Proceedings of Machine Learning Research 106:1–16, 2019 Machine Learning for Healthcare

Predicting Sick Patient Volume in a Pediatric Outpatient Setting using Time Series Analysis

Grace Guan GRACEGUAN@PRINCETON.EDU
Barbara E. Engelhardt BEE@PRINCETON.EDU
Department of Computer Science
Center for Statistics and Machine Learning
Princeton University
Princeton, NJ, USA

Abstract

Reducing patients' medical wait times by improving resource and staffing allocation is an important area of focus in hospital operations management. Two ways to decrease wait times are to adjust staffing or to limit the number of non-urgent visits to reflect a predicted volume of sick patients. Currently, this problem has been approached by both generalized linear models and time series models, and has mainly been researched in the context of adult emergency departments. We analyze sick visit data over a nine year period from one pediatric group (PG) that serves over 30,000 sick infants, children, and adolescents yearly in a walk-in and appointment-based out-patient clinic. The PG currently schedules staff and well-child appointments assuming a constant number of sick visits daily despite weekly and seasonal cycles in the data. We develop time series models to estimate the volume of sick patients that the PG can expect on any given day, so that clinicians can be allocated and the number of well-child appointments scheduled in advance can be adjusted according to predictions. First, we find that recurrent neural network (RNN) models are able to capture the seasonality of the data and perform substantially better than state-of-the-art models, including constant predictions. Next, we find that previous days' data can be used to perform outbreak detection by identifying error outliers. Lastly, we find improvements in prediction when modeling sick patients as a mixture of disease types, because disease types are concentrated differently throughout the year. Resource allocation based on these findings can be expanded upon to reduce wait time by improving staffing at pediatric emergency departments and outpatient clinics.

1. Introduction

Medical wait times are economically inefficient as they represent lost productivity for patients. In a study of 21.4 million outpatient electronic health records from over 2,500 practices in 2013, the median wait time was 4 minutes, a fifth of visits had wait times longer than 20 minutes, and half of those visits had wait times longer than 30 minutes (Oostrom et al., 2017). Time spent waiting to see a doctor influences a patient's health care experience and can lead to dissatisfaction with the medical establishment.

Lengthy clinical wait times are a symptom of sub-optimal processes within the health care system. Long wait times are caused by over-crowding and under-staffing of medical facilities that exceed peak volume in times of high demand. But the problems of over-crowding and under-staffing can be ameliorated. Specifically, models can be used to analyze the effect of staffing and scheduling on reducing wait times, as well as predict future staffing needs. In the case of British Columbia Cancer Agency's ambulatory care unit, simple configuration changes in resource allocation and scheduling achieved a 70% reduction in patient wait times (Santibanez et al., 2009).

Our paper focuses on building models to predict sick patient volume in a pediatric outpatient setting, so that staffing and well-child appointments can be adjusted accordingly to minimize patient wait time while not having more on-duty clinicians than necessary. The current literature in predicting patient volume varies from descriptive statistics to advanced time series models; however, only one study focuses on pediatric care. Pediatric medicine may differ from general medicine in terms of how physician time is spent, how medications are prescribed, and how infectious disease spreads through the population, among other attributes (Wiley et al., 2002; Mossong et al., 2008). The one study focusing on an urban pediatric emergency department uses descriptive statistics to show that the median daily volume peaked on Mondays and in the winter (Krinsky-Diener et al., 2017). Compared with non-holidays, the authors found that holidays had fewer patient visits. Our study differs in that the clinic is a pediatric outpatient clinic in small suburban setting where most of the patients attend one of a handful of local schools with synchronized schedules.

Most of the literature focuses on linear and time series models for predicting patient volume, though model performance depends on the data. Calegari et al. (2016) forecast daily emergency department (ED) visits in Porto Alegre, Brazil, finding that the seasonal autoregressive integrated moving average (SARIMA) model performed best for predictions of urgent patients. Similarly, Luo et al. (2017) forecast daily hospital outpatient visits for a large hospital in Chengdu, China, finding that the SARIMA model performed best when combined with an exponential smoothing (SES) model. SARIMA has already been used for malaria outbreak prediction, among other medical disease prediction tasks involving time series (Permanasari et al., 2013). On the other hand, Jones et al. (2008) and Marcilio et al. (2013) find that, for their emergency department (ED) visit data, generalized linear models (GLMs) without time series perform better than SARIMA. It is still an open question whether time series or GLMs are more appropriate for prediction of sick patient volume in specific medical time series tasks.

While the focus of related work has been on linear models, Jones et al. (2008) attempted to use neural networks to forecast daily patient volumes, finding that those models were difficult to interpret and were less accurate than linear models. Neural networks are worth re-investigating as a decade has passed since this work, and many advances have been made in non-linear time series modeling and the interpretability of neural networks.

1.1. Technical significance

Here, we create neural network models to investigate the effects of non-linear and recurrent layers on the time series prediction of sick patient volume in a pediatric outpatient setting. In doing this, we expand our feature space to include weather and calendar features as in past literature (Batal et al., 2008; Holleman et al., 1996; Marcilio et al., 2013). Moreover, we decompose the appointment data by ICD category into visits with and without a respiratory diagnosis. This allows us to investigate outbreaks of specific types of infectious diseases.

We compare our prediction results to state-of-the-art linear models for different forecasting tasks on held-out time windows on data from the same pediatric group.

