Predictive Modeling for Telemedicine Service Demand
Agni Kumar¹, Nancy Hung¹, Yuhan Wu², Robyn Baek³, Amar Gupta⁴

Affiliations: ¹Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA; ²Biological Sciences Department, Wellesley College, Wellesley, MA, USA; ³SOC Telemed, Reston, VA, USA; ⁴Institute for Medical Engineering and Science & Computer Science, Artificial Intelligence Lab, and Electrical Engineering and Computer Science Department, MIT, Cambridge, MA, USA

Corresponding Author: Agni Kumar, Massachusetts Institute of Technology, Departments of EECS, Mathematics, Email: agnik@mit.edu

Keywords: Alteplase, Evidence-Based Care, Predictive Modeling, rt-PA Telemedicine, Stroke, Teleneurology, t-PA

Section: Original Clinical Research

Introduction: Emergency teleneurology care has grown in magnitude, impact, and validation. Stroke is a leading cause of death in the United States, and timely treatment of stroke results in better outcomes for patients. Telenurology provides evidence-based care to patients even when a board-certified neurologist is not physically on site. Determining staffing demand for telemedicine consultation for a specific period of time is an integral part of the decision-making activities of providers of acute care telemedicine services. This study aims to build a forecasting model to predict consultation demand to optimize telemedicine provider staffing. Such forecasting models acquire added importance in emergency situations such as the current COVID-19 pandemic.

Materials and methods: This study trained consultation data of SOC Telemed, a private telemedicine provider, from 411 hospitals nationwide and involving 97,593 incidents of consultations. The forecasting model analyzes multiple characteristics, including hospital size (number of beds), annual volume, patient demographics, time of consultation, and reason for consultation.

Results: Several regression techniques were used to demonstrate a strong correlation between these features and weekly demand with $R^2 = 0.7821$. Reason for consultation in the past week was the strongest predictor for the demand in the next week with $R^2 = 0.7899$.

Conclusion: A predictive model for demand forecasting can optimize telemedicine resources to improve patient care and help telemedicine providers decide how many physicians to staff. The goal of the forecasting model is to improve patient care and outcomes by providing physicians timely and efficiently to meet the consultation demand.
The ability to predict demand and calculate expected volume allows telemedicine providers to schedule physicians in advance. This mitigates the clinical risk of excess patient demand and long waiting time, as well as the financial risk of scheduling a surplus of physicians.

Telemedicine is the delivery of healthcare services in situations where the physician and patient are not at the same geographic location. Telemedicine relies on three closely coordinated components: healthcare tasks performed by humans on-site, healthcare tasks performed by humans off-site, and healthcare tasks performed by computers without direct human involvement.

Teleneurology is a medical specialty within telemedicine. It can exist as a local hub or facility connected electronically to a larger hospital, neurologist, or specialist to provide consultations. For example, for SOC Telemed, an American telemedicine company, the audiovisual conferencing equipment consists of a camera with pan, tilt, and zoom capabilities, allowing the neurologist to examine and speak to the patient in a hospital. Teleneurology consultations include diagnosing and treating headaches, dementia, strokes, multiple sclerosis, and epilepsy. The hospital’s emergency department physician refers patients to a teleneurologist by calling SOC Telemed to initiate a consultation.

In spite of advances in stroke treatments, stroke remains the fifth most common cause of death and the leading cause for long-term adult disability, worldwide. Brain damage is directly correlated with time, because a patient with ischemic stroke typically loses 1.9 million neurons, 14 billion synapses, and 12 km of myelinated fibers every minute that blood is not flowing into the brain. Approximately, 87% of acute strokes are ischemic—a situation involving a blood vessel carrying blood to the brain being blocked by a blood clot. Patients with ischemic stroke are generally good candidates for intravenous (IV) alteplase (Acteplase®, tissue plasminogen activator, t-PA, rt-PA), a thrombolytic agent that helps reverse or prevent disability from ischemic stroke if administered within 3 hours—a guideline recommended timeline window for thrombolytic therapy according to the American Stroke Association. Alteplase IV is the only thrombolytic agent approved by the Food and Drug Administration for acute ischemic stroke. The time between stroke and patient receiving alteplase is inversely correlated with chances for recovery with no or minimal deficits. However, if misdiagnosed, administering alteplase in a hemorrhagic stroke patient poses severe risks to the patient’s life.

