Automatic Forecasting of Radiology Examination Volume Trends for Optimal Resource Planning and Allocation

Anton S. Becker[1] · Joseph P. Erinjeri[2] · Joshua Chaim[1] · Nicholas Kastango[3] · Pierre Elnajjar[4] · Hedvig Hricak[5] · H. Alberto Vargas[1]

Abstract

The aim of the study was to evaluate the performance of the Prophet forecasting procedure, part of the Facebook open-source Artificial Intelligence portfolio, for forecasting variations in radiological examination volumes. Daily CT and MRI examination volumes from our institution were extracted from the radiology information system (RIS) database. Data from January 1, 2015, to December 31, 2019, was used for training the Prophet algorithm, and data from January 2020 was used for validation. Algorithm performance was then evaluated prospectively in February and August 2020. Total error and mean error per day were evaluated, and computational time was logged using different Markov chain Monte Carlo (MCMC) samples. Data from 610,570 examinations were used for training; the majority were CTs (82.3%). During retrospective testing, prediction error was reduced from 19 to < 1 per day in CT (total 589 to 17) and from 5 to < 1 per day (total 144 to 27) in MRI by fine-tuning the Prophet procedure. Prospective prediction error in February was 11 per day in CT (9934 predicted, 9667 actual) and 1 per day in MRI (2484 predicted, 2457 actual) and was significantly better than manual weekly predictions ($p = 0.001$). Inference with MCMC added no substantial improvements while vastly increasing computational time. Prophet accurately models weekly, seasonal, and overall trends paving the way for optimal resource allocation for radiology exam acquisition and interpretation.

Introduction

Resource planning is a critical component of success in a radiology department. At a strategic level, the ability to accurately estimate future exam volumes is a crucial input to physician recruitment and retention strategies, which are especially essential in the context of increasing clinical demand and physician burnout [1]. At an operational level,

radiologists' roster is usually scheduled weeks to months in advance when the exact number and type of examinations to be performed are not known. "Coordinated visits," where a patient attends the radiology department on the same day as their clinic appointment, add an additional layer to the planning process. These visits are increasingly popular as a tool for improving patient satisfaction, but create additional demands for immediate radiologist availability in order to produce results in a timely manner. Finally, while some useful information can be derived from an institution's maximum "scanning capacity" (e.g., maximum number of exams that can be performed per scanner multiplied by the total number of scanners), the increasing recognition of the crucial value of re-interpretation of images submitted from outside institutions by subspecialized radiologists [2–5] places additional strain on a radiology department's ability to deal with the clinical workload in a timely manner.

Forecasting is the process of using past data to make predictions about future states, usually by mathematical/statistical methods analyzing longitudinal time-series data. The field has made great progress on a technical level in the past years [6] and has become increasingly available

✉ Anton S. Becker
beckera1@mskcc.org

1 Department of Radiology, Body Imaging Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA

2 Department of Radiology, Interventional Radiology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA

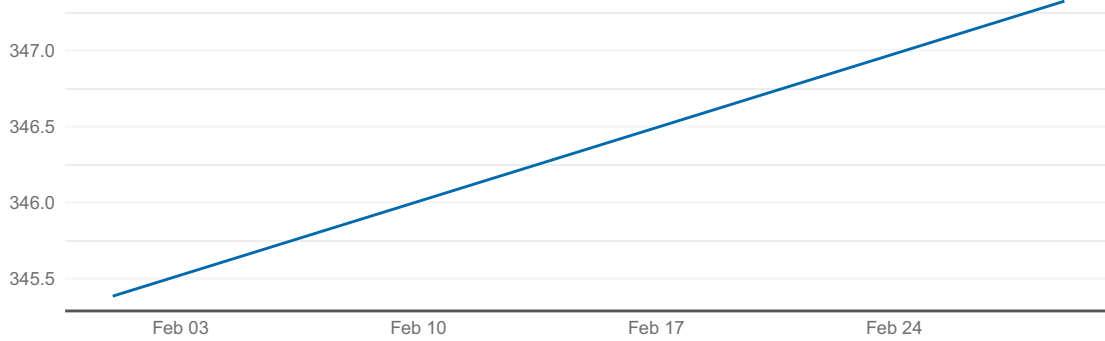3 Department of Strategy and Innovation, Memorial Sloan Kettering Cancer Center, New York, NY, USA