1.2. Roadmap

The structure of the paper is as follows. First, we describe the pediatric outpatient data and its unique challenges to predictive modeling in Section 2. In Section 3, we describe the models and time scales used to predict the number of sick-child visits the PG can expect on a given day. Next, Section 4 presents the results of our experiments on different prediction tasks. Section 5 concludes with directions for future work. The full code used for the experiments in this paper, including saved model weights, is available online.[1]

2. Pediatric outpatient clinic data

2.1. Clinical background

We work with an anonymous suburban Pediatric Group (PG) in the tri-state area. Currently, the PG sees approximately 30,000 patients a year across multiple office locations. Over the past decade, the PG has received many negative reviews on Yelp, a site where customers can freely review businesses. The negative reviews are concerned with billing practices, wait times, and front office staffing. Moreover, personal communications with clients of the PG reveal frustrations with the availability of well-child appointments and wait times. Thus, applying our approaches to the PG may improve patient experience.

The PG has three different types of appointments: (1) walk-in sick-child visits on non-holiday weekdays, for acute issues that need to be seen within 24 hours. These appointments are similar to urgent care visits and can be booked no more than 24 hours in advance; (2) well-child visits, which are annual check-ups that can be booked up to 3 months in advance; and (3) non-physician based visits, such as flu shots and lab tests. There are a fixed, constant number of well-child visits (2) each day. Besides a 20-minute time blocked off for responding to phone calls, the remainder of a pediatricians time is intended to be used for walk-in, sick-child visits (1). In this project, we ignore non-physician based visits (3) as they can be handled by nurses and lab technicians rather than by full-time pediatricians.

There is a need to accurately predict the number of sick visits per day, as the number of well visits made available in advance should be adjusted according to the predicted number of sick visits. On a day where there are more sick visits than predicted, too many well visits would have been booked, leading to longer wait times for patients, more rushed patient-doctor interactions, and an overall diminished patient experience. Alternately, on a day with fewer sick visits than predicted, there will be gaps in physicians schedules that could have accommodated well visits. We focus on minimizing appointment overbooking rather than underbooking.

2.2. Clinical data

The PG provided day-level patient visit data from January 1, 2010 to September 20, 2018 from their electronic health record software, with each data point representing a single sick

[1] Available at: https://github.com/guanzgrace/predicting-patient-volume

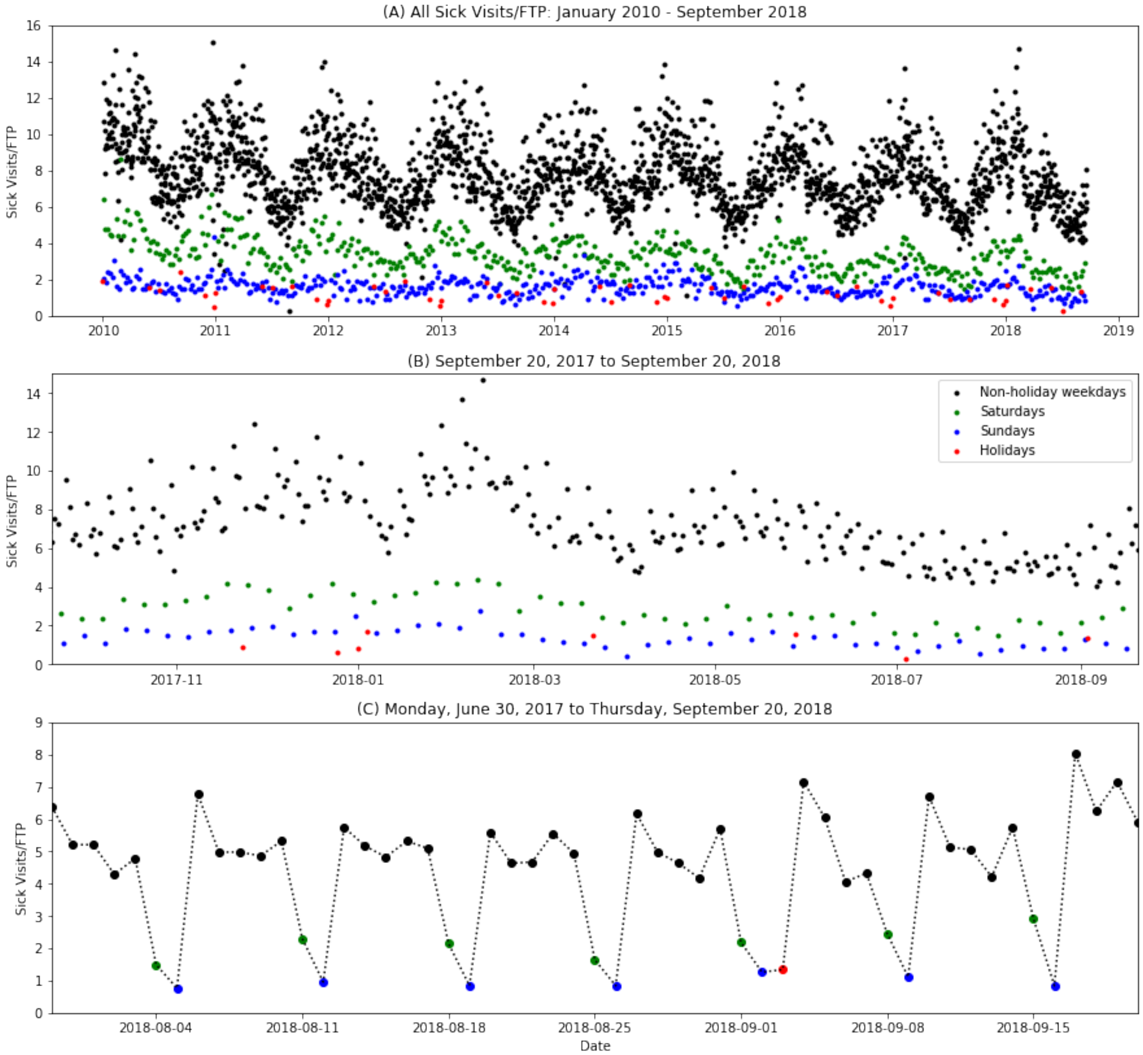


Figure 1: Daily walk-in sick visits per full time pediatrician for different time intervals. Panel A) January 2010-September 2018 (8.8 years); Panel B) September 2017-September 2018 (1 year); and Panel C) June 2018-September 2018 (7.5 weeks). The x-axis is the date; the y-axis is the number of sick-child appointments per full time pediatrician.

visit. The PG did not provide well-appointment data. Each data point contained billing code and name, ICD-9 or ICD-10 diagnosis code and name, transaction date, and location of appointment. To protect patient anonymity, time of visit was not recorded, nor were patients given identification numbers.

