applied sciences

MDPI

*Article*

# Outcome Prediction for Patients with Bipolar Disorder Using Prodromal and Onset Data †

**Yijun Shao [1,2], Yan Cheng [1,2], Srikanth Gottipati [1] and Qing Zeng-Treitler [1,2,*]**

[1] Biomedical Informatics Center, George Washington University, Washington, DC 20037, USA
[2] Washington DC VA Medical Center, Washington, DC 20422, USA
* Correspondence: zengq@gwu.edu; Tel.: +1-202-994-0477
† This paper is an extended version of our abstract published in the Conference Proceedings of American Society for Clinical Pathology (ASCP) 2019 Annual Meeting: Innovations in Personalized Medicine from Biomarkers to Patient-Centered Care, Phoenix, AZ, USA, 11–13 September 2019.

**Abstract:** Background: Predicting the outcomes of serious mental illnesses including bipolar disorder (BD) is clinically beneficial, yet difficult. Objectives: This study aimed to predict hospitalization and mortality for patients with incident BD using a deep neural network approach. Methods: We randomly sampled 20,000 US Veterans with BD. Data on patients' prior hospitalizations, diagnoses, procedures, medications, note types, vital signs, lab results, and BD symptoms that occurred within 1 year before and at the onset of the incident BD were extracted as features. We then created novel temporal images of patient clinical features both during the prodromal period and at the time of the disease onset. Using each temporal image as a feature, we trained and tested deep neural network learning models to predict the 1-year combined outcome of hospitalization and mortality. Results: The models achieved accuracies of 0.766–0.949 and AUCs of 0.745–0.806 for the combined outcomes. The AUC for predicting mortality was 0.814, while its highest and lowest values for predicting different types of hospitalization were 90.4% and 70.1%, suggesting that some outcomes were more difficult to predict than others. Conclusion: Deep learning using temporal graphics of clinical history is a new and promising analytical approach for mental health outcome prediction.

**Keywords:** prediction; bipolar disorder; deep neural network; support vector machine



**Citation:** Shao, Y.; Cheng, Y.; Gottipati, S.; Zeng-Treitler, Q. Outcome Prediction for Patients with Bipolar Disorder Using Prodromal and Onset Data. *Appl. Sci.* **2023**, *13*, 1552. https://doi.org/10.3390/app13031552

Academic Editors: Krzysztof Ejsmont, Aamer Bilal Asghar, Yong Wang and Rodolfo Haber

Received: 6 December 2022
Revised: 13 January 2023
Accepted: 18 January 2023
Published: 25 January 2023



## 1. Introduction

Outcome prediction can facilitate physician and patient decision making as well as disease management. In the management of serious mental illnesses, the accurate and early identification of patients at high risk of mortality and other adverse event will enable a targeted monitoring and interventions to improve prognosis. While many risk indices and calculators have been developed for a wide range of diseases and conditions, the outcome prediction of mental illness, especially, serious mental illnesses such as bipolar disorder, remains very challenging [1,2].

A wide array of features has been used in predictive modeling in mental health, ranging from early treatment response to genetics. The AUC of these predictive models, however, rarely reached 80%, which some argued is the threshold of clinical utility [1]. For instance, with data from mood disorder patients with first lifetime episodes of major depression, Tondo et al. used multivariate, logistic modeling and Bayesian methods to predict a later diagnosis of unipolar versus bipolar depressive disorder. It reported correct classification rates of 64–67% of cases and an AUC of 0.72 [3].

Challenges faced by outcome prediction in patients with mental illnesses include inadequate sample size, heterogeneity of the patients sharing the same diagnosis, and individual choices and disparities [1,4]. However, we can potentially overcome these challenges with larger samples, more nuanced feature representations, and state-of-the-art machine learning methods.

The key research question in this study was how well (in terms of AUC and accuracy) the adverse outcome (all-cause mortality and number of all-cause hospitalizations) of bipolar disorder (BD) can be predicted after the initial diagnosis. All-cause mortality and all-cause hospitalizations (especially, re-hospitalizations) are of great interest in clinical research. The AUC of the predictive models in various patient populations is often in the lower 70% range, only occasionally reaching 80% [5–8]. As regards mental health populations, however, it was difficult for us to find any studies that predicted mortality and hospitalization, even though a number of studies identified mental illness as the significant risk factor for all-cause mortality and all-cause hospitalization [9–11].

One related study predicted a change in diagnosis from major depression to bipolar disorder after antidepressant initiation. It reported a mean AUC of 76% (ranging from 0.73 to 0.80) [12]. A study predicted clinically relevant changes in bipolar disorder outside the clinic walls based on pervasive technology interactions via smartphone typing dynamics. The results showed a predictive accuracy of around 90% [13]. Busch et al. predicted remission using clinical trial data in BP patients with an AUC of 80% at baseline and of 75% at 1-year follow-up. The AUC was reduced to 67% at both baseline and 1-year follow-up when routinely collected electronic data were used [14]. While machine learning is increasingly used in clinical research [15], prediction models for bipolar, especially with hospitalization and mortality as outcomes, have not been previously reported. This makes our study novel and important.

