Predicting Substance Misuse Admission Rates via Recurrent Neural Networks

Matthew J. Howard  
Computer Science and Engineering  
UC Santa Cruz  
Santa Cruz, CA, USA  
matthoward@ucsc.edu

Rakshit Agrawal  
Computer Science and Engineering  
UC Santa Cruz  
Santa Cruz, CA, USA  
ragrawa1@ucsc.edu

Abstract—Substance misuse affects millions of American adults each year, including 19.7 million adults who battled substance use disorders in 2017 [1]. Substance use disorders contribute immensely to the prevalence of disease, mental health disorders, homelessness, and cost American society greater than $740 billion annually in health care, crime, and lost workplace productivity. Beyond the annual fiscal burden cast by substance misuse at the national level, the introduction and/or spread of substance misuse within communities may cripple local economies and destroy community stability. Identifying at-risk communities may allow policy makers to focus efforts on identifying local causal factors of substance misuse trends and subsequently assist them in making informed policy decisions to help curb the spread of substance use.

In this work, we present a sequential statistical model built on Recurrent Neural Networks for predicting geographic locations that present high future risk for increased substance misuse cases. Our model leverages 17 years (2000-2016) of patient-level admission data via the Treatment Episode Data Set for Admissions (TEDS-A) from the United States Substance Abuse and Mental Health Services Administration (SAMHSA) that catalogues anonymized demographic data, mental health conditions, and drugs of abuse, amongst other information, for patients that are admitted to publically-funded substance misuse treatment facilities within designated metro- and micro-politan areas. Our model leverages the temporal (yearly) and spatial (geographical) structure of the data to predict how past trends influence future substance abuse admission rates, achieving 70% binary classification accuracy in predicting geographic regions expected to see a year-over-year increase in substance use admissions.

I. INTRODUCTION

Substance misuse disorders directly affect millions of American adults each year, gravely impacting individuals, communities, and the nation at large. According to the National Survey on Drug Use and Health (NSDUH) [1] conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), approximately 19.7 million adults aged 12 or older had a substance use disorder related to drug use activity within the previous year. An evaluation by the National Institute on Drug Abuse places the economic burden of illicit drug use and prescription misuse to the U.S. at more than $740 billion per year related to health care, crime, and lost workplace productivity [3]. The fiscal impacts are grave; however, the immeasurable detriment to individual lives, family cohesion, and local community stability is perhaps even more harrowing.

Given the societal implications, detecting drug misuse trends remains an important challenge for governments and legislators. The U.S. previously established several efforts toward addressing this challenge, including the establishment of the Drug Abuse Warning Network (DAWN) [4] in 1976, and subsequent re-establishment in 2018 after discontinuation in 2011. The goal of DAWN is to provide a nationwide public health surveillance system to monitor substance use crises and function as an early warning system for detecting outbreaks as well as monitoring the geographic, temporal, and demographic distributions of these detected outbreaks.

In alignment with the goals of DAWN, and with the United Nations Global Goals1 Target 3.5 for prevention and treatment of substance abuse, we are proposing a sequential statistical model utilizing Recurrent Neural Networks (RNNs) to predict substance use treatment admission trends for Core-Based Statistical Areas (CBSAs) – metro- and micro-politan areas defined by the U.S. Office of Management and Budget – from anonymized, temporal, patient-level admissions data. Our model augments patient-level data with CBSA-aggregated data to train the proposed RNN architecture across consecutive years to predict future trends. The aim of our method is to provide a measure for geographic regions at risk of future increases in substance misuse trends, and potentially serve medical and legislative sectors in monitoring these trends.

We apply our method to over 22 million patient substance misuse admission records collected between 2000-2016 and spanning 477 CBSA geographies from the Treatment Episode Data Set for Admissions (TEDS-A) from SAMHSA [2] that catalogues demographics, substances used upon admission, mental health conditions, and other information.