4 Department of Radiology, Informatics Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA

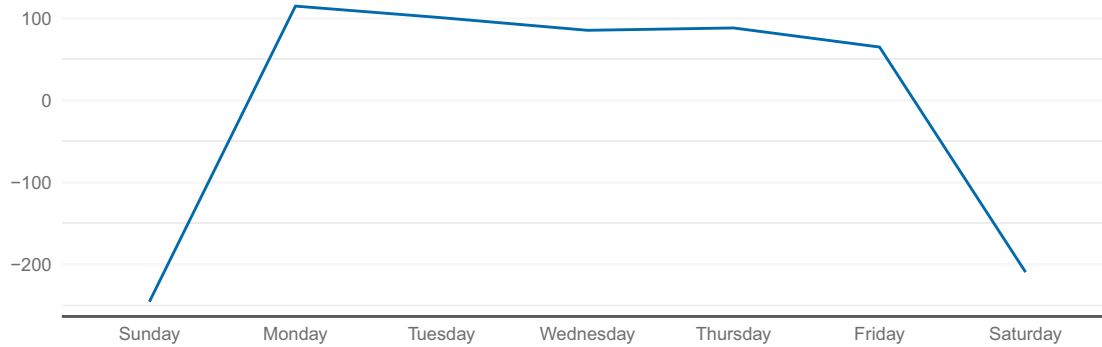5 Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

a) Number of examinations over time
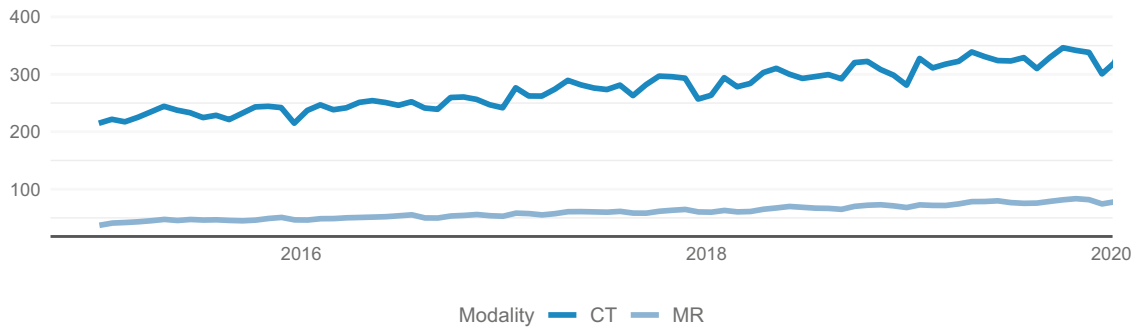
b) Overall model component

c) Weekly model component
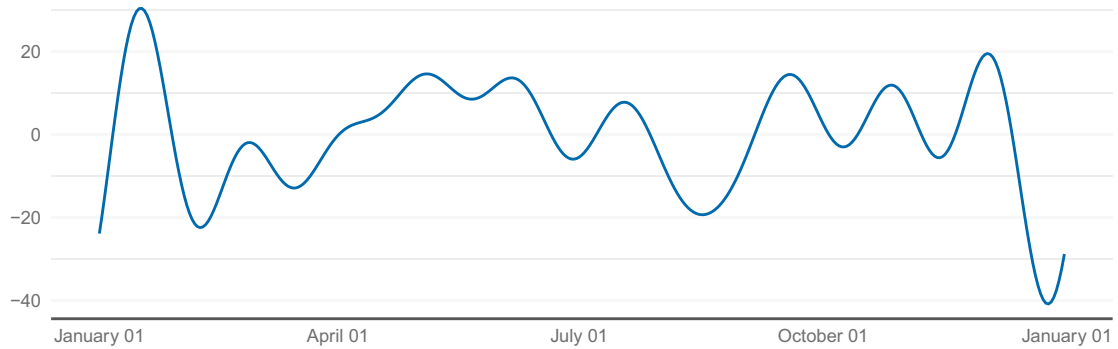

d) Yearly model component

◄**Fig. 1 a** Smoothed curve of examination volume over a 5-year period, showing an overall growth pattern as well as seasonal repeating patterns, which are detected by the model as shown in the trend decompositions in **b–d**. The *y*-axis represents the effect of the respective seasonal component on the predicted *y* value