In this study, we experimented with adverse outcome prediction in 20,000 BD patients, using data collected during the 1-year prodromal period before the initial BD diagnosis and during the first week after the diagnosis to predict 1-year post-diagnosis outcomes. Different from the traditional predictive models, our study exhibits innovation in two ways: (1) by using a temporal graph of nuanced features to represent clinical data during the prodromal and onset periods and (2) by applying a deep neural network learning algorithm to the temporal images, since deep learning has demonstrated superior performance over traditional machine learning algorithms [16].

The data source for the study was the clinical data warehouse (CDW) of the US Veteran Affairs (VA) health system. The VA has the "America's largest integrated health care system, providing care at 1298 health care facilities, including 171 medical centers and 1113 outpatient sites of care of varying complexity, serving 9 million enrolled Veterans each year" [17]. The CDW contains data domains for vital signs, pharmacy, laboratory, and inpatient and outpatient diagnoses, procedures, and text notes.

## 2. Materials and Methods

### 2.1. Study Population

The study population was BD patients in the US VA health system, who were identified using ICD codes (ICD9 codes: 296.4X, 296.5X, 296.6X, 296.7, 296.80, and 296.89; ICD10 codes: F31.X). A total of 207,838 patients were identified using the following criteria: (1) at least two encounters with an ICD of BD; (2) first BD diagnosis after 2001; (3) being in the VA healthcare system for $\geq$12 months before the first BD diagnosis; (4) being in the VA healthcare system for $\geq$12 months after the first BD diagnosis.

### 2.2. Clinical Features

In this study, we used features including prior hospitalizations, diagnoses, procedures, medications, note types, vital signs, lab results, and BD symptoms to build the prediction models. Since the note types are often correlated with diagnosis, we did not include note types in the potential variable list. We also leveraged existing terminologies and instruments to group individual data items into groups or themes. The presence of a BD symptom was determined by the presence of a keyword representing the symptom in the notes. The keywords came from standardized psychiatric measurement scales that are typically used in clinical trials. These include psychosis (PANSS), mania (YMRS, HAM-D), depression (HAM-D), sleep (MADRS), drug-induced movement disorders (SAS, BARS,

and AIMS), suicide risk (C-SSRS), and other such symptoms [18–25]. The keywords were grouped by hand, and the groups were called "BD symptoms". These features were obtained from unstructured (text) data. All the potential features are defined in Table 1.

**Table 1.** Potential features included for outcome prediction, with their definition.

| Features | Definition | Data Source |
|---|---|---|
| Hospitalization | Started from the corresponding admission date and ended at the corresponding discharge date | Admission/discharge data from the hospitalization table |
| Diagnosis groups | Collapsed ICDs to the first-level categories (e.g., cardiovascular system disorder, etc.). Mental disorder group included one additional level of details (e.g., dementias, alcohol-induced mental disorders, etc.) [26] | Primary and secondary diagnostic ICD9 codes |
| CPT groups | Grouped the CPT codes into seven categories (e.g., anesthesia, surgery, etc.) [27] | CPT codes |
| Vitals signs | Focused on five types of vital sign standardized values of each vital sign into a score between 0 and 1, with 0 for normal value, and 1 for the most extreme value | Vital sign data |
| Lab results | The top 10 most frequent lab analyses (e.g., carbon dioxide, etc.) Standardized values of each lab testing into a score between 0 and 1, with 0 meaning the population mean, and a higher score meaning farther away from the population mean | Lab data |
| BD symptoms | Extracted keywords from the notes to identify BD symptoms. The keywords came from a set of instruments [18–25]. The keywords were grouped into BD symptoms. | Medical notes |

### *2.3. Temporal Data Representation*

The patient data were represented in a temporal image (matrix) with a time window of 1 year before and 1 week after the initial BD diagnosis. In Figure 1, the x-axis represents the time, measured in weeks. The y-axis represents the variable list. For vital signs and lab tests, the pixel values were grey-scaled, representing the level of abnormality. For all other features, the pixel values were binary, representing the absence/presence of the corresponding clinical event or observation.


**Figure 1.** An example of a temporal image for an individual patient.

### 2.4. Outcome Variables

We collected the all-cause mortality and number of all-cause hospitalizations within 1 year after the diagnosis. For the prediction models, we focused on the prediction of multiple hospitalizations combined with mortality as the composite outcome, because death is a competing risk for hospitalization. Therefore, the outcome variable was binary-valued, with 1/0 representing the presence/absence of the adverse outcome. We experimented with 4 different levels of "adverseness": n+ hospitalizations combined with mortality, for n = 1, 2, 3, 4. The corresponding binary outcomes were abbreviated as "0 vs. 1+/mort", "1− vs. 2+/mort", "2− vs. 3+/mort" and "3− vs. 4+/mort", respectively.

