few limitations. First, it is desirable to extend the training and validation of the GDRO to other mental disease conditions. Moreover, this model had approximately a 10% lower AUC than the disease-specific BDRO and SDRO. Access to additional training data from other patients cohorts may help improve the performance of GDRO and its transfer learning potential to disease-specific DROs for other

mental disease conditions. Second, the medical notes used in the development of the DROs were annotated according to the GAF score. The GAF scale was widely used and was shown to align with several other scales such as the psychiatric Apgar scale [47] and the Zung depression scale [48]. However, it is being replaced by the WHODAS scale. Therefore, future work should investigate the development of DROs using the new WHODAS scale once sufficient samples become available. Third, the results show that the three token-level embeddings produced DROs with comparable AUCs. However, these DROs had different sensitivities and specificities. In the short term, future work should investigate these differences and propose automated calibration mechanisms that can help select the appropriate cut-off threshold for each type of token-level embedding. In the long term, it may be beneficial to develop a language model that is specifically trained with routine care medical notes.

6. Conclusions

An automated method for classifying the medical notes of patients with psychiatric disorders according to their impairment severity can help improve healthcare for the patients and enhance resource allocation for healthcare institutions. Several previous studies demonstrated the utility of ML models in identifying symptoms, predicting readmission, and assessing treatment outcomes for psychiatric patients. The present study extends these studies to the assessment of the functioning impairment severity in daily activities for two cohorts of patients suffering from bipolar disorder and schizophrenia. The proposed DROs can be deployed in healthcare systems and used to monitor these patients. Patients assigned to the severe impairment level may need additional evaluation using other CROs and laboratory tests. Future work will explore the application of the proposed framework to other mental disease conditions, will consider training and validating the framework with medical notes that are scored using the more recent WHODAS impairment scale, and will seek additional training data in order to improve the accuracy and generalizability of the proposed DROs.

Author Contributions: Conceptualization, R.K.K., M.R.I.P., F.C., P.R.D., M.A.B. and Z.B.M.; formal analysis, R.K.K., M.R.I.P., F.C., P.R.D., M.A.B. and Z.B.M.; methodology, R.K.K., M.R.I.P., F.C., P.R.D., M.A.B. and Z.B.M.; validation, R.K.K., M.R.I.P., F.C., P.R.D., M.A.B. and Z.B.M.; and writing— review and editing, R.K.K., M.R.I.P., F.C., P.R.D., M.A.B. and Z.B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

Institutional Review Board Statement: This study was approved by the Review Board of Indiana University (IRB number: 2011632512). We confirm that all of the methods were performed in accordance with the relevant guidelines and regulations.

Informed Consent Statement: The Ethics Committee of Indiana University waived the need for informed consent due to the retrospective nature of the study.

Data Availability Statement: The data that support the findings of this study are available from the Regenstrief Institute, but restrictions apply to the availability of these data, which were used under a research agreement for the current study and so are not publicly available. Data are, however, available from the corresponding author upon reasonable request and with permission of the Regenstrief Institute.

Acknowledgments: The authors would like to thank Jarod Baker and Anna Roberts of the Regenstrief Institute for their support. The high-performance computing services used in this study are supported in part by Lilly Endowment, Inc. through its support for the Indiana University Pervasive Technology Institute.

Conflicts of Interest: Ben Miled has a financial interest in DigiCare Realized and could benefit from the results of this research. Boustani serves as a Chief Scientific Officer and Co-Founder of BlueAgilis, Inc. and as the Chief Health Officer of DigiCare Realized, Inc. He has equity interest

in Blue Agilis, Inc.; DigiCare Realized, Inc.; Preferred Population Health Management LLC; and MyShift, Inc. (previously known as RestUp, LLC). He serves as an advisory board member for Acadia Pharmaceuticals; Eisai, Inc; Biogen; and Genentech. These conflicts have been reviewed by Indiana University and have been appropriately managed to maintain objectivity. The remaining authors declare no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

ATC	Anatomical therapeutic chemical classification system
BDRO	Bipolar digital-reported outcome
BERT	Bidirectional encoder representations from transformers
CRO	Clinician-reported outcome
DRO	Digital-reported outcome
GAF	General assessment of functioning
GRU	Gated recurrent unit
HAN	Hierarchical attention network
INPC	Indiana network for patient care
PRO	Patient-reported outcome
SDRO	Schizophrenia digital-reported outcome