and integrated within existing analytics software. These advances were fueled by a multitude of applications, for example, stock prices in financial forecasting, or user fluctuations and server-load for web-based computer applications in information technology companies [7]. Similar to these applications, the number of radiology exams shares some strong seasonal and periodic characteristics, some of which may be obvious (e.g., holidays, weekdays vs. weekends), some less so (e.g., seasonal weather fluctuations). This lead to the hypothesis that a forecasting algorithm modeling daily, weekly, monthly, and seasonal variations would be well-suited to predict radiology workload. The *Prophet* procedure, which is part of the Facebook AI portfolio, is an open-source forecasting algorithm designed for such tasks.

Hence, the purpose of this study was to prospectively evaluate the performance of the *Prophet* forecasting algorithm for radiological examination volume, with the ultimate aim to aid with radiologist coverage planning.

Material and Methods

Data Collection and Experimental Design

This study was Health Insurance Portability and Accountability Act (HIPAA)-compliant and Institutional Review Board-exempt as an institutional quality improvement project. From our radiology information system, we first extracted all examinations in the body imaging service over a 5-year period from January 1, 2015, through January 31, 2020. The sum of daily examinations without any individual patient identifiable information were retrieved on February 1, 2020, and served as a training dataset. The month of January 2020 was withheld during model training to validate and optimize the model.

The reference method we used to compare against *Prophet* was a weekly, manual procedure: A certain number of radiologists are on clinical shifts as a baseline coverage; then, worklists were checked every Friday evening and radiologists moved to or from clinical assignments the following week based on the volume of unread examinations.

Algorithm and Software

The Prophet algorithm is part of the Facebook (R) open-source AI portfolio (https://opensource.facebook.com/#artificial-intelligence). Prophet is a forecasting procedure based on an additive model fitting non-linear trends with daily, weekly, and yearly effects [8]. Furthermore, it allows for inclusion of relevant events such as major holidays. The Prophet implementation (v. 0.5) in R version 3.6.3 was used [9]. The full analysis together with non-patient-related, synthetic data is publicly available at https://github.com/ASBecker/exam_forecast. All computations were performed on a personal computer with a 3.5 GHz CPU (AMD (R) Ryzen Threadripper 1920X 12 Core) and 64 GB RAM.

Data Analysis

CT and MRI examinations were analyzed separately, since they are different assignments in our body imaging roster. Furthermore, the performance on in-house and external examinations (henceforth: "second reads") was analyzed separately and combined. Inpatients were not analyzed separately since they only represent a small fraction of the overall volume, have less underlying variability, and are covered by a constant number of radiologists due to the requirement for timely interpretation. Final performance assessment of the best selected model was performed after 4 weeks using the data from February 1 to 29, 2020. Performance was measured by (1) the total sum of prediction error and (2) the mean prediction error per day. Furthermore, we compared the residual sum of squares (RSS) between manual and *Prophet* forecast with a paired Wilcoxon signed-rank sum test. A two-tailed alpha of 0.05 was used to indicate significant differences.

Prophet allows for setting of various hyperparameters, most of which were left at the default settings after initial empirical trials. However, we evaluated the use of Bayesian inference using different numbers of Markov chain Monte Carlo (MCMC) samples on our prospective test set. We evaluated the incremental accuracy by total error and mean error per day as well as the computational time needed.

Results

Data

The total number of examinations was 610,566 (Fig. 1a), of which 543,386 were performed at our institution and 67,180 were second reads. The majority (82.3%) were CTs, constituting 502,708 examinations (9.6% second reads), whereas MRIs amounted to 107,858 (17.4% second reads). In the validation data (January 2020), there were 13,961 examinations, of which 12,102 were performed at our institution and 1859 were second reads. The majority (81.4%) were again CTs with 11,358 examinations (11.6% second reads). There were 2603 MRIs (20.9% second reads).

