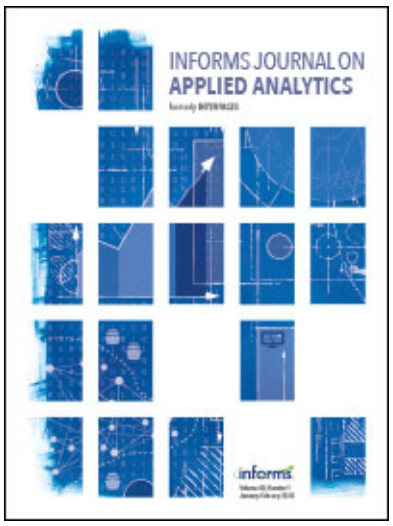

Interfaces

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Business Analytics Assists Transitioning Traditional Medicine to Telemedicine at Virtual Radiologic

Ersin Körpeoğlu, Zachary Kurtz, Fatma Kılınç-Karzan, Sunder Kekre, Pat A. Basu

To cite this article:
Ersin Körpeoğlu, Zachary Kurtz, Fatma Kılınç-Karzan, Sunder Kekre, Pat A. Basu (2014) Business Analytics Assists Transitioning Traditional Medicine to Telemedicine at Virtual Radiologic. Interfaces 44(4):393-410. https://doi.org/10.1287/inte.2014.0752

Interfaces
Vol. 44, No. 4, July–August 2014, pp. 393–410
ISSN 0092-2102 (print) | ISSN 1526-551X (online)

http://dx.doi.org/10.1287/inte.2014.0752

Business Analytics Assists Transitioning Traditional Medicine to Telemedicine at Virtual Radiologic

Ersin Körpeoğlu, Zachary Kurtz, Fatma Kılınç-Karzan, Sunder Kekre
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
{ersinkorpeoglu@cmu.edu, zkurtz@andrew.cmu.edu, fkilinc@andrew.cmu.edu, skekre@cmu.edu}

Pat A. Basu
Doctor on Demand, San Francisco, California 94105, patbasu@gmail.com

Virtual Radiologic (vRad), the largest teleradiology company in the United States, faces the difficult problem of matching more than 400 radiologists with time-varying and seasonal demand. In addition to the constraints that traditional medical facilities face, vRad is subject to supply and demand requirements that are unique to the teleradiology business environment. In this paper, we present a forecasting and capacity-planning model that more accurately assesses demand and plans system capacity to provide better service to vRad's customers. We discuss the underlying reasons for improvement and quantify the impact on vRad's entire system. We explain managerial insights that will help both vRad and other companies in the service sector with similar service-response requirements and demand patterns. We also highlight the implementation challenges our teams faced.

Keywords: business analytics; capacity planning; forecasting; optimization; healthcare industry.
History: This paper was refereed. Published online in *Articles in Advance* May 27, 2014.

With more than 400 radiologists, 2,700 client facilities, and an annual revenue of approximately $300 million, Virtual Radiologic (vRad) is the largest national provider of radiology in the United States. The company processes eight million radiology images annually. By integrating on-site presence with the power of cloud computing, vRad's value proposition is to deliver the best clinical results for its patients by deploying the radiologist with the most relevant specialized training in the least turnaround time, anywhere, anytime.

On a 24/7 basis, through customized high-technology equipment located at client facilities (i.e., clinics and hospitals), the vRad cloud receives medical images (hereafter jobs) and distributes them to be read by vRad radiologists throughout the United States. Figure 1 displays vRad's client and radiologist footprint. Radiologists are vRad's most expensive resource, but they are the cornerstone of its business and its most important asset in delivering its commitment to quality. As in many other medical practices, achieving the best care outcomes in radiology requires assigning each medical image to the right doctor. vRad radiologists are highly specialized, and 75 percent of them are trained in an array of subspecialties, including but not limited to neuroradiology, musculoskeletal radiology, and pediatric radiology. For many traditional hospital radiology departments, having such a wide array of radiologists with special training available at all times is impossible. By contrast, vRad radiologists work flexible hours in the comfort of their homes using their own specialized equipment. vRad optimizes capacity utilization by pooling supply and demand across a large base of clients and providers. In particular, by aggregating the demand of many facilities and pooling its radiologists, vRad is able to maximize the utilization of its radiologists' time in a way that an individual facility staffed with on-site radiologists cannot. Smart technology solutions, a large client base, and better capacity utilization via resource pooling underlie vRad's ambitious business model. Moreover, in 2011, it acquired NightHawk, its largest competitor, an acquisition that nearly doubled its client base and operational volume. Its high-technology orientation and large customer base allow vRad to achieve significantly higher productivity rates without sacrificing quality.