High patient numbers, limited specialized expertise, and geographic determinants remain major barriers that negatively impact implementation of evidence-based stroke management. At the same time, the shortage of neurologists is only exacerbating the existing barriers. Between 2012 and 2025, the demand for and the supply of neurologists are estimated to increase by 3,260 and 1,694, respectively. The supply of neurologists was already short by 1,814 neurologists in 2012, and by 2025, the shortage will increase to 3,380 neurologists. Teleneurology consultations serve an important role in bridging the shortage of neurologists and expanding access to care, so that patients can receive life-saving diagnosis and treatments in a timely manner.

Teleneurology forecasting differs from traditional emergency demand forecasting in that teleneurology providers need to provide service 24/7 to accommodate patients’ needs at any given time, with some requiring a response time of 5 minutes or less. Teleneurology provides
the hospital and patient immediate access to expertise around alteplase administration, considerably shortening door-to-needle time. Door-to-needle time, which is defined as the time between a patient entering the hospital to when the “clot-busting” alteplase drug is administered, is one of the key measurements in neurology. The benchmark “door-to-needle” time for stroke is set at 60 minutes, while many centers have exceeded this benchmark with median times ranging from 20 to 51 minutes.\textsuperscript{10–15} Teleneurology is demonstrated to be associated with reductions in “door-to-needle” time for alteplase.\textsuperscript{16–18} Several studies reveal that teleneurology improves alteplase use and outcomes in ischemic stroke in rural areas and community hospitals.\textsuperscript{19–22}

Given that thrombolytic treatments are decided upon in an urgent manner, clinicians have shown growing interest in using teleneurology to treat stroke patients. Teleneurology can be initiated by the clinician at the hospital, and telemedicine enables an instantaneous and private conversation between neurologists, patients, and clinicians. Ideally, teleneurology operational and service hours should be provided at all times year-round. Determining how many physicians to staff for teleneurology consults in a given period of time is therefore an integral part of the decision-making activities of providers of acute care telemedicine services.\textsuperscript{18,23} Without knowing an expected teleneurology consult request volume, telemedicine providers are at the mercy of physicians’ availability and are blindly scheduling physician shifts to cover as many hospitals as possible for 24/7 coverage. In addition, physicians are staffed on a monthly cadence, which adds further complexity to the scheduling problem without foresight of the projected demand.

The availability of good demand forecasts can lead to more effective allocation of human and physical resources to minimize patient care delays and improve the overall quality of healthcare.\textsuperscript{24} Teleneurology particularly benefits patients who are located in remote or rural areas, to provide equal and timely access to optimal stroke services across geographic regions and the care continuum.\textsuperscript{25}

Existing models employed to forecast patient demand consist of regression-based forecasting models such as linear,\textsuperscript{26–28} logistic\textsuperscript{29,30} and Poisson regression.\textsuperscript{31} Apart from regression-based models, time series methods have been used to forecast patient demand. Within time-series analyses, Auto Regressive Integrated Moving Average (ARIMA)\textsuperscript{32–38} modeling and exponential smoothing (ES)\textsuperscript{39–41} are applied most commonly. Machine learning algorithms such as Support Vector Machine (SVM),\textsuperscript{42} neural networks,\textsuperscript{28,43,44} and Bayesian networks\textsuperscript{45} have also been used.

In particular, the models we propose in this paper are highly interpretable, as they use a small number of important input features based on \textit{a priori} knowledge that can be formalized as \textit{a priori} relevance, and incorporate simple internal processing involving understandable rules, instead of complex learning processes associated with neural networks. Due to the sparse nature of our data, we did not employ rigorous nonregression models, as this could lead to sparse data bias.\textsuperscript{46}

In summary, this work proposes a weekly and monthly forecasting model on teleneurology consultation demand. Regression techniques such as ridge and lasso regression are applied. The model elucidates factors that affect the return on investment for telemedicine services, thus providing a valuable insight to telemedicine service providers.