To account for changes in the number of offices and size of the practice as a whole, we consider the daily number of sick visits across all offices per full-time pediatrician (FTP) rather than the daily raw number of sick visits. The **number of full-time pediatricians**

(FTPs) represents the total number of full-time pediatricians hired on payroll, not the number of pediatricians present on a given day. The "full-time" aspect of FTP indicates one employed pediatrician working for a full 40-hour work week. For example, if there were two pediatricians scheduled to work part time at 20 hours a week each, this would count as only one FTP. Further, the "pediatrician" aspect excludes nurses, lab technicians, and front desk assistants. The number of FTPs is increasing, approximately once per year. As such, the number of FTPs can be used as a proxy for the growth of the practice.

The PG visit data exhibit obvious seasonal and cyclical trends (Figure 1). As one would expect, there are more sick visits in the winter compared to the summer (Figure 1, Panel A). Further, within each week, there appear to be reproducible patterns: Monday has the most visits, followed by Tuesday, Friday, Wednesday, Thursday, Saturday and Sunday, making a "u" shape every week (Figure 1, Panel C, dashed line to emphasize temporality). Zero well-child appointments are scheduled on Saturdays, Sundays, and holidays, yet there are still a non-zero amount of walk-in visits on weekends and holidays (Figure 1, most obvious around January 1, 2018 in Panel B). Holidays and weekends have similar sick visit per FTP volume. For example, Labor Day, which fell on Monday, September 3, 2018, has patient volume of similar magnitude to the previous two days (September 1 and 2), which were weekend days (Figure 1, Panel C with Labor Day denoted in red).

2.3. Non-clinical data

Historical local weather data from the two closest weather stations to the PG were obtained from the National Oceanic and Atmospheric Administration (NOAA) website[2]. Additionally, national and local holiday data were respectively obtained from the Pandas United States Federal Holiday Calendar[3] and from local charter and public school websites. We recoded holidays manually as the PG operates normally on some federal and school holidays.

2.4. Encoding features in the data

One-hot encoding creates a binary random variable for each value of a categorical variable. Because the data contained weekly and seasonal cycles encoded in categorical variables, we included one-hot encoded features for: **day of week**, dichotomous variables for each of Monday through Sunday; **week of year**, dichotomous variables for each of Weeks 1 through 52, through which we believed seasonality of the data would be captured; and **whether or not the practice was closed**. We term these days "holidays," and they include New Year's Day, Memorial Day, July 4, Labor Day, Thanksgiving, and Christmas Day.

Weekly dichotomous variables provided more information and less error compared to seasonal dichotomous variables. School break variables did not have much impact, so they were not included. Further, to account for weather, we included normalized continuous variables for: **temperature**, **wind**, **precipitation**, and **snow**, averaged between the two closest weather stations. For models in which it would be possible to see the previous day's true values in predicting the next day's value, we introduced **daily lag features** as well. Lag features allow pure time series problems to be converted into supervised learning problems. For example, **one day lag** would have the label of day $X-1$ input as a feature for predicting

[2] From: https://www.ncdc.noaa.gov/cdo-web/

[3] From: https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html#timeseries-holiday

day X. Similarly, **seven day lag** would have the labels of days $[X-7, X-6, \ldots, X-1]$ as seven different features for predicting day X.

2.5. Prediction label choices

We predicted the daily **number of sick visits per full time pediatrician (FTP)**, as defined in Section 2.2. While we could have used the number of FTPs as a feature, we chose to use it in normalizing our sick visit counts because this allows our predictions to be appropriately-calibrated across data sets and time windows. Specifically, due to the time series nature of our data, the number of FTPs did not change in our test set (2018) but was larger than the number of FTPs throughout the training and validation sets, leading to poorly-calibrated unnormalized predictions of the 2018 data.

Respiratory diagnosis labeling. The International Statistical Classification of Diseases and Related Health Problems (ICD) is a medical classification standard set by the World Health Organization (2018) (WHO) that allows for accurate statistical analysis of various mortality and morbidity metrics. The PG switched from ICD-9 to ICD-10 diagnosis labeling in late 2015. Because there are few of each individual diagnosis, we clustered the ICD-9 and ICD-10 diagnosis labels into 18 distinct groups (Table 1) based on chapters of the ICD-10 manual (World Health Organization, 2010).

After modeling the number of daily sick visits per FTP, we saw that our predictions were often wrong when there appeared to be slight upticks in the number of sick visits, specifically on Mondays in the winter. Thus, we believed that calendar and weather variables could not capture all of the information presented to us in the data, so we tried to determine a patient mixture we could predict. We assumed that these slight upticks were representative of outbreaks. Outbreaks are normally viral infections that present with respiratory conditions, and are especially contagious in infants, children, and teenagers in school environments.