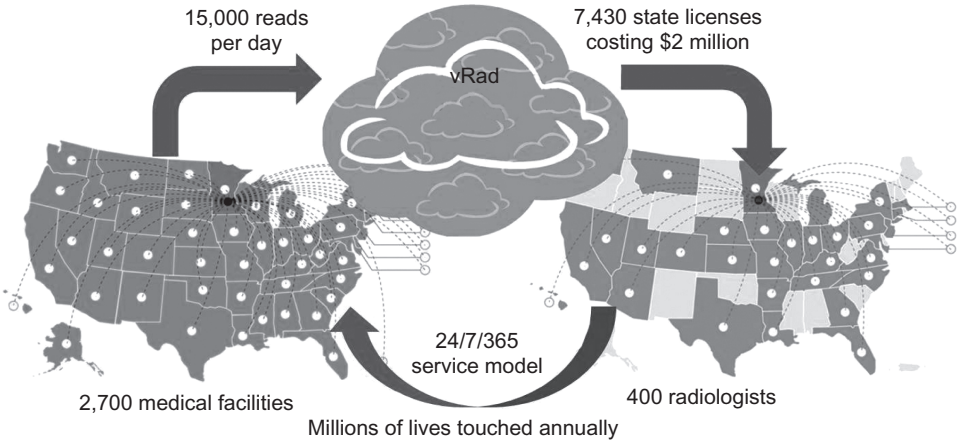

Figure 1: vRad's clients and radiologists are spread throughout the United States.

An average vRad radiologist reads more than 25,000 studies (jobs) annually, a number significantly higher than the industry average, while maintaining an exceptional 99.7 percent accuracy rate (Nicodemus 2010).

The telemedicine industry, with an estimated $3.6 billion value (Monegain 2009), offers flexibility and significant savings in associated medical costs; yet it brings with it operational challenges that were unknown to classical medical practices. In an environment in which a student in rural Georgia can be diagnosed by a pediatrician in a faraway state, the operational complexity reaches a whole new dimension, leading to four major challenges. The first challenge is to effectively manage capacity at the national level. Telemedicine requires a highly complex licensing process: for a doctor in one state to prescribe in another, the doctor must also be licensed in that other state. State licensing usually takes three to eight months and is quite costly. Moreover, hospital networks require doctors to have hospital privileges (credentials) in addition to state licenses to be able to diagnose medical images. Second, at the operational level, daily personnel management is no longer a minor optimization problem faced at the local hospital level; instead, it involves matching hundreds of specialized doctors with thousands of patients. Moreover, as hospitals downsize their workforce capacity and rely on telemedicine at night and on weekends, it causes huge variations in demand patterns for telemedicine companies. Third, the success of the business is strongly tied to its promise of an incomparable gain in efficiency (i.e., the productivity rates of doctors increase substantially because of a higher flow of patients); thus, the effective management of a system of this scale becomes critical in today's competitive industry. Finally, the core competencies of telemedicine practices are their more specialized doctors, faster service, and lower costs. In the medical industry, both service levels and time are of the utmost importance. Telemedicine requires a crucial analysis of the system at the level of minutes instead of the hours or days that companies in manufacturing or other service environments traditionally use. Therefore, there is constant pressure to improve business operations; developing accurate demand forecasts at the level of minutes is particularly critical. Comprehensive models for resource management play a vital role in these time-sensitive environments. The main challenge, therefore, is to optimize the staffing structure to deal with huge temporal variations in daily telemedicine demand.

To assist vRad to overcome these challenges, we developed two resource-management tools. Our first tool is a forecasting model that predicts both short- and long-term demand. Our second tool is an optimization model that allows vRad to match its demand with contracted radiologists to design a robust plan to handle various scenarios. The real-world problem is far too complex to address comprehensively in a single paper. Therefore, we present only our condensed results to highlight our conceptual framework and its basic application and implementation issues. We have also masked the data to protect company confidentiality

and to follow the privacy guidelines of the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

We organized the remainder of the paper as follows. In the *Diagnosis of the Problem and Operational Challenges* section, we describe the managerial problem. In particular, we summarize the operational challenges and the legacy process. Then, in the *Solution Approach* section, we explain our solution approach, which has two phases: (1) data collection and (2) strategic and tactical modeling. In the data-collection phase, we analyze the supply and demand data to understand the characteristics of the business environment and associated trends and to outline a suitable way to model vRad's problem. In the second phase, we forecast the demand and propose a capacity-planning tool. The forecasting methodology and optimization model for capacity planning are detailed in the *Forecasting Model* and *Loading Model* sections. We present a summary of the results and discuss policy implications in the *Impact* section.