**MATERIALS AND METHODS**

Data from SOC Telemed, a private telemedicine provider, was used for this research. The dataset
covered 411 hospitals nationwide and 97,593 incidences of teleneurology consultations from July 2015 to January 2018, where each row of data was timestamped at the minute level. We aggregated the data on a weekly and monthly basis to create the training data used for our models. This dataset was collected through SOC Telemed’s documentation software and documented by the call center agents and neurologists.

When a consultation is requested by a hospital’s clinician, SOC Telemed’s call center agent receives the call and begins documenting information about the patient and reason for consultation. Then, SOC Telemed’s neurologist begins the teleconsultation with the patient and documents clinical notes, including diagnosis and treatment recommendations. Within teleneurology, 53,000 female patients (55%) received consultations, compared to 43,000 (45%) male patients. Geographically, the dataset shows a heavier representation of hospitals in the Southeast, accounting for 35% of the total number of clients. Consultations were predominantly focused on stroke, followed by transient ischemic attack (TIA), encephalopathy, seizures, and stroke requiring tPA administration.

Although alteplase stroke did not occur with high frequency, it is the condition with the highest acuity due to the time-sensitive need for precise care. If patients arrive at the emergency department too late, they cannot be given alteplase as the treatment would be less effective. No correlation was found between gender and tPA stroke. A normal-like distribution was found between age and tPA stroke, with median centered around the 61- to 70-year age group. Using polynomial regression fitting, alteplase administration and the months of the year follow a cubic relationship with $R^2 = 0.7821$, as shown in Figure 1. This demonstrates seasonality in our data on a monthly basis, which is an expected trend. A slightly higher influx of patients was observed for fall and winter months, compared with those of spring and summer months. The number of patients who experienced stroke in this population varies between 311 in March and peaks at 593 in December. Although there are historically mixed results on the relationship between seasonality and stroke occurrences, a meta-study cites that 48.4% of studies evaluated on stroke reported higher occurrences of stroke during fall and winter months, whereas 39% of studies observed the opposite trend. Thus, seasonality may be present in the data.

RESULTS

Ordinary least squares (OLS) linear regression and polynomial regression models were applied to search for correlations between features and demand. Ridge regression and the lasso were also tested to find the best fit model. Ridge regression was applied for fine-tuning the complexity of the model with a regularization term, while lasso regression suits sparse data by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, causing certain coefficients to be set to zero and reducing the problem to a simpler model that does not include those coefficients. Plots depicting the weekly
and monthly telemedicine service demand aggregated across all partnering hospitals, over a period of approximately 1 year, are shown in the figures below.

In Figure 2, the data are aggregated on a weekly basis and plotted against the number of consultations that occurred during that week. The weekly demand (Figure 2) ranges from around 440 consultations in early July 2015 to 1051 consultations in late January 2018, showing that demand for telemedicine services has steadily increased over this time period. The correlation between number of consultations and data collected on a monthly time interval (Figure 3) also demonstrates a similar upward trend. Consultations range from 2220 in October 2015 to 4415 in January 2018. Demand is noted to spike slightly around the holiday periods. In the monthly demand plot (Figure 3), dips and peaks are observed once or twice a month.

Figures 2 and 3 depict the weekly and monthly telemedicine service demand aggregated across all partnering hospitals, over a 1-year period. The weekly data and monthly data comprise aggregations of minute-level data collected between January 2017 and February 2018. Aggregations at the monthly level reveal the slight presence of seasonality within monthly demand due to the fluctuations on a monthly basis, in line with the trend associated with tPA administrations in Figure 1. Thus, across time, the tPA administration rate is fairly constant.