Thus, of interest to us were the visits that have a respiratory diagnosis (our label 8) in attempts to find a methodology that could help in identifying infectious disease outbreak. Therefore, two additional labels we chose were the daily **number of sick visits that had a respiratory diagnosis per FTP** and the daily **number of sick visits that had a non-respiratory diagnosis per FTP**; combining models predicting these two labels give us the ability to more precisely model respiratory disease outbreaks.

2.6. Our prediction problem: Daily number of sick visits per FTP

We aim to develop predictive models for the daily number of sick visits per FTP, the daily number of sick visits with a respiratory diagnosis per FTP, and the number of sick visits without a respiratory diagnosis per FTP, given calendar information (day of week, week of year, and holiday) and weather information. We test our time series model on different forms of prediction that use prior days' outcomes in prediction as well. Our goal is for our predictive models to perform better than the current state-of-the-art linear regression and time series models to bridge the gap between the inaccurately predicted number of sick visits and the actual number of sick visits in pediatric outpatient clinical data. To do this, our models take advantage of three predictive patterns in the data that impact the daily number of sick visits. Our models capture the cyclical seasonality of the data, identify outbreaks of infectious diseases, and separately model mixtures of disease types.

	ICD 10	**ICD 9**	**Abbreviated Diagnosis Group**	**% of visits**
1	A00-B99	001-139	infectious & parasitic diseases	6.54%
2	C00-D48	140-239	neoplasms	0.10%
3	D50-D89	280-289	diseases of the blood & blood-forming organs	0.14%
4	E00-E90	240-279	endocrine, nutritional & metabolic diseases	0.24%
5	F00-F99	290-319	mental & behavioral disorders	2.49%
6	G00-G99	320-389	diseases of the nervous system & sense organs	9.02%
7	I00-I99	390-459	diseases of the circulatory system	0.14%
8	J00-J99	460-519	diseases of the respiratory system	33.55%
9	K00-K93	520-579	diseases of the digestive system	2.56%
10	L00-L99	680-709	diseases of the skin & subcutaneous tissue	5.81%
11	M00-M99	710-739	diseases of the musculoskeletal system	2.13%
12	N00-N99	580-629	diseases of the genitourinary system	1.33%
13	O00-O99	630-679	complications of pregnancy & childbirth	0.00%
14	P00-P96	760-779	certain conditions of the perinatal period	1.58%
15	Q00-Q99	740-759	congenital anomalies	0.26%
16	R00-R99	780-799	symptoms, signs, & ill-defined conditions	17.48%
17	S00-T98	800-999	injury & poisoning	7.94%
18	other	other	other	8.68%

Table 1: Groupings used to categorize the ICD diagnosis codes. We model patients in group 8 (respiratory diseases) separately from all other diseases.

3. Methods

Current method (constant model). This model is comparable to the method used at the PG. We computed the average number of visits for five categories of days: (1) weekends and holidays in any season; weekdays in (2) spring, (3) summer, (4) fall, and (5) winter.

Linear regression (baseline model). Linear regression models follow the general form $y = X\beta + \epsilon$ where X is a matrix of predictor variables (Section 2.4) and y, β, and ϵ are vectors representing the response variable (number of sick visits per pediatrician), regression coefficients for the predictor variables, and independent Gaussian error terms respectively. Our linear regression model used the scikit-learn implementation.[4] We found that having one linear regression model for all of the data over-predicted weekend and under-predicted weekday visits. Thus, we trained two separate linear regression models, one on weekday data only and one on weekend data only, and combined the results of these two models.

Time series SARIMA (TS). The SARIMA time series model allows for forecasting of seasonal trends and moving average processes. We chose the appropriate parameters for our time series data (cycle length = 7, degree of differencing = 0, moving average window = 1, and seasonal order = 1). Our time series model, which contained all endogenous variables

[4] From: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

(Section 2.4), used the Statsmodels implementation.[5] We used Powell's method in fitting the model, as the maximum likelihood optimization failed to converge.

Non-linear ReLU neural network (NN). We used a neural network with an input dense layer of 64 units, and one hidden layer of 64 units, both with the non-linear ReLU activation function. We chose this architecture because it achieved the best performance on the validation set when predicting all sick visits (as opposed to purely respiratory or non-respiratory visits). The validation metric we used was mean squared error (VMSE). We trained this model over 1000 epochs, with an early stopping function that stopped training if VMSE had not decreased after 10 consecutive epochs, restoring the weights from the epoch with lowest VMSE. This model used a sequential TensorFlow implementation[6] with the commonly used Adam optimizer[7] with learning rate 10^{-4}.

Recurrent LSTM neural network (RNN). Recurrent neural networks (RNNs) can capture temporal behavior in data and have been used for generating output sequences by predicting each subsequent step using the output of the past step as input (Graves, 2014). RNNs have had success in clinical applications, successfully predicting diagnosis, medication orders, and visit time (Choi et al., 2016). One RNN architecture is Long Short-term Memory (LSTM), which mitigates the problem of storing past output data over longer periods of time (Hochreiter and Schmidhuber, 1997). We composed a RNN with two hidden LSTM layers[8] of 32 units. We chose this architecture in the same way we chose the NN architecture. We trained this model over 300 epochs, with an early stopping function that stopped training if VMSE had not decreased after five consecutive epochs, restoring the weights from the epoch with lowest VMSE. The input, forget, and output layers of this RNN used the sigmoid activation function. The cell state and hidden output state used the ReLU activation function. We implemented this model with TensorFlow.

3.1. Time scales for predictions

Our use case involves predicting staffing for well-appointments, which must be scheduled in advance. Hence, our main prediction task was predicting almost 10 months of held-out data from all of the training data. However, in actuality, there would be little need to predict so far into the future. Instead, staff planning follows a two-time-scale approach, where an initial round of staffing is done several months in advance, and additional staff can be called in the day before or day of. We were interested in seeing how much prediction performance would be improved if more recent data were used within both of these timescales.