### 2.5. Deep Neural Network

In this study, we predicted the outcomes using a deep neural network (DNN) model. DNN, an increasingly popular machine learning method, is sometimes superior to classic statical analyses in performance. Classic statistical analyses draw conclusions at the population level; DNN models can predict the outcomes at the individual level. Classic statistical models need to satisfy assumptions such as linear relationships between predictors and outcomes, normality for residual errors, or independence among predictors; DNN models do not rely on these assumptions. In addition, DNN techniques can be designed to accommodate the temporal sequence of observations with irregular time intervals, which is a common feature in clinical data. DNN has demonstrated the ability to model highly nonlinear relationships without extensive feature engineering, learn from massive amounts of data, and perform better than linear/logistic regression and traditional machine learning methods in many tasks [16,28–31].

In this study, the network was implemented using a Python library called Theano [32] together with a helper library called Lasagne [33]. The network had a total of 7 layers (Figure 2). Starting from the input layer, the 2nd layer was a convolutional layer, followed by a max pooling layer. The 4th layer was again a convolutional layer and again followed by a max pooling layer. The 6th layer was a fully connected layer, followed by the output layer. All the weights and biases involved in the layers were learned from the training data.

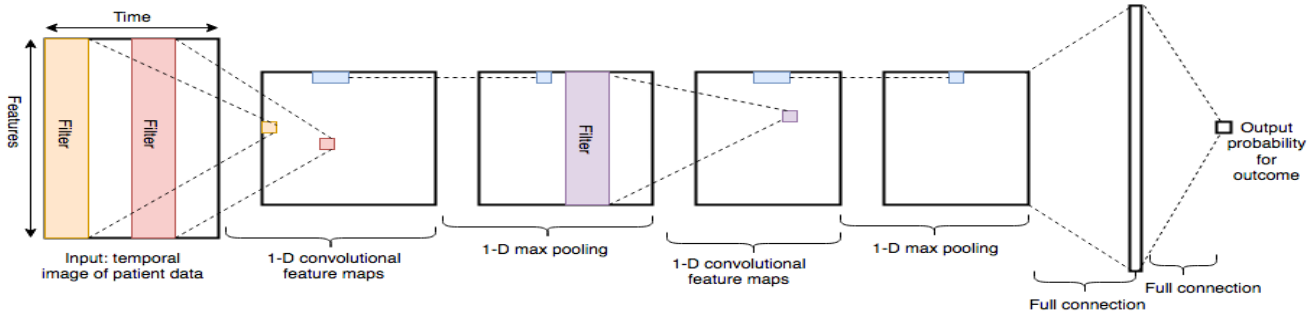


**Figure 2.** Illustration of the deep neural network.

The technical definition of each layer is as follows. The input layer took the temporal images $X = (x_{ij})(1 \le i \le 84, 1 \le j \le 54)$ as inputs. The second layer was a convolutional layer. It involved 10 filters $W^{(k)} = \left(w_{ij}^{(k)}\right)$ $(1 \le k \le 10,\ 1 \le i \le 84,\ 1 \le j \le 3)$, which were all $84 \times 3$ matrices (in Figure 2, two filters, colored in yellow and red, respectively, are displayed), as well as 10 biases $b_k$. The entries of the 10 matrices were the weights of this layer. The 10 filters were applied to the input layer through a time window of width 3, which was sliding in the time direction (from left to right in Figure 2). Mathematically, the output of the convolutional layer was a $10 \times 52$ matrix $Y = (y_{kl})$ obtained by

$$y_{kl} = f\left(\sum_{i=1}^{84}\sum_{j=0}^{2} x_{i,l+j} w_{i,3-j}^{(k)} + b_k\right) \quad (1)$$

where $f$ is the rectified linear function (i.e., $f(x) = max(0, x)$). The third layer was a max pooling layer. This layer had no weights, and all it did was take the maximum from a sliding time window of length 3 for each row. Mathematically, the output of this max pooling $Z = (z_{km})$ was computed as:

$$z_{km} = max_{m \le l \le m+2} y_{kl} \quad (2)$$

Therefore, the output $Z$ was a $10 \times 50$ matrix. The fourth layer was another convolutional layer. Its definition was similar to that of the second layer, except that it took the output $Z$ of the third layer as input. It also had 10 filters and 10 biases. The fifth layer was another max pooling layer, whose definition was the same as that of the third layer. The output of the fifth layer was a $10 \times 46$ matrix. The sixth layer was a full connection layer, meaning that each entry of the input (from the fifth layer) and each entry of the output were together associated with a weight and a bias. We set the output to be a vector $P = (p_i)$ of dimension 15. Mathematically, if we denote the output of the fifth layer as $Q = (q_{kl})$, the output is