OLS regression output demonstrated a slope of 0.202485 with y-intercept value as 845.707773 and correlation coefficient as 0.0357 (Figure 4). The time axis is given in the days elapsed since the start of data collection in January 2017. The low correlation coefficient output prompted running a model with additional input variables. Several orders of fitted polynomials were explored. Training, testing, and cross-validation were then executed to generate accuracy scores,
whose values were no longer highly variable in successive iterations due to augmentations of independent variables set in the system. Care was taken to minimize overfitting.

OLS linear regression was used to find the coefficient values that maximize $R^2$ and minimize the sum of squared residuals. OLS gives a coefficient for each predictor provided, including terms with little predictive power, a model has low bias but high variance results. We have the potential to improve our model by trading some of that variance with bias to reduce our overall error. This trade comes in the form of regularization, in which we modify our cost function to restrict the values of our coefficients.

After experimentation, it was observed that a slight change to the model’s input arguments yielded relatively large changes in the outputted weights. Ridge (Figure 5) and lasso regressions (Figure 6) were set to a specific regularization ($\alpha$) to reduce this variation. In particular, lasso regression adds an additional

![Figure 5—Ridge coefficients as a function of the regularization.](image)

![Figure 6—Lasso coefficients as a function of the regularization.](image)
term to the cost function, namely, the sum of the coefficient values multiplied by a constant ($\alpha$). This additional term penalizes the model for having coefficients that do not explain a sufficient amount of variance in the data. It also has a tendency to set the coefficients of bad predictors to zero, which makes lasso useful in feature selection. Ridge regression also adds an additional term to the cost function, but instead sums the squares of coefficient values and multiplies it by some constant $\alpha$. Compared to lasso regression, this regularization term decreases the values of coefficients but is unable to force a coefficient to exactly zero. This makes ridge regression’s use limited with regard to feature selection, but it in turn is able to select groups of highly correlated features. When $\alpha$ is very large, the regularization effect dominates the squared loss function, causing the coefficients (each in a different color on the plot) to tend to zero. From these figures, we observe that as $\alpha$ tends toward zero and the solution toward OLS, coefficients vary widely. In both plots, the faster a coefficient is shrinking, the less important it is for prediction.

Features used to train the new regression models were: time zone, visit initiated, state, gender, age, service line, reason for consult, provider diagnosis, hospital type, bed count, stroke certifications by level, and total ER visits to predict telemedicine consultations (Figure 7). Running the linear regression on these features, most of the data points matched the predicted value with $R^2 = 0.7821$ (Table 1). Lasso regression on the same features yielded an $R^2$ of 0.7695, a value slightly less than that associated with OLS Linear Regression. Ridge regression on all inputs yielded slightly better results with an $R^2 = 0.7829$ (Table 2) and provided the best overall result. This result was expected because ridge regression penalizes incorrectly trained and labeled predictions.

![Figure 7—Visualization of relation between measured and predicted demand.](image)

### Table 1. Results for linear, ridge, and lasso regression models

<table>
<thead>
<tr>
<th>Regression</th>
<th>$R^2$</th>
<th>Rank of Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge</td>
<td>0.7829</td>
<td>1</td>
</tr>
<tr>
<td>Linear (OLS)</td>
<td>0.7821</td>
<td>2</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.7695</td>
<td>3</td>
</tr>
</tbody>
</table>

OLS, ordinary least squares.

Executing the weekly prediction linear regression model with just Provider Diagnosis, the model yielded $R^2 = 0.7433$. Using the same model with Service Line in place of Provider Diagnosis, the accuracy increased to $R^2 = 0.7698$. Weekly prediction with only Reason for Consult performed the best, even better compared to training all the features together with $R^2 = 0.7899$. Interestingly, weekly forecasting with Reason for Consult and Provider Diagnosis increased the model accuracy to $R^2 = 0.8057$. While it was expected that Reason for Consult and Provider Diagnosis yield good results, it was unexpected that the Service Line would also provide an equally good prediction (Table 2).

### DISCUSSION

This study was envisaged to predict future teleneurology consultation demand with the
The objective of optimizing hospital staffing to reduce both financial risk for telemedicine providers and clinical risks for patients. The analysis demonstrates a strong correlation between the month and week of a telemedicine consultation request and the predicted number of consultations using linear, ridge, and lasso regression models.