In the following section, we summarize related literature on monitoring and predicting drug misuse trends. Additionally, we provide a formal definition of our prediction problem, and provide a detailed description of the TEDS-A dataset that we use to train our model. An overview and description of our statistical model is given, followed by an evaluation of our prediction results and discussion of the implications of these results. Finally, we discuss directions for extensions of our model and highlight avenues for future work.

1The Global Goals: https://www.globalgoals.org/
II. RELATED WORK

A. Analyzing Drug Misuse Trends

Analysis of drug misuse relies on identifying meaningful trends that characterize and highlight potential causal factors for the misuse. Crane and Cai [5] analyzed emergency department visits relating to underage drinking and discovered that certain weekly times (Friday and Saturday nights from 9pm to 6am) experienced significantly higher cases of adolescent emergency visits than other times of the week. In addition, the authors showed that emergency visits for young adults typically began later (between 12am and 6am), indicating a significant difference in alcohol consumption behavior between these two age groups. Spiller et. al. [9] examined social, geographic, and demographic trends using poison center information in conjunction with census data, identifying that as poverty and unemployment rates increased, prescription opioid use rates also increased.

B. Machine Learning for Predicting Drug Misuse Patterns

Beyond analysis, several statistical machine learning models have been proposed in relation to predicting substance use disorder and related trends, including the comparison and combination of several models (logistic regression, random forests, and neural networks) for predicting treatment success [6]. Unfortunately, the authors limit their data to only one population (Hispanic origin), include only non-structural demographic and case-specific data, and utilize only 45% of their original dataset to gain a marginal 3% relative increase in predictive accuracy over logistic regression.

Hassanpour et. al. [7] propose a deep-learning method for identifying substance abuse risk for individuals by extracting social media information in conjunction with a substance use screener assessment to train their model. Che et. al. [8] also utilize deep learning methods to classify opioid drug uses into different dependency levels, significantly improving classification accuracy over standard statistical learning techniques such as logistic regression, support-vector machines, and random forests.

III. PROBLEM DEFINITION, MOTIVATION, AND DATA

In this work, we seek to develop a deep-learning model for predicting temporal trends in substance abuse facility admissions across defined geographical regions within the U.S. Specifically, we aim to build a model that leverages both patient-level and geographic data features to more accurately predict potential increases (and decreases) in substance misuse admission levels for geographic areas in which data is available. Importantly, we do not seek to quantify drug misuse trends outside of admission rates; however, we believe that developing such a model may assist in identifying causal societal factors that influence trends in drug misuse.

Recurrent Neural Network models are deep learning models that are acutely situated for sequential learning tasks, such as temporal prediction and classification. In particular, Long Short-Term Memory (LSTM) networks [11] are RNNs that are especially well-suited for these tasks, as they are able to capture long-term dependencies in sequential data.

We formulate the problem of predicting substance misuse admissions as a binary classification task. In the prediction task, we assume we are provided a priori all annual admissions data for all considered geographic regions for a finite number of consecutive years, and wish to make predictions for the first year following the range provided. For binary classification, the goal is to correctly predict whether or not a geographic region will experience an increase in admissions the following year.

A. Data: TEDS-A (2000-2016)

The Treatment Episode Data Set for Admissions (TEDS-A) for 2000-2016 provides 31,406,891 anonymized patient admission records from publicly-funded substance use treatment facilities for 811 U.S. CBSAs. In addition to the year and CBSA in which a record was recorded, each record includes demographic information (age, race, sex, etc.), substances and administration routes used by the subject, referral sources, and psychiatric and economic items. In this work, we limit the data used for analysis to 477 CBSAs within the contiguous U.S. that appear in all 17 years of the admission data, resulting in 22,612,190 (72%) remaining records. Each variable in the dataset is encoded categorically.