Dynamic. Dynamic models forecast held-out data from January 1, 2018 to September 20, 2018, after being trained on data from January 1, 2010 to December 31, 2016. For the NN and RNN models, we validated on data from January 1, 2017 to December 31, 2017 (approximately 12.5% of the full data set). For the constant, baseline, and TS models, since there were no hyperparameters to tune for the models, we withheld the validation data; essentially, the models never saw any data from 2017.

[5] From: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

[6] From: https://www.tensorflow.org/api_docs/python/tf/keras/models/Sequential

[7] From: https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam

[8] From: https://keras.io/layers/recurrent/

One-Month-Ahead (OMA). OMA prediction uses all available past data to predict each month (i.e., 30 days ahead) in our test set. We implemented OMA for our NN and RNN models with the following walk-forward validation method. First, we compiled a base model using the saved weights from our dynamic prediction models. Next, we increased the learning rate to 10^{-3} and trained the base model on the validation set (all of 2017) for one epoch. We then used this model to predict the label of each day of January 2018. For future months, we trained the model for one epoch on the true values of each month before predicting the next month. According to this protocol, we train on January 2018, predict February 2018; and then we train on February 2018, predict March 2018; and so on.

One-Step-Ahead (OSA). OSA prediction uses the true value of each past day in predicting the next day's output. We implemented OSA for our TS, NN, and RNN models. The TS OSA model used the default implementation for forecasting within the StatsModel package. The NN and RNN OSA method was analogous to our OMA method, except with the walk-forward step size being one day rather than one month.

Limitations in our prediction models. Our OMA and OSA models are limited in terms of apples-to-apples comparisons and in their overall performance. Because walk-forward validation does not make use of a validation set, models that implement it are not directly comparable with models that use a validation set. Various means of trying to include validation data (using an extreme time window, such as 2010, as our validation set, or keeping 2017 as a fixed validation set but adding the former test data from 2018 to the training set, or walking forward the validation set) did not make practical sense. Thus, we do not compare our dynamic models to our OMA and OSA ones because the comparison cannot trivially be made fair. Next, using the saved weights of the base trained models is suboptimal for our OMA and OSA models. The ideal early stopping function used in training the base model should calculate VMSE the same way that mean squared error (MSE) is calculated for the test set. In the case of one month ahead and one step ahead, that would be through walk-forward validation. Thus, a model with an early stopping function that matches how MSE is calculated should improve our results; the error we report for our OSA and OMA NN and RNN models is thus an upper bound of those true errors.

4. Results

4.1. Evaluation approach

The models were evaluated by mean squared error (MSE). MSE measures the average of the squares of the errors, or the difference between the actual (Y) and predicted ($\hat{Y}$) values of n data points, as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \quad (1)$$

4.2. Evaluation

First, we wanted to see if our non-linear models could capture seasonality in the data. To do this, we quantified how each model performed on our dynamic prediction task of

Model (Days of Lag)		All Visits MSE	Resp. Only MSE	Non-Resp. Only MSE
Dynamic				
	Constant	2.447	0.671	1.095
	Baseline	1.578	0.474	0.906
	TS	1.727	0.505	0.941
	NN	1.355 (0.02)	0.489 (0.00)	0.754 (0.01)
	RNN	1.085 (0.01)	0.444 (0.00)	0.363 (0.01)
OMA				
	NN	1.245 (0.03)	0.463 (0.00)	0.430 (0.02)
	RNN	1.029 (0.01)	0.436 (0.00)	0.425 (0.01)
OSA				
	TS	0.701	0.255	0.326
	NN	0.780 (0.02)	0.324 (0.01)	0.314 (0.00)
	RNN	0.854 (0.01)	0.378 (0.00)	0.296 (0.01)
	NN (1)	0.693 (0.01)	0.254 (0.01)	0.311 (0.02)
	RNN (1)	0.767 (0.03)	0.288 (0.01)	0.293 (0.00)
	NN (7)	0.791 (0.03)	0.216 (0.01)	0.438 (0.02)
	RNN (7)	0.562 (0.01)	0.166 (0.00)	0.281 (0.00)

Table 2: The MSE of various models tested on almost 10 months of held out pediatric visits (January 1 to September 20, 2018). Standard deviation over five trials is included in parentheses for the non-deterministic NN and RNN models. *Days of Lag* represents the number of days prior to the prediction day whose labels are included as features in the model. For the respiratory only (Resp. Only) column, the models were trained, validated, and evaluated only on visits with a respiratory diagnosis. The same holds for the Non-Resp. Only column for visits with a non-respiratory diagnosis.

training on data from 2010-2016, validating or withholding data from 2017, and predicting almost 10 months of held out data from 2018. The constant, baseline, and TS models were linear, whereas the NN and RNN models were non-linear. For this task, none of the models had information about outbreaks or patient mixtures. The non-linear models performed substantially better than the linear baseline, TS, and constant models. Specifically, the NN and RNN models had respective average MSEs of 1.355 and 1.085 over five trials[9], with respective standard deviations of 0.02 and 0.01. The RNN and NN models' MSEs are substantially lower (one sided t-test compared to baseline MSE, $p \leq 7.68 \times 10^{-6}$ (NN), $p \leq 2.03 \times 10^{-8}$ (RNN) with $df = 4$) compared to 1.578, 1.727, and 2.447 for the baseline, TS, and constant models (Table 2, Dynamic × All Visits). The success of both the NN and RNN models, especially the RNN model, suggests that there are additional components of the data beyond calendar variables that are useful in prediction across time.

[9] We define a trial as an independent training-evaluation run.