Furthermore, the model can be adjusted to address changes in environment. For example, COVID-19 has significantly diminished the volume of teleneurology consultation. Instead of using the past year’s data, we could shorten the time frame of data to as little as a couple of weeks to reflect and adapt to the rapidly changing environment.

The combination of input features, such as hospital characteristics, and consult characteristics, which include Provider Diagnosis, Service Line, and Reason for Consult, yielded forecasted consults demand, with 78% overall accuracy with ridge regression. Overall, this weekly and monthly prediction model informs how to staff providers on their monthly schedules for SOC Telemed’s needs.

Ongoing work includes improving the accuracy of the current predictive model by augmenting each training vector with the Service Line, Reason for Consult, and Provider Diagnosis. We will continue to work around dataset limitations, which include the dataset consisting entirely of data from one telemedicine provider and its clients, for a single practice, only over a 1-year time period. Ideally, we would like more

### Table 2. Key fields, example fields, and descriptions used in data analysis

<table>
<thead>
<tr>
<th>Field</th>
<th>Field Definition</th>
<th>Example Categories</th>
<th>Definition of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reason For Consult</td>
<td>Reason for initiating telemedicine</td>
<td>TIA</td>
<td>Transient Ischemic Attack</td>
</tr>
<tr>
<td></td>
<td>consultation</td>
<td>Acute ischemic stroke</td>
<td>Requires brain imaging for diagnosis within 60 min in order to intervene</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Migraine</td>
<td>Throbbing/pulsing head pain involuntary and rhythmic shaking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tremor</td>
<td></td>
</tr>
<tr>
<td>Provider Diagnosis</td>
<td>Clinical diagnosis from telemedicine</td>
<td>Stroke/tPA Stroke</td>
<td>Stroke that requires tPA administration</td>
</tr>
<tr>
<td></td>
<td>consultation</td>
<td>Seizure</td>
<td>Sudden, uncontrolled electrical disturbance in the brain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Encephalopathy</td>
<td>Brain degeneration likely caused by repeated head traumas</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Migraine</td>
<td>Throbbing/pulsing head pain</td>
</tr>
<tr>
<td>Service Line</td>
<td>Immediacy for consultation scheduling</td>
<td>Neuro Routine</td>
<td>Middle priority</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neuro Emergency</td>
<td>Highest priority</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neuro Routine —</td>
<td>Lowest priority</td>
</tr>
<tr>
<td></td>
<td>Scheduled</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TIA: transient ischemic attack, tPA: tissue plasminogen activator
information on the extent to which telemedicine was used in consults over various hospitals, patient demographics data (to analyze which hospitals get greater telemedicine service demand and if this is correlated with patient demographics), and notes of which partnering hospitals specialize in which service lines. Noting that the spikes and dips in demand over the year are not dramatic, we look to incorporate an anomaly detection component into the model, to search specifically for points and microtrends that impact the demand for telemedicine service the most at specific points over time.

Planned future work includes performing predictions on different time scales, such as a model that looks for trends on a daily cadence instead of weekly or monthly. Although the weekly and monthly models are more aligned with the real-world use case, daily fluctuations could facilitate insights on providing a more granular model for staffing resources.

Accurate personnel forecasting is more important in teleneurology settings because of the higher prevalence of alteplase administration in teleneurology, compared to traditional neurology. A case study on a hospital’s alteplase use showed that IV alteplase increased from three to 34 administrations from 2009 to 2018. This study found that the increase in consultations and IV alteplase coincides with the adoption of teleneurology. From April 2019 through March 2020, SOC Telemed’s teleneurologists administered alteplase to 3,271 patients out of 6,548 acute ischemic stroke patients who presented to emergency departments with time last known well of 4.5 hours or less, indicating a 50% alteplase administration rate. Therefore, teleneurology service is an effective way to provide emergency consultations to hospitals of different demographics. In rural areas, the alteplase administration rates have increased to 55% when teleneurology is offered as an option. In 2014, data from Massachusetts General Hospital (MGH) website showed that 207 out of 532 stroke consults over telemedicine received IV alteplase as compared to a 4% national rate.