Data Limitations: There are several limitations of the TEDS-A dataset highlighted by SAMHSA alongside the publication of the data [2]. First, facility inclusion in the TEDS dataset is affected by local certification, licensure, and public funding. Second, some reporting facilities report admissions financed by public and private funds, while others provide only those funded by public funds. Additionally, admissions are defined differently across different states, and thus absolute numbers between geographic regions are incomparable. These limitations highlight the need for geo-centric models that more adequately capture region-specific trends.

B. Motivation

Given the social influence of substance use, the problem of substance misuse is inherently societal, and thus dependent upon the geography of the society. Figure 1 demonstrates this dependency by mapping 1-year annual changes in admissions for reporting geographies within the TEDS-A dataset. In Figure 1a, widespread admission increases are seen across much of the contiguous U.S. from 2000 to 2001, with major increases clustered in the central U.S. Annual changes in substance use admissions vary widely year-over-year, however, such as from 2015 to 2016 (Figure 1d) where major increases are exhibited in the Carolinas, while major decreases are seen in the Mid-Atlantic. These distinct “pockets” of changes further emphasize the geographic influence on substance misuse across the nation. Additionally, the non-linear nature of the increases/decreases in admissions with respect to time also motivates the use of powerful statistical models, such as RNNs, to learn how these non-linear relationships evolve year-over-year.
IV. METHODS

Below, we outline our data preparation for predicting admission rate increases in the binary classification setting, and then we discuss the design of our sequential deep learning model.

A. Feature Extraction & Labeling

For the prediction task, we select $N$ consecutive years of data (e.g., 2000-2004) from a yearly-partitioned dataset $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_N, ..., \mathcal{D}_{|\mathcal{D}|}\}$ to construct input features to our model, where $\mathcal{D}_k$ represents all records from the $k$th year in the dataset. Then, we construct prediction labels from $\mathcal{D}_{N+1}$, the first year following the selection of input data (e.g., 2005).

Individual records are vectors of $F$ categorical features in the form $x = <x_1, ..., x_F>$. For model training, each categorical feature is “one-hot encoded” by converting each categorical feature value into a unique vectorized representation. For example, a feature with 3 categories (e.g., “red”, “green”, and “blue”) would encode each of the values as $<1,0,0>$, $<0,1,0>$, and $<0,0,1>$, respectively. This encoding is also known as a “one-of- $k$” encoding. Most statistical models require numeric input and assume an ordinal relationship between values, thus encoding variables in this way provides the advantage of allowing such a relationship to exist, at the cost of higher-dimensional input vectors.

From the dataset records, we generate two varieties of features and sequences. The first variety concerns CBSA-level features that aggregate patient records for a specific CBSA region. The second variety concerns combined patient-CBSA features and couples patient-level features with the aggregated CBSA-level features. For our final model, we utilize only feature sequences; however, we generate individual features to compare against non-sequential baseline models and a non-sequential version of our proposed model.

1) CBSA-level Features and Sequences: For each yearly partition in our dataset, $\mathcal{D}_k$, we further partition our data by selecting all records in $\mathcal{D}_k$ that were generated in a specific CBSA, denoted $\mathcal{D}_{k,c}$. For each partition $\mathcal{D}_{k,c}$ consisting of all admission records associated with CBSA $c$ and year $k$, we compute aggregated feature vectors, $\sigma_{k,c} = <\bar{x}_1, ..., \bar{x}_F>$, by taking the mean for all variables across all records in the partition, as shown in Figure 2a. In addition, each aggregated feature vector is concatenated with the year and all previous years admissions totals, with admission totals for year $k$ in a CBSA $c$ denoted by $C_{k,c} = |\mathcal{D}_{k,c}|$. Thus, the resulting CBSA-level feature vectors have the form: $\sigma_{k,c} = <\bar{x}_1, ..., \bar{x}_F, k, C_{1,c}, ..., C_{k,c}>$

For sequential models, CBSA-level feature sequences are constructed for each CBSA by concatenating in ascending yearly order each CBSA-level feature vector for all considered years. Specifically, a sequence $\mathbf{X}_c$ for years $k$ through $k+N$ is defined as $\mathbf{X}_c = <\sigma_{k,c}, \sigma_{k+1,c}, ..., \sigma_{k+N,c}>$