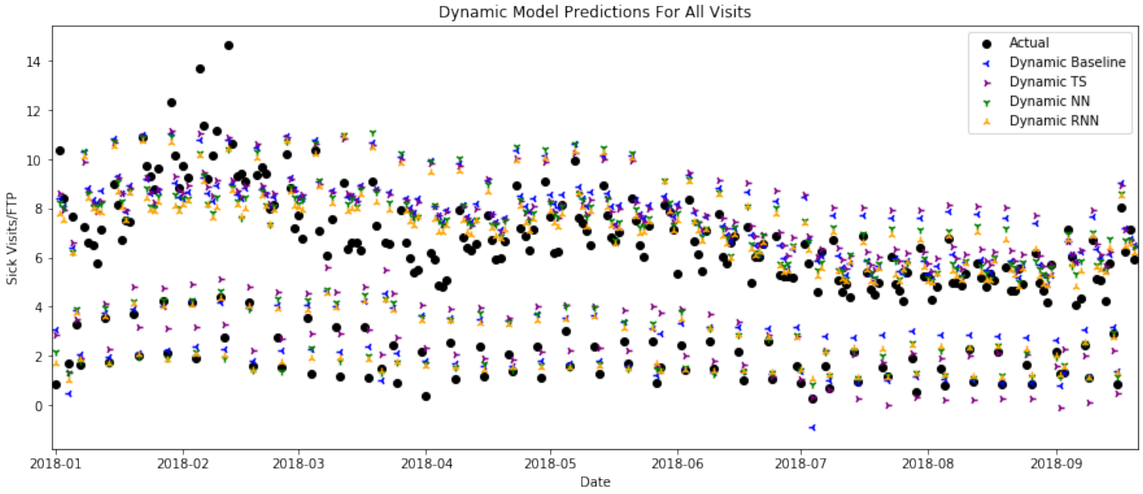
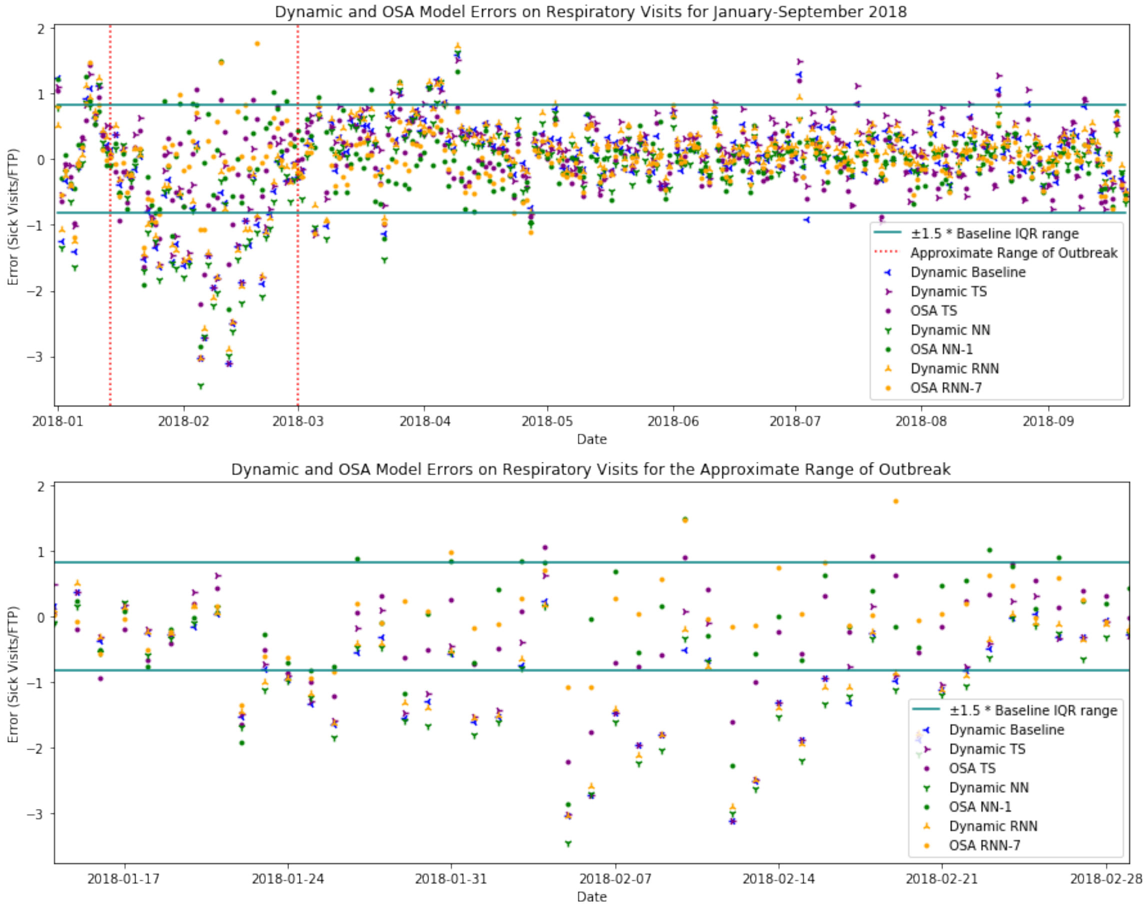


Figure 2: Scatter plot illustrating each dynamic model's predictions (Baseline blue, TS purple, NN green, RNN orange). The underpredictions in February and the overpredictions in March and April are clear. Time is on the x-axis; number of sick visits per FTP is on the y-axis.