Diagnosis of the Problem and Operational Challenges

We can categorize the main drivers of vRad's business as follows:

- Cost of radiologists on shift.
- Costs related to overutilizing or underutilizing radiologists.
- Cost of licenses and credentials needed to run the business.
- Opportunity cost of poor service (i.e., cost of losing client business).

Clearly, the items listed above have some conceptual overlap. For example, reducing the number of radiologists is likely to go hand in hand with an increase in overtime and licensing costs and a reduction in gross revenue. The key to vRad's success is in keeping an ideal balance between these main drivers. This balance must be achieved in the context of addressing the operational challenges of matching jobs with radiologists while taking into account factors such as licensing requirements and subspecialty expertise.

In the following section, we provide an exploratory analysis of vRad's supply and demand, followed by an explanation of performance metrics and our overall assessment of the problem.

Analytics of Supply

vRad contracts with its radiologists to work on predetermined days and at predetermined hours. Although the company needs to have the appropriate radiologists available on shift to process the images promptly, it creates the radiologist schedules in advance. Therefore, in any given week, the list of radiologists that are contracted to work is fixed. In emergencies (i.e., when the composition and (or) number of radiologists cannot cover demand), vRad requests some of its off-shift radiologists to work additional shifts. Given the frequency of these last-minute schedule adjustments, the company has a dedicated scheduling department to handle these issues.

A hard constraint is that each radiologist must be licensed in each state for which he (she) can read jobs. In addition, radiologists must have separate credentials at the hospitals sending jobs to vRad. This results in an extremely complex and expensive licensing and credentialing process, involving a total of 7,500 state licenses (for all radiologists) and tens of thousands of facility credentials. Figure 2 depicts the distribution of the number of radiologists licensed per state and state licenses per radiologist. Moreover, for a given typical day in 2012, Figure 3 displays the actual state-level demand against the scaled available capacity per state, where the capacity of each radiologist is allocated proportionally to the states in which he (she) is licensed based on the total demand from those states. Such an apportioning does not truly reflect the actual assignment of capacity used to cover state-level demand. Nonetheless, it gives a hint of the mismatch between demand and capacity in certain states (e.g., states 12 and 39).

Although some jobs may be read by nearly all radiologists, other jobs can be read only by radiologists who are licensed for relatively uncommon subspecialties. This further complicates the process of matching capacity with demand.

Analytics of Demand

The time required for a radiologist to process a job depends on the job type (e.g., MRI scan versus x-ray). vRad measures jobs in terms of work units per job; the (fixed) number of work units per job is used as a proxy for the amount of radiologist time needed to process each job. Thus, our demand forecast takes the form

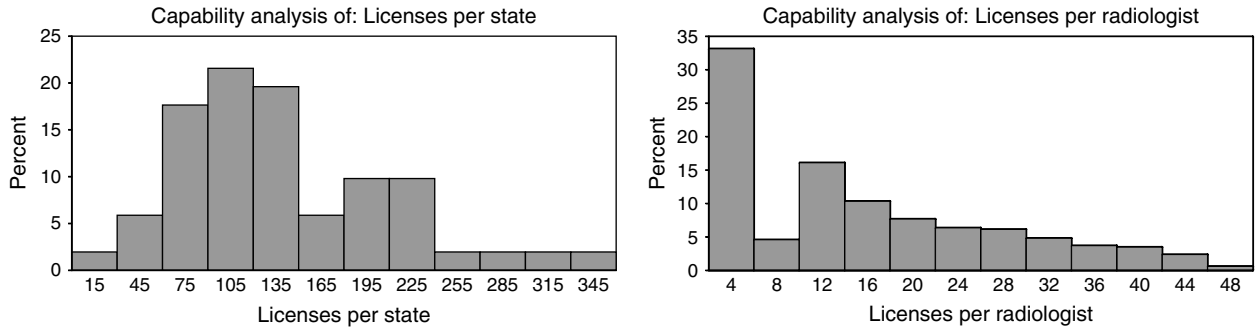

Figure 2: The graphs depict the distributions for (1) the number of radiologists licensed per each state and (2) the number of state licenses per radiologist. The vertical axes represent the percentage of states with the given number of licensed radiologists and the percentage of radiologists with the given number of licenses, respectively. High variances in the distributions are evident in both graphs.

of a prediction of the number of work units in each future period. Designing a forecast involves significant challenges because the demand in work units is noisy and nonstationary, and it has a seasonal, weekly, and daily cyclic structure.