2) Combined Patient-CBSA Features and Sequences: In addition to CBSA-level features, we consider patient-level features in order to provide more information to our model. For each patient admission record in the considered input data partitions, we generate a patient-level feature vector $x$ and concatenate the patient-level feature vector with a CBSA-level feature vector $\sigma_{k,c}$ corresponding to the (year, CBSA) pair in...
each feature vector in which, in non-sequential models, we assign a positive label (1) to the fully-connected layers, we apply a 25% dropout [12] layer before the final classification layer. All fully-connected layers utilize Rectified Linear Unit activation functions, while the classification output layer uses a sigmoid activation function. In this work, we refer to this proposed neural network model as NonSeq P-NN, or simply P-NN. For our proposed sequential model (Seq P-NN), we extend this base model by inserting two additional LSTM layers of size 125 units prior to the dropout layer.

C. Baseline Models

We consider two baseline models to compare against our proposed RNN model: a logistic regression classifier and a support-vector machine (SVM).

**Logistic Regression**: Logistic Regression is a statistical model widely used to model dependencies between predictor variables (features) and a binary response variable. In logistic regression, probabilities are formed for the response variables with respect to the varying predictor variables, with the model seeking to maximize the log-likelihood of these probabilities to decide the most-probable outcome given the set of inputs.

**Support-Vector Machine (SVM)**: Support-Vector Machines (SVMs) [10] are supervised learning models that represent feature vectors as points in space and seek to divide categories of data (according to labels) by a maximal gap/margin. Specifically, SVMs learn to separate categories of data according to this principle by training on provided data, and predict unseen inputs by determining which side of the gap the input falls on according to the parameters learned by examining the training data. Importantly, SVMs are effective at separating data that does not exhibit linear relationships by learning non-linear separation criteria.

V. EXPERIMENTS AND EVALUATION

We conduct two types of experiments: non-sequential and sequential. In each experiment type, we define training and testing subsets of our data, where the training set is used to train model parameters and the testing set is held-out data used to evaluate the performance of models. In each experiment, the training set consists of all feature vectors/sequences within the $N$ years of selected input data, and the testing set consist of all feature vectors/sequences from the $N + 1$’th year, where $N$ is varied for different experimental runs.

A. Non-Sequential Prediction Experiments

In the non-sequential experiments, feature vectors are supplied individually along corresponding labels to train the model, and the goal is to correctly predict the assigned label;
for unseen testing data. For the first experiment, we provide feature vectors to train each model from $N = 5, 10,$ and 15 years of data. In this experiment, we compare 3 variants of input features to our models: (1) CBSA-level features only; (2) all past-years admissions features only; and, (3) CBSA-level features with all past-years admissions features (as described in Section IV-A.1). For each value of $N$, each model is provided training data beginning from the start year $N_{\text{start}}$ through year $N_{\text{start}} + N - 1$, and then evaluated on year $N_{\text{start}} + N$. More specifically, each model is evaluated on the first year following the last year on which it was trained. For example, a training data point $x_{k,c}$ for past-years admissions features only for a specific CBSA ($c$) and sample year ($k$) in the training range ($N_{\text{start}}$ to $N_{\text{start}} + N - 1$) would provide a vector of past admissions totals representing admission counts in that CBSA for each year leading up to the sample year, while the corresponding training label for evaluation is represented as a 1 or 0 if admission counts increased/decreased in the sample year itself from the prior year. All combinations of $k$ and $c$ with $k$ in the training range are provided as training data, while all combinations of $k$ and $c$ where $k = N_{\text{start}} + N$ are provided as testing data. This segregation of training and testing data is utilized to simulate the practical setting in which only previous years' admissions data is available for forecasting future admissions rates. Further, the purpose of this experiment is to measure the impact of each feature type (CBSA-level and past labels) in comparison with the joint feature representation we propose.