The methodology of dynamic predictions 10 months in advance is not particularly helpful for scheduling clinicians for walk-in visits that cannot be predicted so soon. Specifically, dynamic models cannot adapt parameter values based on more recent data. Predictions from our dynamic models show that none of them accounted for an uptick of visits around February and a lull in visits in March and April (Figure 2). Thus, we wanted to see if giving our models additional, more recent data would improve performance. To do this, we turned to OMA and OSA predictions, which respectively use all past data in predicting the next month or day. OSA-based models perform substantially better than their respective OMA-based models—for the NN, average MSE over five trials improved from 1.245 to 0.780 when moving from OMA to OSA predictions; for the RNN, average MSE over five trials similarly improved from 1.029 to 0.854 (Table 2, All Visits). This suggests that our walk-forward validation-based models are appropriately capturing outbreaks as recent information appears to be more useful than all information from previous years. Further, both OSA- and OMA-based NN and RNN models perform better than their respective dynamic models, though the NN and RNN models are not directly comparable to the dynamic models because they see additional data (Table 2, All Visits).

In OSA predictions, since the NN and RNN models see all data up to and including the past day in predicting the next day, we wanted to see if adding additional lag features would improve the performance of our NN and RNN models compared to the TS model, which already uses a form of lag (cycle length = 7) as part of the SARIMA architecture. There are almost infinite choices of lag features (such as different amounts of past data or moving averages), but we chose to use straight lagged data of 1 day and 7 days as input features to our NN and RNN models to best match what the TS model uses. Adding lag improved the NN and RNN OSA models. The OSA TS model outperformed the NN and



Dynamic and OSA Model Errors on Respiratory Visits for January-September 2018

Dynamic and OSA Model Errors on Respiratory Visits for the Approximate Range of Outbreak

Figure 3: Scatter plot comparing each model's errors (Baseline blue, TS purple, NN green, RNN orange; Dynamic Y OSA ·) on respiratory visits only. Teal lines highlight outliers compared to the baseline. The underpredictions during the approximate range of the outbreak (in red) are nearly all due to respiratory visits.

RNN models without lag, with an MSE 0.701 compared to average MSEs of 0.780 and 0.854 over five trials for the NN and RNN models respectively (Table 2, All Visits). However, given one day of lag, the average MSE of the NN model improved to 0.693; similarly, given seven days of lag, the average MSE of the RNN model improved to 0.562. The improved performance of the NN and RNN models with more recent data as well as lag features suggests that previous days' data quantifying outbreaks is important for this prediction task. The unexpected peaks and valleys of our data is missed by the dynamic models.

Finally, we examined the ability of our models to handle the data as mixtures of disease types. To do this, we split the data into visits with a respiratory diagnosis and visits with a non-respiratory diagnosis, as it was the most common diagnosis category. For dynamic predictions, the RNN model was able to predict non-respiratory visits extremely well; however, the NN and RNN models performed about the same as the baseline on predicting respiratory

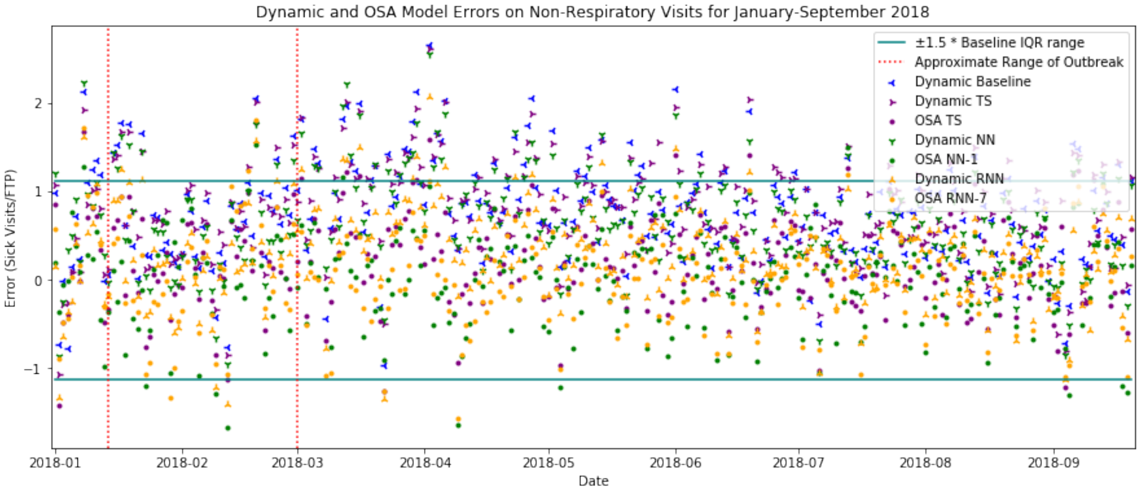


Figure 4: Scatter plot comparing each model's errors (Baseline blue, TS purple, NN green, RNN orange; Dynamic Y OSA ·) on non-respiratory. Teal lines highlight outliers compared to the baseline. The overpredictions throughout all of 2018 are largely non-respiratory.

visits (Table 2, Dynamic and Resp. Only and Non-Resp. Only). A graphical comparison of these models reveals that their errors may suggest outbreaks of certain types of diseases, as it becomes clear that respiratory visits account for most of the error throughout the approximate range of the outbreak (mid-January through the end of February; Figure 3, between red dotted lines). Compared to other models, the OSA RNN with 7 days of lag (orange dot) performed the best on this prediction task as it was closest to 0 and rarely outside 1.5 times the inter-quartile range of the baseline predictions' error throughout the outbreak range. Moreover, a graphical representation of the errors shows that nearly all of the overpredictions throughout 2018 were non-respiratory; however, the OSA RNN with 7 days of lag again performed best on this task with an average MSE over five trials of 0.281, compared to MSEs of 0.311 for the best NN model and 0.326 for the time series model (Figure 4).