Teleneurology provides a potential solution to address the shortage of neurologists, and the demand predictive model for consultations is helpful for teleneurology providers in this context. For neurology, estimates for the number of physicians required per 100,000 population ranges between 1.4 and 2.3. These ratios are used as a starting point to assess the community need for physicians. The supply of neurologists is predicted to grow by 11% between now and 2025, while demand is projected to grow by 16%. Teleneurology treatment is highly cost-effective, particularly for severe stroke. The incremental cost-effectiveness ratio is $108,363/QALY in the 90-day horizon and $2,449/QALY in the lifetime horizon. Teleneurology is also demonstrated to be more patient-specific, sensitive, and to have higher predictive values than traditional telephone-based consultations. Teleneurology services are performed as reliably as bedside neurologists.

Teleneurology is recommended by the American Heart Association and American Stroke Association in instances when specialists are not available in person. Based in part on the experience with the current COVID-19 pandemic, we believe strongly that these guidelines should and will be applicable to all situations. In fact, one of the co-authors of this paper is aware of a hospital that caters to potential cases of stroke both on on-site basis and
via telemedicine. The available response time for the neurologist to advise the patient (from the time that the request was initiated) is actually lower in the telemedicine case, as a neurologist is always on duty around the clock. In the other case, it takes time for the neurologist to walk from other duties to see the concerned patient or to come from home during nighttime.

This research could be used as a foundation to study telemedicine consultations involving other medical specialties. Furthermore, patient demographics data could be incorporated to identify hospitals with greater telemedicine service demand and to test whether demand is correlated with patient demographics. Considering that the spikes and dips in demand over the year are not dramatic, an anomaly detection component could be incorporated into the model to search for microtrends affecting telemedicine service demand at certain time points.

Acknowledgments
The authors thank Hai Tran, Chief Operating Officer (COO) and Chief Financial Officer (CFO) at SOC Telemed, for delivering a guest lecture for the Telemedicine and Telehealth for Enhancing Global Health course at Massachusetts Institute of Technology (MIT), for sparking brainstorming discussions on ways to enhance and optimize telemedicine consultation demand, and for providing valuable insights into the day-to-day operations of the largest US provider of acute care telemedicine services. We also thank Dr Rafael Palacios (Visiting Scholar, MIT Energy Initiative) for comments and suggestions on the evolving manuscript. This paper originated from a student paper that was written as part of the above course, developed and taught at MIT in 2018, 2019, and 2020, which provided a unique opportunity to learn from visionaries in industry, academia, and government on strategic, medical, policy, technical, business, legal, and economic aspects of telehealth. The authors are grateful for the inputs and suggestions from the diverse base of distinguished group of speakers and participants in the course, whose ideas have been incorporated in this paper.

Conflicts of Interest
Apart from the disclosure in the funding statement below, the authors confirm that there are no known conflicts of interest associated with this publication and that there has been no financial support for this work that could have influenced its outcome. During the past 18 months, Dr Gupta has been compensated as a consultant by the MITRE Corporation for keynote addresses on telemedicine at a closed-door session to senior officials of department federal government agencies in Washington, D.C., McDermott Will Emery for a keynote address on the subject of artificial intelligence (AI) in Healthcare: The Role of Medical Boards at an event co-organized by the Federation of State Medical Boards, the Business Incubators Accelerators of The Saudi Technology Development and Investment Company (via MIT) for a globally live-streamed keynote address, co-organized by MIT Hacking Machine, on “Using Telemedicine and Artificial Intelligence to Provide Better Healthcare for All: Opportunities and Challenges” and by Walmart Health for work on Telemedicine and Telehealth.

Funding Statement
Dr Amar Gupta is an employee of MIT and was partly funded from a research contract signed by Philips. Robyn Baek is a full-time employee of SOC Telemed.
REFERENCES


35. Afifal M, Yalaoui F, Dugardin F, et al. Forecasting the emergency department


Copyright Ownership: This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, adapt, enhance this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0.