Figures 3 and 4 demonstrate the results from our binary classification prediction experiments, comparing our base neural network architecture (NonSeq P-NN) against SVM and logistic regression for the $N = 5$ and $N = 10$ cases, varying the start year of the input data as the data permits. For $N = 15$, we report results in Table II.

### TABLE II

<table>
<thead>
<tr>
<th>Input Feature Variants</th>
<th>SVM</th>
<th>LogReg</th>
<th>NonSeq P-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBSA-features</td>
<td>0.5702</td>
<td>0.5220</td>
<td>0.6080</td>
</tr>
<tr>
<td>Past-years’ features</td>
<td>0.5094</td>
<td>0.4969</td>
<td>0.5283</td>
</tr>
<tr>
<td>CBSA + past-years’</td>
<td>0.5639</td>
<td>0.5011</td>
<td>0.5849</td>
</tr>
</tbody>
</table>

### B. Sequential Prediction Experiments

In the sequential experiments, feature sequences are supplied alongside corresponding labels to train the model, and the goal is to correctly predict the assigned label for unseen testing data that occurs in the year following the last year used to train the model. Again, we compare the same 3 variants of input features to our models: (1) CBSA-level features only; (2) all past-years admissions features only; and, (3) CBSA-level features with all past-years admissions features. In addition
to these feature variants, we also compare the effectiveness of our sequential model (Seq P-NN) against the same model trained with combined patient-CBSA features (SeqPatient P-NN) as described in Section IV. We provide CBSA-level feature sequences to train each model for $N = 5, 10,$ and $15$ years. As in the non-sequential experiments, we evaluate at each value of $N$ for all possible values of start year $N_{\text{start}}$.

Figures 5 and 6 show comparison results from our binary prediction experiments, comparing our proposed sequential models Seq P-NN and SeqPatient P-NN against our non-sequential model P-NN for $N = 5$ and $N = 10$. For $N = 15$, we report results in Table III.

### Table III

| Sequential Classification Accuracy ($N = 15$, Start Year 2001) |
|---|---|---|
| CBSA-features  | 0.6080 | 0.6373 | 0.6709 |
| Past-years’ features  | 0.5283 | 0.5975 | 0.6331 |
| CBSA + past-years’ features  | 0.5849 | 0.6457 | 0.6835 |

### VI. Discussion

In this section, we discuss the results from our experiments and propose explanations for our findings.

#### A. Non-Sequential Prediction

In our non-sequential experiments for binary classification of substance use admissions increases, we compared logistic regression and SVM baselines against our proposed non-sequential neural network architecture, NonSeq P-NN. In examining our results, we first look at the impacts of our feature variants.

For experiments with 5 years of training data (Figure 3), we note an exceptional variance for all models in the case where only CBSA-level features are provided (Figure 3a). The proposed NonSeq P-NN model achieves significant gains over both logistic regression for the 2006 start year; however, our NonSeq P-NN model vastly underperforms both baselines for input data starting the following year, 2007. This marked variance highlights the important difficulty for prediction models of generalizing trends outside of the available training data, and that short-term “snapshot” features are not enough on their own to provide this generalizability. In contrast, our NonSeq P-NN model exhibits much more stable prediction accuracies when trained with only past-years’ features (Figure 3b), consistently outperforming both baseline models across all start years and highlighting the added benefit of temporal features. More interestingly, when pairing these two feature variants together (Figure 3c) our model is able to leverage the additional information to improve prediction accuracy. In the case of start years 2001 and 2011, the joint CBSA and past-years’ features allow our NonSeq P-NN model to...
markedly improve prediction accuracy over both of the feature variants independently. Both SVM and logistic regression models exhibit highly varying prediction accuracies regardless of feature inputs, with slight improvements in both cases when past-years’ features are provided. Unlike NonSeq P-NN, the baseline models are unable to fully utilize the joint features to their advantage.