Model Selection

Data for the entire study period is shown in Fig. 1a. Seasonal fluctuations and continuous growth can be appreciated, as well as "dips" around major holidays. Individual components of the *Prophet* model are depicted in Fig. 1b–d.

Backtesting of data from the last month of the training period ("backtest-month") (Fig. 2) showed that the prediction tended to underpredict during the week and overpredict on weekends. Since there is a clinical and operational penalty when reports are delayed, a slight over-allocation of radiologists on a given day is generally favored over under-allocation. Thus, rather than using the predicted point estimate, the upper bound of the prediction 80% confidence interval during weekdays and the lower bound on weekends ("hilo-estimate") were selected. With this slight adjustment to the standard Prophet procedure, the total prediction error for CT was reduced from a total of -587 (mean -19 examinations/5.2% per day) to 50 examinations for the whole backtest-month (mean 2 examinations/0.4% per day), and for MRI was reduced from a total of -144 (mean -5 examinations/5.6% per day) to 21 examinations for the whole month (mean 1 examination/0.8% per day). A negative error means that more examinations were predicted than actually occurred, while a positive error means fewer examinations were predicted than actually occurred.

Model Evaluation: Forecast

The model was prospectively evaluated in the 4 weeks of February 2020 (test month), using all data until the end of January for training of the algorithm. Estimate of total CT examinations was 9942, and actual number was 9553 (mean error of 10 examinations/2.3% per day, illustrated in

Fig. 3); estimate of total MR examinations was 2485, and actual number was 2397 (mean error of 2 examinations/4.5% per day).

After the *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2 or COVID-19) pandemic had led to an initial decrease in clinical activity from March to May and a subsequent rebound effect in June and July, the same model (trained with data until January 2020) was again tested prospectively in August 2020: Estimate of total CT examinations was 10,326, and actual number was 10,009 (mean error of 10 examinations/3.2% per day); estimate of total MR examinations was 2610, and actual number was 2482 (mean error of 4 examinations/5.1% per day).

Manual examination prediction yielded a median underestimation of -14 (IQR -39.25–41.38), whereas the prophet prediction slightly overestimated the examination volume by a median of 7.1 (IQR -11.3–30) per day. The *Prophet* procedure was significantly more accurate than the manual forecast ($p = 0.001$).

Model Evaluation: Markov Monte Carlo Chain

Forecasting accuracy was evaluated on the test month (February 2020) with increasing number of MCMC samples from 250 to 32.000. Zero samples is the default setting at which a maximum a posteriori estimation is performed, which is computationally very efficient (< 1 s run time) but yields slightly less accurate forecasts. This setting was used in all forecasts above. The results of MCMC sampling are summarized in Table 1.

For CT, the forecast becomes more accurate when in-house exams and 2nd reads are combined. Beyond 2500 samples, there was no improvement in accuracy, and even this improvement was rather modest with 17 examinations
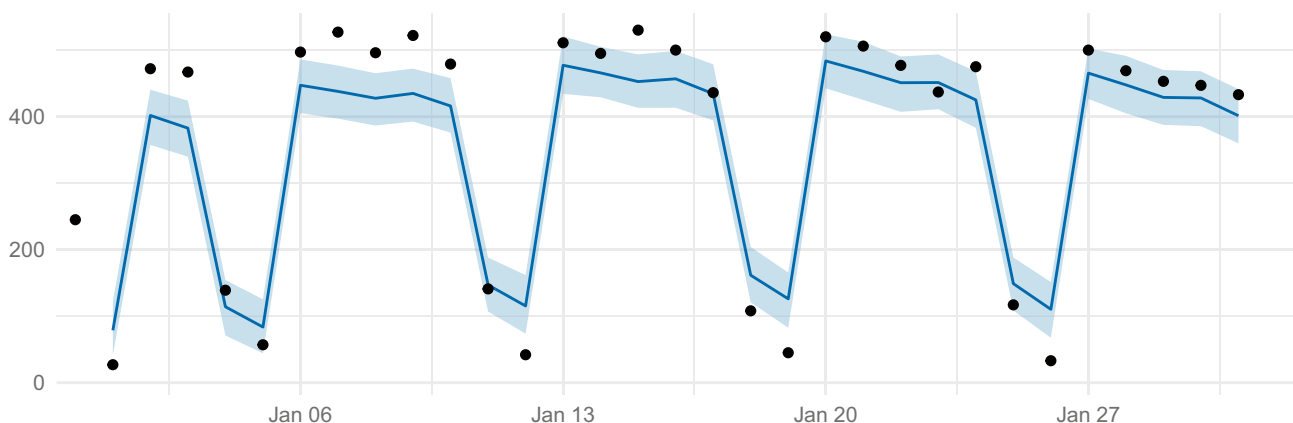