The superior performance of the OSA RNN with 7 days of lag features suggests that certain models perform better on different prediction tasks, and these models could be combined over more data for greater effect. Overall, these results suggest that the combination of the three components that we built into the RNN models—nonlinear time series dynamics, outbreak detection through walk-forward validation and lag features, and mixture modeling to represent dynamics of different disease types—is essential for properly encoding these data.

5. Discussion

We analyzed nine years' worth of sick visit data from a suburban pediatric group in the tri-state area. Our goal was to improve their predictions of the volume of daily sick-patient visits so that they could adjust staffing and well-child visits accordingly. We predicted the number of sick visits per full time pediatrician using calendar and weather features. Comparing RNNs with generalized linear models, time series autoregressive models, and neural network models without recurrent layers, we found that the RNNs had the best predictive power for this problem. In particular, modeling patient sick visits as a mixture of patient subtypes in a nonlinear model, and including outbreaks of specific disease types enables actionable predictive capabilities beyond existing methods.

Our RNN models capture the cyclical seasonality of the data. While time series and generalized linear models are both widely used in predicting patient visits, one model has not been shown to be substantially better than the other. Though both of our non-linear models perform better than linear regression and time series models, the RNN model is most able to encode the temporal nature of the data. While RNNs proved effective in this suburban pediatric outpatient setting, more work needs to be done to determine whether settings that do not have a major seasonal component and do not generally have outbreaks (oncology, orthopedics) also find them useful for forecasting.

Using more recent information improves model performance. Our dynamic models performed poorly on test data from February 2018 due to an uptick of sick visits during that month. However, models using walk-forward validation as well as lag features are able to take advantage of more recent information to make predictions for the next day. Specifically, the OSA RNN with 7 days' lag features performed best out of all of our models in predicting all visits, including the uptick in February 2018.

Distinct disease types contribute separately to the daily number of sick visits. The performance of our models was improved by predicting sick visits with a respiratory diagnosis separate from sick visits with a non-respiratory diagnosis. Future work should investigate disease mixtures of other types, such as our category 9. Category 9, diseases of the digestive system, includes norovirus, another common infectious disease among pediatric patients. Our models can be extended with additional data and more fine-grained modeling of patient subtypes. More accurate prediction of patient volume 30-60 days in advance would allow for further improvements in staffing and availability of well-child appointments.

Acknowledgments

All of the data used in this study came from the generous pediatric group, who has asked to remain anonymous. Compute credits were provided through AWS Educate. Funding and conference travel were generously provided by the Princeton University Center for Health and Wellbeing's Global Health Program and the Center for Statistics and Machine Learning at Princeton University. GG thanks Yannis Karakozis and Jack Burdick for helpful comments. BEE was supported by NIH R01 HL133218 and an NSF CAREER AWD1005627. BEE is on the SAB for Celsius Therapeutics and Freenome, and is currently employed by Genomics Plc.

References

Holly Batal, Jeff Tench, Sean McMillan, Jill Adams, and Phillip S. Mehler. Predicting Patient Visits to an Urgent Care Clinic Using Calendar Variables. *Academic Emergency*

Medicine, 2008.

Rafael Calegari, Flavio S. Fogliatto, Filipe R. Lucini, Jeruza Neyeloff, Ricardo S. Kuchenbecker, and Beatriz D. Schaan. Forecasting Daily Volume and Acuity of Patients in the Emergency Department. *Computational and Mathematical Methods in Medicine*, 2016.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *Proceedings of Machine Learning for Healthcare*, 2016.

Alex Graves. Generating Sequences With Recurrent Neural Networks. 2014.

Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

Donald Holleman, Renee Bowling, and Charlane Gathy. Predicting daily visits to a waik-in clinic and emergency department using calendar and weather data. *Journal of General Internal Medicine*, 1996.

Spencer Jones, Alun Thomas, R Scott Evans, Shari J Welch, Peter Haug, and Geregory Snow. Forecasting Daily Patient Volumes in the Emergency Department. *Academic Emergency Medicine*, 2008.

Miriam Krinsky-Diener, Konstantinos Agoritsas, Jennifer H. Chao, and Richard Sinert. Predicting Flow in the Pediatric Emergency Department: Are Holidays Lighter? *Pediatric Emergency Care*, 2017.

Li Luo, Le Luo, Xinli Zhang, and Xiaoli He. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. *BMC Health Services Research*, 2017.

Izabel Marcilio, Shakoor Hajat, and Nelson Gouveia. Forecasting Daily Emergency Department Visits Using Calendar Variables and Ambient Temperature Readings. *Academic Emergency Medicine*, 2013.

Jol Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*, 2008.

Tamar Oostrom, Liran Einav, and Amy Finkelstein. Outpatient Office Wait Times and Quality of Care for Medicaid Patients. *Health Affairs*, 2017.

Adhistya Erna Permanasari, Indriana Hidayah, and Isna Alfi Bustoni. SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence. *International Conference on Information Technology and Electrical Engineering*, 2013.

Pablo Santibanez, Vincent S. Chow, John French, Martin L. Puterman, and Scott Tyldesley. Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation. *Health Care Management Science*, 2009.

James Wiley, Susan Fuchs, Sarah Brotherton, Georgine Burke, William Cull, Janet Friday, Harold Simon, Ethan Jewett, and Holly Mulvey. A comparison of pediatric emergency medicine and general emergency medicine physicians' practice patterns: Results from the Future of Pediatric Education II Survey of Sections Project. *Pediatric Emergency Care*, 2002.

World Health Organization. *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. World Health Organization, 2010. URL https://icd.who.int/browse10/2010/en.

World Health Organization. International Classification of Diseases (ICD) Information Sheet. *World Health Organization*, 2018. URL https://www.who.int/classifications/icd/factsheet/en/.