$$p_i = f\left(\sum_{all\ k,l} q_{kl} v_{kli} + b_{kli}\right) \quad (3)$$

where $u_{kli}$ and $b_{kli}$ are the weights and biases, respectively, and $f$ is the rectified linear function. The last layer of the network was a full connection layer as well. The output $O$ was a scalar computed as

$$O = g\left(\sum_{1 \le i \le 15} p_i u_i + b_i\right) \quad (4)$$

where $u_i$ and $b_i$ are the weights and biases, respectively, and $g$ is the sigmoid function (i.e., $g(x) = e^x/(1 + e^x)$). The function $g$ made the output $O$ be between 0 and 1; hence, it was a probability value.

### *2.6. Training and Testing of the DNN*

The studied 20,000 patients were divided randomly into 3 sets: a training set (70%), a validation set (10%), and a testing set (20%). The training set was used to build the prediction model, during which all the parameters (i.e., the weights and biases) in the network were updated, so that the prediction outputs approximated the true outcomes. The parameters were initially set to be random values and were updated iteratively using a common approach called mini-batch stochastic gradient descent. Specifically, the training set was randomly divided into small groups of 100, with each group called a mini-batch. The temporal data of each mini-batch (i.e., 100 temporal images) were supplied to the input layer to compute 100 output probabilities, which were compared with the corresponding 100 binary outcomes to calculate a distance. The distance was calculated using a binary cross-entropy metric. We then added L1 and L2 regularizations to the distance to make an objective function. The purpose of the added L1 and L2 regularizations was to make the parameters sparse and to prevent the parameters from being too large (in absolute value), respectively. Then, the parameters were updated to reduce the objective function. One pass over the training set is called an epoch. Since the training set had 14,000 patients, the parameters would be updated 140 (=14,000/100) times within one epoch.

After one epoch ends, the training could be continued on another epoch. However, the training must be stopped at some point because the overfitting effect would appear if training continued without stopping. A common approach to prevent overfitting is to use a validation set. In this study, we used the validation set to determine after how many epochs the training should stop. After each epoch of training, we applied the trained model to the pre-set validation set to calculate the performance of the prediction. We chose AUC to measure the performance. Initially, the AUC on the validation set usually showed an increasing trend from one epoch to the next. After some epochs of training, the increasing trend stopped, and a decreasing trend would appear. Note that an increasing trend does not mean the value is always increasing. Occasionally, the value would decrease slightly during one or more epochs but would increase again after that and reach a higher level than before, and that is still counted as an increasing trend. To find when the decreasing trend

started, we looked for the number of epochs such that the AUC at that point was higher than all the AUCs before and also higher than 10 consecutive AUCs afterward. Then, we selected the model corresponding to the AUC value and thus found the final model, which was then applied to the testing set to report the performance in terms of AUC and accuracy.

### 2.7. Support Vector Machine

For comparison, we also trained a support vector machine (SVM) model with a linear kernel for the prediction of the outcomes. The model was implemented using a Python machine-learning library called Scikit-Learn [34]. We used non-temporal features, which were the same as in the temporal images but aggregated over time by taking maximum values along the rows. The training and testing of the SVM model used the same training, validation, and testing set as for the DNN model. The SVM model had only one hyper-parameter "C". We adjusted the C-value using the validation set for optimal AUC. Then, we tested the model on the testing set.

## 3. Results

We identified a total of 346,511 patients with at least two encounters with a BD diagnosis. After we applied the inclusion/exclusion criteria, 207,838 patients remained. Among these patients, 20,000 were randomly selected for the analysis. The baseline characteristics of the randomly selected 20,000 BD patients are described in Table 2. The mean age of the patients was 48.8 (SD 13.5) years. Most of the patients were male, who accounted for 84.2% of them; about 73.4% were Whites, followed by Blacks (17.2%), unknown (7.5%), and others (2.0%); 5.4% were Hispanics.

**Table 2.** Baseline characteristics of the randomly selected 20,000 BD patients.

| Demographics | Mean/N | STD/% |
|---|---|---|
| Age (Years) | 48.8 | 13.5 |
| Male | 16,842 | 84.2% |
| Race White | 14,673 | 73.4% |
| Race Black | 3434 | 17.2% |
| Race Others | 395 | 2.0% |
| Race Unknown | 1498 | 7.5% |
| Ethnicity Hispanics | 1072 | 5.4% |
| Ethnicity Non-Hispanics | 17,896 | 89.5% |
| Ethnicity Unknown | 1032 | 5.2% |

The outcome hospitalizations of the cohort had various types as indicated by the admission ICD codes. The distributions of the hospitalization types among the outcome hospitalizations and the patients are summarized in Table 3. To provide a nuanced analysis of the DNN performance, the hospitalizations were first divided into two larger groups, i.e., mental and non-mental, and then divided into smaller groups. The smaller groups were the same as the diagnosis groups listed in Table 1. Here, we only list the smaller groups which had 100 or more associated patients. The result showed that 64% of the outcome hospitalizations were caused by mental disorders, and 69% of the hospitalized patients had such a hospitalization as an outcome.