Fig. 2 Retrospective test for January predicted CT numbers (blue line) with 80% confidence interval (shaded blue area). Actual number of examinations on a given day are represented by the black points

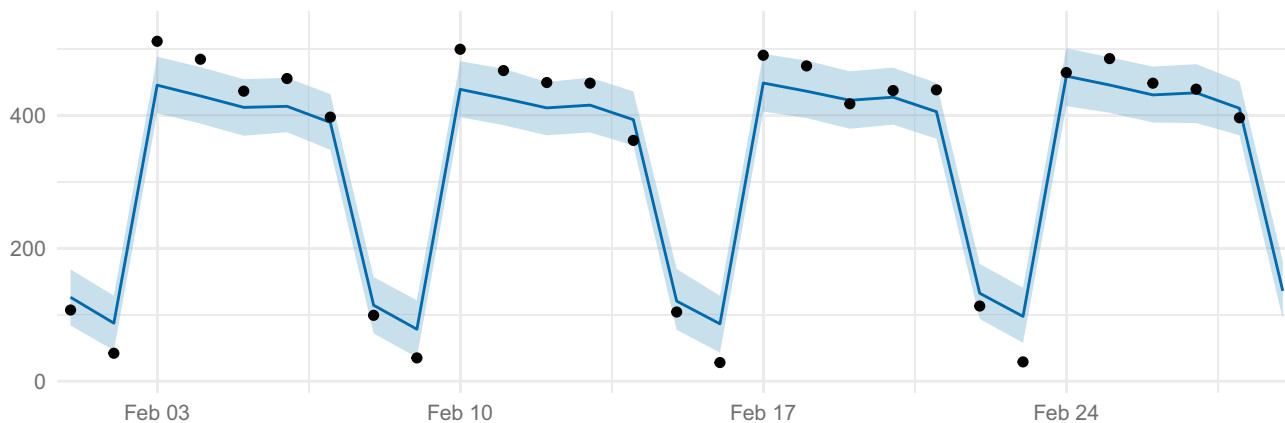## Evaluation: Daily examination prediction for February 2020

Fig. 3 Prospective evaluation for February predicted CT numbers (blue line) with 80% confidence interval (shaded blue area). Actual number of examinations on a given day are represented by the black points

(0.07%) difference over the whole test month (at an increased computation time from < 1 s to > 20 min). In MRI, there was overall no improvement beyond 500 samples; however, due to a general underprediction of in-house examinations and overprediction of 2nd reads, those two predictions would be complimentary and result in a lower overall error. The mild improvement in accuracy with a large number of MCMC samples in the in-house forecast comes at the price of long computation times increased by several orders of magnitude.

Discussion

In this study, we evaluated the use of the state-of-the-art forecasting *Prophet* algorithm to predict radiology examination volumes. We found that the algorithm captures weekly, seasonal, and overall trends and allows for better radiologist allocation compared to manual planning. Custom fine-tuning of the *Prophet* procedure allowed us to tailor the forecasts to favor slight overestimation, whereas MCMC-sampling vastly increased computation time while not substantially increasing performance.