Moving to 10 years of training data (Figure 4), both baseline models as well as the NonSeq P-NN model demonstrate more stable and improved prediction results. This improvement in prediction results is a direct result of the increased amount of training data provided. Further, we notice similar feature variant impacts as in the $N = 5$ setting, where the combination of CBSA and past-years’ features outperforms either feature variant on its own. In all cases, our proposed NonSeq P-NN outperforms the baseline models in predictive accuracy.

Overall, our proposed NonSeq P-NN model is able to outperform the logistic regression and SVM baselines; however, the observed predictive accuracies do not far exceed a random prediction model, achieving 60% average classification accuracy for both $N = 5$ and $N = 10$. This result suggests that more powerful prediction models (such as sequence models) are necessary to extract more information from the underlying data.

B. Sequential Prediction

In our sequential experiments for binary classification of substance use admissions increases, we compare our non-sequential P-NN model (NonSeq P-NN) with a CBSA feature-based sequential model (Seq P-NN) and a combined patient-CBSA feature-based sequential model (SeqPatient P-NN).

For experiments with $N = 5$ and $N = 10$ years of training data (Figures 5 and 6), we notice a similar additive effect on performance of our feature variants, with the combined CBSA and past-years’ features outperforming both variants on their own. This additive benefit highlights the importance of incorporating geographic features into our training process, as it provides more contextual information for the model to learn and generalize from. In all experimental runs, our proposed SeqPatient P-NN model outperforms both Seq P-NN and our standard non-sequential P-NN model, achieving 70% classification accuracy in the best case (Figure 6c, start year 2004). SeqPatient P-NN outperforms the other models due to to the added patient-level information supplied by the combined patient-CBSA features. By introducing more training samples, we allow SeqPatient P-NN to extract significant information from the individual patient records that is not captured by the aggregated CBSA feature vectors alone.

Along the same line, a practical deployment of a prediction model should seek to incorporate as much information as possible to facilitate stronger predictive accuracy. For $N = 15$ (Table III), we include all available TEDS-A admissions data to train our model and achieve 68.4% prediction accuracy. This performance exhibits a stark improvement from the non-sequential models in which our best-performing model, P-NN, achieved only 58.5% classification accuracy.

C. Results Summary

From our experiments, we highlight four important contributions toward predicting substance use admission rates. First, the introduction of geographic (CBSA) features provides both non-sequential and sequential models with contextual information that assists in more accurately predicting future admissions trends. Second, neural network models demonstratively outperform standard baseline classification models in this context by more successfully utilizing all available features. Third, predicting trends necessitates the use of historical information, and the performance of our sequential models showcases the predictive power that is gained by encoding temporal information with RNNs. Lastly, forecasting trends remains a difficult problem when limited temporal information is available; however, we show that by combining individual records with aggregated feature sequences we can significantly increase the performance and robustness of our prediction models at several timescales.

VII. Conclusion and Future Work

In this paper, we introduced a Recurrent Neural Network model for predicting substance use admission rate increases across geographic entities in the U.S using the TEDS-A dataset. In the best performing case, we are successfully able to predict if substance use admissions rates will increase/decrease year-over-year in 70% of considered geographic regions. In considering all available admissions data from TEDS-A (2000-2016), we achieve 68.4% classification accuracy in predicting 2016 admission rates across considered regions. Our models emphasize the importance of both contextual geographic information and individual admissions records in predicting admissions rates.

In conclusion, we highlight several avenues for future research. The task of predicting substance use admissions rates is inherently difficult given the large number of causal societal factors (e.g., financial status, social environment, life experiences, etc.) that affect the disposition of a single individual or community. For this reason, incorporating additional resources such as substance overdose rates, economic data, social media information, and other sources may provide further contextual information to augment our geographic aggregates. In addition, encoding more geographical structure such as distances between geographic regions could help capture how trends diffuse and spread around geographic centers.

References