**Table 3.** Number of hospitalizations and patients by hospitalization type.

| Hospitalization Type | Number of Hospitalizations | Number of Patients |
|---|---|---|
| **All-cause** | **10,961** | **5276** |
| **Mental** | **7035 (64%)** | **3659 (69%)** |
| Episodic mood disorders (296) | 2321 (21.2%) | 1613 (14.7%) |
| Alcohol dependence syndrome (303) | 1156 (10.5%) | 737 (6.7%) |

**Table 3.** *Cont.*

| Hospitalization Type | Number of Hospitalizations | Number of Patients |
|---|---|---|
| Adjustment reaction (309) | 664 (6.1%) | 514 (4.7%) |
| Schizophrenic disorders (295) | 811 (7.4%) | 450 (4.1%) |
| Drug dependence (304) | 504 (4.6%) | 398 (3.6%) |
| Drug-induced mental disorders (292) | 368 (3.4%) | 271 (2.5%) |
| Alcohol-induced mental disorders (291) | 308 (2.8%) | 225 (2.1%) |
| Depressive disorder not elsewhere classified (311) | 254 (2.3%) | 221 (2%) |
| Nondependent abuse of drug (305) | 234 (2.1%) | 200 (1.8%) |
| Other nonorganic psychoses (298) | 126 (1.1%) | 104 (0.9%) |
| Anxiety dissociative and somatoform disorders (300) | 117 (1.1%) | 101 (0.9%) |
| **Non-mental** | **3604 (33%)** | **2158 (41%)** |
| Symptoms signs and ill-defined conditions (780–799) | 518 (4.7%) | 434 (4%) |
| Diseases of the circulatory system (390–459) | 547 (5%) | 401 (3.7%) |
| Diseases of the digestive system (520–579) | 408 (3.7%) | 321 (2.9%) |
| Supplementary classification of factors influencing health status and contact with health services (v01–v91) | 352 (3.2%) | 301 (2.7%) |
| Injury and poisoning (800–999) | 341 (3.1%) | 291 (2.7%) |
| Diseases of the respiratory system (460–519) | 338 (3.1%) | 267 (2.4%) |
| Diseases of the musculoskeletal system and connective tissue (710–739) | 186 (1.7%) | 165 (1.5%) |
| Diseases of the genitourinary system (580–629) | 170 (1.6%) | 143 (1.3%) |
| Endocrine nutritional and metabolic diseases and immunity disorders (240–279) | 191 (1.7%) | 130 (1.2%) |
| Diseases of the nervous system and sense organs (320–389) | 140 (1.3%) | 121 (1.1%) |
| Neoplasms (140–239) | 150 (1.4%) | 120 (1.1%) |

We trained the DNN models and SVM models for the prediction of the four levels of adverse outcomes. The performance of the models on the testing set is reported in Table 4. In particular, the AUCs of the models are plotted in Figure 3 for direct comparison.

**Table 4.** Performance of the predictive models on the testing set.

| Outcome | DNN AUC | DNN Accuracy | SVM AUC | SVM Accuracy |
|---|---|---|---|---|
| 0 vs. 1+/mort | 0.750 | 0.767 | 0.740 | 0.757 |
| 1− vs. 2+/mort | 0.776 | 0.867 | 0.770 | 0.861 |
| 2− vs. 3+/mort | 0.794 | 0.921 | 0.780 | 0.920 |
| 3− vs. 4+/mort | 0.806 | 0.946 | 0.796 | 0.949 |