Patient volume forecasting has been attempted in various settings, for example, notoriously erratic emergency room visits and admissions, with modest success [10]. Boyle et al. achieved a ~ 7% mean daily error of emergency room presentations, but only 11% error for daily admissions [11]. Similarly, Zhang et al. achieved a ~ 7% mean error [12] when modeling radiologic patient flow in an emergency room setting, whereas our fine-tuned *Prophet* procedure had a daily mean error between 2.3 and 5.1%. The slightly better performance is likely a compound effect of the more advanced technology used, the larger number of data points in our dataset, and the majority of our examinations being planned

outpatient appointments. In the radiological literature, the problem of no-shows and same-day cancellations has been thoroughly investigated in order to understand the reasons and propose new interventions to minimize them [13, 14]. While Zhang et al. have recently studied examination volume in the emergency room of a hospital [12], there are no studies using machine learning for radiology examination volume forecasting and resource planning on a departmental level.

Machine learning approaches fare better in high-volume settings with ample training data [15]. In our study, this was evident by the slightly lower performance when trying to predict second reads from outside CT examinations separately. *Prophet* produced more accurate estimates when in-house examinations and outside reads were combined. This may be due to some co-correlation between the two distributions; i.e., when overall activity at an institution (e.g., outpatient clinic visits, inpatient census) is higher, there will be both more in-house and outside examinations.

The *Prophet* procedure is designed specifically for time series forecasting tasks with strong periodic and seasonal effects [8]. However, it may underperform in settings with simple periodic (i.e., quarterly) effects but otherwise without strong seasonality [16]. Papacharalampous and colleagues directly compared temperature and precipitation forecasts of *Prophet* and other forecasting methods. They found that *Prophet* outperformed naïve forecasts and yielded comparable results to other state-of-the-art forecasting algorithms [17]. Furthermore, Zunic and colleagues recently demonstrated the versatility and robustness of the *Prophet* procedure by applying it in a real-world environment (retail business) [18]. Our results are in line these findings. We expect our results to be reproducible in other bottleneck problems in healthcare with seasonal fluctuations, such as

Table 1 Evaluation of forecast accuracy for CT and MRI examination volume with different numbers of MCMC samples. A negative error means that more examinations were predicted than actually occurred, while a positive error means fewer examinations were predicted than actually occurred

MCMC samples	Modality	In-house			Outside (2nd reads)			Combined		
		Total error	Mean error	Run time	Total error	Mean error	Run time	Total error	Mean error	Run time
0	CT	−205.68	−6.86	0.0	−127.67	−6.08	0.0	−227.71	−7.59	0.0
250	CT	−186.66	−6.22	2.4	−115.63	−5.51	5.5	−226.08	−7.54	2.6
500	CT	−201.87	−6.73	6.0	−114.08	−5.43	6.4	−218.49	−7.28	5.6
750	CT	−193.06	−6.44	9.0	−121.85	−5.80	6.9	−217.04	−7.23	8.4
1000	CT	−205.58	−6.85	11.8	−116.92	−5.57	9.2	−218.62	−7.29	14.3
1250	CT	−198.78	−6.63	18.1	−116.13	−5.53	11.0	−224.82	−7.49	15.6
1500	CT	−208.15	−6.94	17.2	−118.81	−5.66	12.2	−226.38	−7.55	16.6
1750	CT	−199.34	−6.64	18.1	−118.63	−5.65	13.8	−217.97	−7.27	18.1
2000	CT	−203.90	−6.80	23.6	−118.16	−5.63	15.4	−216.75	−7.23	22.4
2500	CT	−193.93	−6.46	24.4	−116.92	−5.57	20.2	−211.16	−7.04	26.3
3000	CT	−212.21	−7.07	30.8	−115.38	−5.49	23.0	−213.18	−7.11	30.6
3500	CT	−199.35	−6.64	32.9	−117.64	−5.60	26.3	−218.02	−7.27	34.3
4000	CT	−205.01	−6.83	40.5	−117.74	−5.61	29.5	−219.61	−7.32	39.1
8000	CT	−200.96	−6.70	75.3	−116.65	−5.55	55.5	−214.88	−7.16	72.5
32,000	CT	−203.52	−6.78	287.2	−117.26	−5.58	206.6	−216.84	−7.23	293.5
0	MR	333.24	11.11	0.0	−30.86	−1.47	0.0	362.22	12.07	0.0
250	MR	318.06	10.60	6.7	−26.85	−1.28	5.7	346.00	11.53	6.7
500	MR	314.46	10.48	8.1	−28.41	−1.35	5.2	345.82	11.53	8.0
750	MR	316.14	10.54	9.4	−27.65	−1.32	5.7	346.70	11.56	9.9
1000	MR	318.53	10.62	12.5	−28.81	−1.37	6.8	347.54	11.58	12.9
1250	MR	317.53	10.58	16.4	−28.96	−1.38	7.7	347.04	11.57	15.1
1500	MR	318.56	10.62	16.2	−28.07	−1.34	8.7	347.26	11.58	16.5
1750	MR	317.53	10.58	18.2	−28.38	−1.35	10.2	347.34	11.58	18.2
2000	MR	318.04	10.60	22.0	−29.15	−1.39	10.8	347.01	11.57	20.8
2500	MR	317.78	10.59	26.1	−28.40	−1.35	13.0	347.46	11.58	25.9
3000	MR	315.17	10.51	29.9	−29.01	−1.38	14.8	349.26	11.64	29.5
3500	MR	317.04	10.57	32.8	−28.38	−1.35	17.1	345.92	11.53	33.0
4000	MR	317.41	10.58	37.2	−27.96	−1.33	18.7	348.63	11.62	36.4
8000	MR	323.33	10.78	64.1	−28.64	−1.36	34.4	347.53	11.58	66.8
32,000	MR	322.90	10.76	258.3	−28.80	−1.37	138.8	348.02	11.60	253.2