References

- Moreno-Küstner, B.; Martin, C.; Pastor, L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS ONE* **2018**, *13*, e0195687.
- Lish, J.D.; Dime-Meenan, S.; Whybrow, P.C.; Price, R.A.; Hirschfeld, R.M. The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *J. Affect. Disord.* **1994**, *31*, 281–294. [\[CrossRef\]](#)
- Patel, K.R.; Cherian, J.; Gohil, K.; Atkinson, D. Schizophrenia: Overview and treatment options. *Pharm. Ther.* **2014**, *39*, 638.
- Fonseka, T.M.; Bhat, V.; Kennedy, S.H. The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. *Aust. N. Z. J. Psychiatry* **2019**, *53*, 954–964. [\[CrossRef\]](#)
- Corcoran, C.M.; Cecchi, G.A. Using language processing and speech analysis for the identification of psychosis and other disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2020**, *5*, 770–779. [\[CrossRef\]](#)
- AlHamed, F.; Ive, J.; Specia, L. Predicting moments of mood changes overtime from imbalanced social media data. In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, Seattle, WA, USA, 15 July 2022; pp. 239–244.
- Spitzer, R.L.; Kroenke, K.; Williams, J.B.; Patient Health Questionnaire Primary Care Study Group; Patient Health Questionnaire Primary Care Study Group. Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA* **1999**, *282*, 1737–1744. [\[CrossRef\]](#)
- Young, R.C.; Biggs, J.T.; Ziegler, V.E.; Meyer, D.A. A rating scale for mania: Reliability, validity and sensitivity. *Br. J. Psychiatry* **1978**, *133*, 429–435. [\[CrossRef\]](#)
- Kroenke, K.; Spitzer, R.L.; Williams, J.B. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* **2001**, *16*, 606–613. [\[CrossRef\]](#)
- Kay, S.R.; Fiszbein, A.; Opler, L.A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* **1987**, *13*, 261–276. [\[CrossRef\]](#)
- Frances, A.; Pincus, H.A.; First, M.B. Global Assessment of functioning scale (GAF). In *Diagnostic and Statistical Manual for Mental Disorders*, 4th ed.; (DSM-IV); American Psychiatric Association: Washington, DC, USA, 2006.
- Aas, I. Global Assessment of Functioning (GAF): Properties and frontier of current knowledge. *Ann. Gen. Psychiatry* **2010**, *9*, 20. [\[CrossRef\]](#)
- Gold, L.H. DSM-5 and the assessment of functioning: The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0). *J. Am. Acad. Psychiatry Law Online* **2014**, *42*, 173–181.
- Ustun, T.B.; Kostanjsek, N.; Chatterji, S.; Rehm, J.; World Health Organization. *Measuring Health and Disability: Manual of WHO Disability Assessment Schedule WHODAS 2.0*; World Health Organization: Geneva, Switzerland, 2010.
- Kotei, E.; Thirunavukarasu, R. A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. *Information* **2023**, *14*, 187. [\[CrossRef\]](#)
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
- Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.

18. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
19. Turc, I.; Chang, M.; Lee, K.; Toutanova, K. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. *arXiv* **2019**, arXiv:1908.08962. Available online: <https://arxiv.org/abs/1908.08962> (accessed on 1 June 2023).
20. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78.
21. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.
22. Yang, W.; Xie, Y.; Lin, A.; Li, X.; Tan, L.; Xiong, K.; Li, M.; Lin, J. End-to-End Open-Domain Question Answering with BERTserini. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, 2–7 June 2019; pp. 72–77.
23. Liu, Y.; Lapata, M. Text Summarization with Pretrained Encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3730–3740.
24. Mulyar, A.; Uzuner, O.; McInnes, B. MT-clinical BERT: Scaling clinical information extraction with multitask learning. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2108–2115. [[CrossRef](#)]
25. Hu, P.; Lin, C.; Su, H.; Li, S.; Han, X.; Zhang, Y.; Mei, J. Bluememo: Depression analysis through twitter posts. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 5252–5254.
26. Jeong, L.; Lee, M.; Eyre, B.; Balagopalan, A.; Rudzicz, F.; Gabilondo, C. Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in schizophrenia. *Psychiatr. Res. Clin. Pract.* **2023**. [[CrossRef](#)]
27. Kshatriya, B.S.A.; Nunez, N.A.; Resendez, M.G.; Ryu, E.; Coombes, B.J.; Fu, S.; Frye, M.A.; Biernacka, J.M.; Wang, Y. Neural language models with distant supervision to identify major depressive disorder from clinical notes. *arXiv* **2021**, arXiv:2104.09644.
28. Zhang, T.; Schoene, A.M.; Ji, S.; Ananiadou, S. Natural language processing applied to mental illness detection: A narrative review. *Npj Digit. Med.* **2022**, *5*, 46. [[CrossRef](#)]
29. Adhikari, A.; Ram, A.; Tang, R.; Lin, J. DocBERT: BERT for Document Classification. *arXiv* **2019**, arXiv:1904.08398.
30. Gao, S.; Alawad, M.; Young, M.T.; Gounley, J.; Schaefferkoetter, N.; Yoon, H.J.; Wu, X.C.; Durbin, E.B.; Doherty, J.; Stroup, A.; et al. Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3596–3607. [[CrossRef](#)] [[PubMed](#)]
31. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune bert for text classification? In Proceedings of the China national conference on Chinese Computational Linguistics, Kunming, China, 18–20 October 2019; Springer Nature Switzerland AG: Cham, Switzerland, 2019; pp. 194–206.
32. Wang, L.; Lakin, J.; Riley, C.; Korach, Z.; Frain, L.N.; Zhou, L. Disease trajectories and end-of-life care for dementias: Latent topic modeling and trend analysis using clinical notes. In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 3–7 November 2018; American Medical Informatics Association: Washington, DC, USA, 2018; Volume 2018, p. 1056.
33. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Francisco, CA, USA, 12–17 June 2016; pp. 1480–1489.
34. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
35. Zhang, N.; Jankowski, M. Hierarchical BERT for medical document understanding. *arXiv* **2022**, arXiv:2204.09600.
36. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*, 2nd ed.; World Health Organization: Geneva, Switzerland, 2004.
37. Goldman, H.H.; Skodol, A.E.; Lave, T.R. Revising axis V for DSM-IV: A review of measures of social functioning. *Am. J. Psychiatry* **1992**, *149*, 9.
38. Bird, S. NLTK: The natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia, 17–18 July 2006; pp. 69–72.
39. World Health Organization. Collaborating centre for drug statistics methodology. In *Guidelines for ATC Classification and DDD Assignment*; World Health Organization: Geneva, Switzerland, 2013; Volume 3.
40. Imming, P.; Buss, T.; Dailey, L.; Meyer, A.; Morck, H.; Ramadan, M.; Rogosch, T. A classification of drug substances according to their mechanism of action. *Die-Pharm.-Int. J. Pharm. Sci.* **2004**, *59*, 579–589.
41. Cho, K.; Merriënboer, B.; Gulcehre, C.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN encoder–decoder for Statistical Machine Translation. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014.
42. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Mandrekar, J.N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316. [[CrossRef](#)] [[PubMed](#)]
44. Webb, C.A.; Cohen, Z.D.; Beard, C.; Forgeard, M.; Peckham, A.D.; Björngvinsson, T. Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *J. Consult. Clin. Psychol.* **2020**, *88*, 25. [[CrossRef](#)]

45. Librenza-Garcia, D.; Kotzian, B.J.; Yang, J.; Mwangi, B.; Cao, B.; Lima, L.N.P.; Bermudez, M.B.; Boeira, M.V.; Kapczinski, F.; Passos, I.C. The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neurosci. Biobehav. Rev.* **2017**, *80*, 538–554. [[CrossRef](#)]
46. Chandran, D.; Robbins, D.A.; Chang, C.K.; Shetty, H.; Sanyal, J.; Downs, J.; Fok, M.; Ball, M.; Jackson, R.; Stewart, R.; et al. Use of natural language processing to identify obsessive compulsive symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder. *Sci. Rep.* **2019**, *9*, 14146. [[CrossRef](#)]
47. Dimsdale, J.E.; Jeste, D.V.; Patterson, T.L. Beyond the global assessment of functioning: Learning from Virginia Apgar. *Psychosomatics* **2010**, *51*, 515–519. [[CrossRef](#)]
48. Hall, R.C. Global assessment of functioning: A modified scale. *Psychosomatics* **1995**, *36*, 267–275. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.