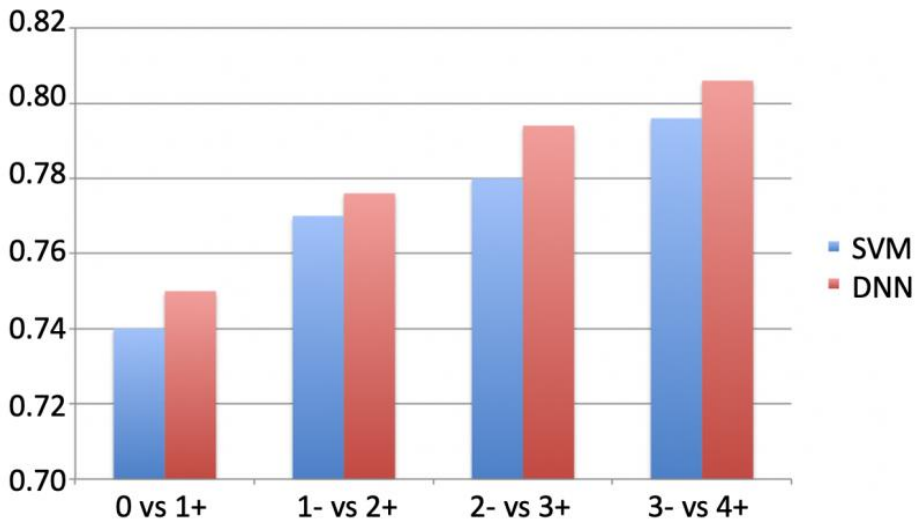


**Figure 3.** AUCs of the DNN model and SVM model.

It was also interesting to ask how the models performed in predicting hospitalizations only vs. mortality, as well as how the models (for predicting "1- vs. 2+/mort") performed in predicting different hospitalization types. The results showed that the prediction of mortality was better than that of hospitalizations only (AUC 0.814 vs. 0.774), and the prediction of non-mental illnesses was better than that of mental illnesses (Table 5). However, on the finer-grained level (smaller group), the prediction of hospitalizations caused by alcohol-induced mental disorders was the best among all.

**Table 5.** Performance comparison across different hospitalization types.

| Hospitalization Type | AUC |
|---|---|
| **All-cause** | **0.776** |
| **Mental** | **0.779** |
| Episodic mood disorders (296) | 0.765 |
| Alcohol dependence syndrome (303) | 0.812 |
| Adjustment reaction (309) | 0.762 |
| Schizophrenic disorders (295) | 0.865 |
| Drug dependence (304) | 0.824 |
| Drug-induced mental disorders (292) | 0.821 |
| Alcohol-induced mental disorders (291) | 0.904 |
| Depressive disorder not elsewhere classified (311) | 0.811 |
| Nondependent abuse of drug (305) | 0.834 |
| Other nonorganic psychoses (298) | 0.711 |
| Anxiety dissociative and somatoform disorders (300) | 0.701 |
| **Non-mental** | **0.826** |
| Symptoms signs and ill-defined conditions (780–799) | 0.832 |
| Diseases of the circulatory system (390–459) | 0.841 |
| Diseases of the digestive system (520–579) | 0.855 |
| Supplementary classification of factors influencing health status and contact with health services (v01–v91) | 0.837 |
| Injury and poisoning (800–999) | 0.814 |
| Diseases of the respiratory system (460–519) | 0.852 |
| Diseases of the musculoskeletal system and connective tissue (710–739) | 0.822 |
| Diseases of the genitourinary system (580–629) | 0.854 |
| Endocrine nutritional and metabolic diseases and immunity disorders (240–279) | 0.902 |
| Diseases of the nervous system and sense organs (320–389) | 0.745 |
| Neoplasms (140–239) | 0.772 |

## 4. Discussion

Prior predictive modeling efforts in the psychiatric domain had difficulty achieving an AUC or accuracy of 80%. Using an image representation of clinical data during the prodromal and onset period and a deep neural network, we were able to achieve accuracies of 0.766–0.949 and AUC of 0.745–0.815. In comparison, using the same features without temporal representation and a traditional machine learning algorithm, lower accuracies (0.757–0.949) and AUC (0.740–0.796) were obtained. This demonstrated the feasibility of predicting the outcomes of serious mental illnesses.

The ability to predict a patient outcome at the time of the initial BD diagnosis has several implications. Patients who are predicted to have poor outcomes can be closely monitored and considered for additional interventions by clinicians. The predictive results may be shared with the patients to allow for a shared decision making. The prediction of poor outcomes may also be used to motivate the patients to adhere to the prescribed treatments, since treatment adherence is a frequent problem in BD patients.

Hospitalization and mortality, combined or as separate outcomes, are clinically very important. Few prior studies reported bipolar outcome prediction [12–14]; however, they did not use hospitalization and mortality as outcomes. This prevented us from performing a direct comparison in terms of accuracy or AUC. There is evidence that predicting outcomes using routinely collected clinical data, as we used in this study, is challenging [14].

Machine learning generally benefits from a large sample size. The fact that the sample size in our experiment was 20,000 helped to boost the predictive performance. From the VA database, we could pull in several times more BD patients. This suggests that the predictive accuracy and AUC could increase. We will use a larger sample size (e.g., n = 100,000) in a future study.

Features and learning algorithms are also important factors for predictive modeling. The features include both clinical features (e.g., diagnoses) and temporal features. Given the prior literature, we included several types of coded data features (i.e., hospitalization, diagnosis, procedures, vitals, and lab results). We did not conduct variable selection but did group the numerous diagnoses and procedures based on existing terminologies. We also only included a small number of common lab tests. One of the advantages of deep learning is it does not require elaborate feature engineering, which is why we skipped the temporal feature extraction. At the same time, the prevalence of individual ICD or CPT codes tends to be very low, and thousands of ICD and CPT codes exist. If we do not aggregate diagnoses and procedures, the temporal image will be very large and mostly blank, which is not suitable for either any human inspection or deep learning. For example, for the same reason, given a large number of words, word2vec is often used for a reduction in dimensionality before applying a deep learning algorithm to a text.