surgical procedures or radiotherapy treatments. Lastly, these methods may also prove useful in allotting resources in academic research, for example, by forecasting occupancy of laboratory animal housing facilities or eligible patients for enrollment in clinical trials.

Limitations

There are several limitations that need to be considered. First, this was a single-center study. The robustness of the method and applicability to other settings will have to be investigated tailored to the individual institution's setting and the unique patient population they serve [19]. To facilitate this process, we provide ready-to-implement code in a public *GitHub* repository with sample data, which needs to be swapped in order to train the model to the new institutional data.

Second, we only investigated a single algorithm. Even more accurate forecasts may be possible with other methods, such as novel deep recurrent neural networks [20]. However, deep neural networks require more computational resources for training and implementation for the planning routine is more difficult [20]. More importantly, the accuracy of the current method was already suitable for resource planning without the use of such more computationally expensive tools.

Third, *Prophet* allows for limited addition of external information (i.e., holidays). Events that predictably lead to a "dip" in examinations such as maintenance of scanners would have to be manually factored in, unless they can be added as "custom holidays" (for example recurring annual conferences). This point is also relevant in the current situation: While this article is being written and under review, the COVID-19 pandemic is entering a second or third wave in many countries. In extraordinary situations such as these, of course, the information to predict scan volume lies completely outside of the historical data and cannot be modeled with traditional forecasting methods.

Conclusion

Prophet produces accurate radiological examination volume predictions and is a useful tool for clinical operations planning.

Declarations

References

1. Harolds, J.A.; Parikh, J.R.; Bluth, E.I.; Dutton, S.C.; Recht, M.P. Burnout of radiologists: frequency, risk factors, and remedies: a report of the ACR Commission on Human Resources. *Journal of the American College of Radiology* **2016**, *13*, 411–416.

2. Gollub, M.J.; Panicek, D.M.; Bach, A.M.; Penalver, A.; Castellino, R.A. Clinical importance of reinterpretation of body CT scans obtained elsewhere in patients referred for care at a tertiary cancer center. *Radiology* **1999**, *210*, 109–112, https://doi.org/10.1148/radiology.210.1.r99ja47109.

3. Wibmer, A.; Vargas, H.A.; Donahue, T.F.; Zheng, J.; Moskowitz, C.; Eastham, J.; Sala, E.; Hricak, H. Diagnosis of extracapsular extension of prostate cancer on prostate MRI: impact of second-opinion readings by subspecialized genitourinary oncologic radiologists. *American Journal of Roentgenology* **2015**, *205*, W73–W78.