One limitation of the dataset we used is that it included predominantly male patients. Another is that we did not have the lifelong record of each patient. While we did require a minimum of 1 year of history, the incident bipolar diagnoses in our dataset may not be the true onset of the disease.

This study represents a first step in applying deep learning and temporal images to mental health risk prediction with promising results. There are many alternatives that we have yet to explore. For example, the choice of a 1-year prodromal period is somewhat arbitrary, and we could examine 2-year and 3-year prodromal periods in the future. We focused on the 1-year outcome in this study. The prediction of longer term (e.g., 3-, 5-, 10-year) outcomes would be of interest in future studies.

There are several ways to represent temporal information. We only aggregated the data by week. In some prior research, data closer to an event (in our case, the event would be the BD diagnosis) could be shown with finer granularity (e.g., day or hour), and data further from an event could be shown with coarse granularity (e.g., week or month).

## 5. Conclusions

Applying deep learning and temporal graphics is a promising approach to predicting adverse health outcomes of BD patients. Depending on the number and type of hospitalization as outcomes, the best accuracies and AUCs reached over 90%. However, single hospitalizations in a year and hospitalizations due to other nonorganic psychoses and anxiety dissociative and somatoform disorders were harder to predict.

**Author Contributions:** Q.Z.-T. conceived the study design and supervised the findings of this work; Y.S. developed methods to create temporal images and apply deep neural network learning algorithms and drafted the manuscript; Y.C. created the cohort and prepared the data; S.G. offered advice on the study design and commented on the analysis. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (IRB) of Washington DC Veterans Affairs (#1595522 approved on 18 December 2018).

**Informed Consent Statement:** Patient consent was waived because it was not practicable to obtain consent from the large numbers of patients for a retrospective study on the existing databases.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We thank the US Department of Veteran Affairs for providing data for this study.

**Conflicts of Interest:** Gottipati, S is currently an employee of Cerevel Therapeutics. The other authors have no conflicts of interest to declare.

## References


1. McMahon, F.J. Prediction of treatment outcomes in psychiatry—Where do we stand? *Dialog. Clin. Neurosci.* **2014**, *16*, 455–464. [CrossRef] [PubMed]
2. Treuer, T.; Tohen, M. Predicting the course and outcome of bipolar disorder: A review. *Eur. Psychiatry* **2010**, *25*, 328–333. [CrossRef]
3. Tondo, L.; Visioli, C.; Preti, A.; Baldessarini, R.J. Bipolar disorders following initial depression: Modeling predictive clinical factors. *J. Affect. Disord.* **2014**, *167*, 44–49. [CrossRef] [PubMed]
4. Evans, T.S.; Berkman, N.; Brown, C.; Gaynes, B.; Weber, R.P. *Disparities within Serious Mental Illness*; Agency for Healthcare Research and Quality (US): Rockville, MD, USA, 2016.
5. Jamei, M.; Nisnevich, A.; Wetchler, E.; Sudat, S.; Liu, E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PLoS ONE* **2017**, *12*, e0181173. [CrossRef] [PubMed]
6. Lassale, C.; Gunter, M.J.; Romaguera, D.; Peelen, L.M.; Van Der Schouw, Y.T.; Beulens, J.W.J.; Freisling, H.; Muller, D.C.; Ferrari, P.; Huybrechts, I.; et al. Diet Quality Scores and Prediction of All-Cause, Cardiovascular and Cancer Mortality in a Pan-European Cohort Study. *PLoS ONE* **2016**, *11*, e0159025. [CrossRef] [PubMed]
7. Upshaw, J.N.; Konstam, M.A.; van Klaveren, D.; Noubary, F.; Huggins, G.S.; Kent, D.M. Multistate Model to Predict Heart Failure Hospitalizations and All-Cause Mortality in Outpatients With Heart Failure With Reduced Ejection Fraction. *Circ. Hear. Fail.* **2016**, *9*, e003146. [CrossRef]
8. Wan, E.Y.F.; Fong, D.Y.T.; Fung, C.S.C.; Yu, Y.T.E.; Chin, W.Y.; Chan, A.K.C.; Lam, C.L.K. Prediction of five-year all-cause mortality in Chinese patients with type 2 diabetes mellitus—A population-based retrospective cohort study. *J. Diabetes Its Complicat.* **2017**, *31*, 939–944. [CrossRef] [PubMed]
9. Kane, K.D.; Yochim, B.P.; Lichtenberg, P.A. Depressive symptoms and cognitive impairment predict all-cause mortality in long-term care residents. *Psychol. Aging* **2010**, *25*, 446–452. [CrossRef]
10. Keyes, C.L.M.; Simoes, E.J. To Flourish or Not: Positive Mental Health and All-Cause Mortality. *Am. J. Public Health* **2012**, *102*, 2164–2172. [CrossRef]
11. Kripalani, S.; Theobald, C.N.; Anctil, B.; Vasilevskis, E.E. Reducing Hospital Readmission Rates: Current Strategies and Future Directions. *Annu. Rev. Med.* **2014**, *65*, 471–485. [CrossRef]
12. Pradier, M.F.; Hughes, M.C.; Jr, T.H.M.; Barroilhet, S.A.; Doshi-Velez, F.; Perlis, R.H. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. *Neuropsychopharmacology* **2020**, *46*, 455–461. [CrossRef] [PubMed]
13. Bennett, C.C.; Ross, M.K.; Baek, E.; Kim, D.; Leow, A.D. Predicting clinically relevant changes in bipolar disorder outside the clinic walls based on pervasive technology interactions via smartphone typing dynamics. *Pervasive Mob. Comput.* **2022**, *83*, 101598. [CrossRef]
14. Busch, A.B.; Neelon, B.; Zelevinsky, K.; He, Y.; Normand, S.-L.T. Accurately Predicting Bipolar Disorder Mood Outcomes. *Med. Care* **2012**, *50*, 311–319. [CrossRef] [PubMed]
15. Thieme, A.; Belgrave, D.; Doherty, G. Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. *ACM Trans. Comput.-Hum. Interact.* **2020**, *27*, 34. [CrossRef]
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
17. The U.S. Department of Veterans Affairs. Available online: https://www.va.gov/health/ (accessed on 11 January 2023).
18. Barnes, T.R.E. A Rating Scale for Drug-Induced Akathisia. *Br. J. Psychiatry* **1989**, *154*, 672–676. [CrossRef]
19. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **1960**, *23*, 56–62. [CrossRef]
20. Kay, S.R.; Fiszbein, A.; Opler, L.A. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr. Bull.* **1987**, *13*, 261–276. [CrossRef]
21. Montgomery, S.A.; Smeyatsky, N.; de Ruiter, M.; Montgomery, D.B. Profiles of antidepressant activity with the Montgomery-Asberg depression rating scale. *Acta Psychiatr. Scand.* **1985**, *72*, 38–42. [CrossRef]
22. Posner, K.; Brent, D.; Lucas, C.; Gould, M.; Stanley, B.; Brown, G.; Fisher, P.; Zelazny, J.; Burke, A.; Oquendo, M.; et al. Columbia-Suicide Severity Rating Scale (C-SSRS). Available online: https://cssrs.columbia.edu/wp-content/upAloads/C-SSRS_Pediatric-SLC_11.14.16.pdf (accessed on 17 January 2023).
23. Simpson, G.M.; Angus, J.W. A rating scale for extrapyramidal side effects. *Acta Psychiatr. Scand.* **1970**, *45*, 11–19. [CrossRef]
24. Guy, W. Abnormal Involuntary Movement Scale (117-AIMS). In *ECDEU Assessment Manual for Psychopharmacology*; National Institute of Mental Health: Rockville, MD, USA, 1976; pp. 534–537.
25. 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] [PubMed]
26. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Centers for Disease Control and Prevention: National Center for Health Statistics. Available online: https://www.cdc.gov/nchs/icd/icd9cm (accessed on 1 December 2022).
27. CPT®Overview and Code Approval. American Medical Association. Available online: https://www.ama-assn.org/practice-management/cpt (accessed on 1 December 2022).

28. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinform.* **2017**, *19*, 1236–1246. [CrossRef] [PubMed]
29. Belle, A.; Thiagarajan, R.; Soroushmehr, S.M.; Navidi, F.; Beard, D.A.; Najarian, K. Big Data Analytics in Healthcare. *Biomed. Res. Int.* **2015**, *2015*, 370194. [CrossRef] [PubMed]
30. Lee, J.-G.; Jun, S.; Cho, Y.-W.; Lee, H.; Kim, G.B.; Seo, J.B.; Kim, N. Deep Learning in Medical Imaging: General Overview. *Korean J. Radiol.* **2017**, *18*, 570–584. [CrossRef]
31. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine Learning for Medical Imaging. *Radiographics* **2017**, *37*, 505–515. [CrossRef]
32. Bergstra, J.; Breuleux, O.; Bastien, F.; Lamblin, P.; Pascanu, R.; Desjardins, G.; Turian, J.; Warde-Farley, D.; Bengio, Y. Theano: A CPU and GPU Math Compiler in Python. In Proceedings of the 9th Python in Science Conference (SCIPY 2010), Austin, RX, USA, 28 June–3 July 2010.
33. Zenodo. Lasagne: First Release. Available online: https://zenodo.org/record/27878#.Y-DRX3YzZPY (accessed on 17 January 2023).
34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830.