4. D'Anastasi, M.; Lakhman, Y.; Miccò, M.; Scelzo, C.; Vargas, H.; Sosa, R.; Chi, D.; Abu-Rustum, N.; Sala, E.; Hricak, H. Mehrwert für Die Patientenversorgung Bei Der Zweitbefundung von Gynäkologischen MRTs Durch Subspezialisierte Radiologen. **2015**, *187*, RK412_2.

5. Woo, S.; Kim, S.Y.; Cho, J.Y.; Kim, S.H. Assessment of deep myometrial invasion of endometrial cancer on MRI: added value of second-opinion interpretations by radiologists subspecialized in gynaecologic oncology. *European radiology* **2017**, *27*, 1877–1882.

6. Spiliotis, E.; Makridakis, S.; Semenoglou, A.-A.; Assimakopoulos, V. Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research* **2020**, 1–25.

7. Tesauro, G.; Jong, N.K.; Das, R.; Bennani, M.N. A hybrid reinforcement learning approach to autonomic resource allocation. In Proceedings of the 2006 IEEE international conference on autonomic computing; IEEE, 2006; pp. 65–73.

8. Taylor, S.J.; Letham, B. Forecasting at scale. *The American Statistician* **2018**, *72*, 37–45, https://doi.org/10.1080/00031305.2017.1380080.

9. R Core Team *R: a language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019

10. Meldon, S.W.; Mion, L.C.; Palmer, R.M.; Drew, B.L.; Connor, J.T.; Lewicki, L.J.; Bass, D.M.; Emerman, C.L. A brief risk-stratification tool to predict repeat emergency department visits and hospitalizations in older patients discharged from the emergency department. *Academic Emergency Medicine* **2003**, *10*, 224–232.

11. Boyle, J.; Jessup, M.; Crilly, J.; Green, D.; Lind, J.; Wallis, M.; Miller, P.; Fitzgerald, G. Predicting emergency department admissions. *Emergency Medicine Journal* **2012**, *29*, 358–365.

12. Zhang, Y.; Luo, L.; Zhang, F.; Kong, R.; Yang, J.; Feng, Y.; Guo, H. Emergency patient flow forecasting in the radiology department. *Health informatics journal* **2020**, *26*, 2362–2374.

13. Rosenbaum, J.I.; Mieloszyk, R.J.; Hall, C.S.; Hippe, D.S.; Gunn, M.L.; Bhargava, P. Understanding why patients no-show: observations of 2.9 million outpatient imaging visits over 16 years. *Journal of the American College of Radiology* **2018**, *15*, 944–950, https://doi.org/10.1016/j.jacr.2018.03.053.

14. Speece, J. Reducing the impact of same-day missed appointments. *Journal of the American College of Radiology* **2019**, *16*, 824–829, doi:https://doi.org/10.1016/j.jacr.2018.12.003.

15. Halevy, A.; Norvig, P.; Pereira, F. The unreasonable effectiveness of data. *IEEE Intelligent Systems* **2009**, *24*, 8–12.

16. Kourentzes, N. Benchmarking Facebook's Prophet 2017.

17. Papacharalampous, G.; Tyralis, H.; Koutsoyiannis, D. Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysica* **2018**, *66*, 807–831.

18. Zunic, E.; Korjenic, K.; Hodzic, K.; Donko, D. Application of Facebook's prophet algorithm for successful sales forecasting based on real-world data. *arXiv preprint arXiv:2005.07575* **2020**.

19. Glover IV, M.; Daye, D.; Khalilzadeh, O.; Pianykh, O.; Rosenthal, D.I.; Brink, J.A.; Flores, E.J. Socioeconomic and demographic predictors of missed opportunities to provide advanced imaging services. *Journal of the American College of Radiology* **2017**, *14*, 1403–1411.

20. Sahoo, B.B.; Jha, R.; Singh, A.; Kumar, D. Long Short-Term Memory (LSTM) Recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica* **2019**, *67*, 1471–1481.