Weekly patterns show that Saturdays and Sundays have about an 18 percent higher average demand than weekdays. At the national level, based on the U.S. Central Time Zone, Figure 4 illustrates the difference between weekdays and weekends by showing the average demand separately for weekdays and weekend days. The left panel considers only emergency jobs, which include jobs that are needed for especially time-sensitive medical conditions and comprise approximately 94 percent of the work units. Nonemergency jobs account for the remainder of the work units. From Figure 4, we can clearly see that the daily pattern depends both on the time of the week (weekend versus

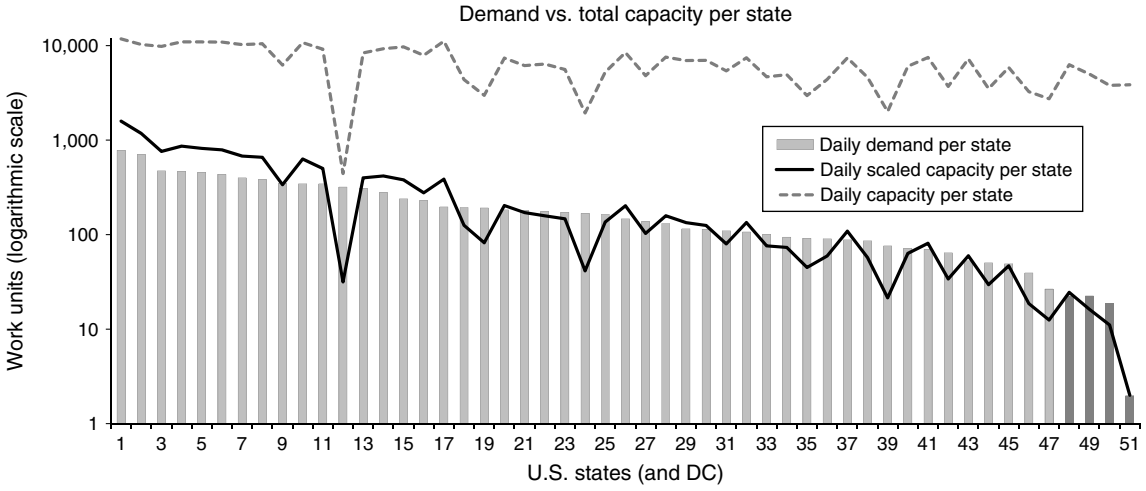

Figure 3: We calculate total scaled service capacity as follows. For each state, we normalize the capacity of each licensed radiologist based on the total demand of the states that the radiologist serves. We then sum the normalized capacities to determine the scaled capacities for each state. The scaled capacity metric measures how fairly licenses are distributed across states; however, it does not indicate whether a state has too many or too few licenses. The substantial gap between demand and total service capacity is calculated as the maximum possible service to a state. For example, in state 48, the realized demand is less than one percent of the allotted capacity. The gap indicates a conservative licensing approach.

Figure 4: The rate at which hospitals request work units depends on many factors. The left panel shows that demand for emergency work is especially high between 8 PM and 5 AM on both weekends and weekdays. The right panel shows that nonemergency work happens primarily during standard weekday working hours, with suppressed demand on weekends.

weekday) and on the type of read request (emergency versus nonemergency).

Figure 4 shows a clear interaction between the weekly variation in demand and the daily variation in demand; the shape of the demand over the hours of the day differs for each day of the week. Furthermore, the distribution of demand across facilities (see Figure 5) is highly skewed; a large number of very small facilities send only a few jobs per day. Because demand may be intermittent, generating an accurate forecast at the individual facility level for shorter time intervals (e.g., 30 minutes) is exceptionally challenging.

Key Performance Metrics

In addition to these operational constraints, two key performance measures contribute to vRad's operational complexity:

1. Turnaround time (TAT): TAT is the time from a job's arrival at vRad's system to the time the completed report on it is returned to the customer. This performance metric is critical for both patient care

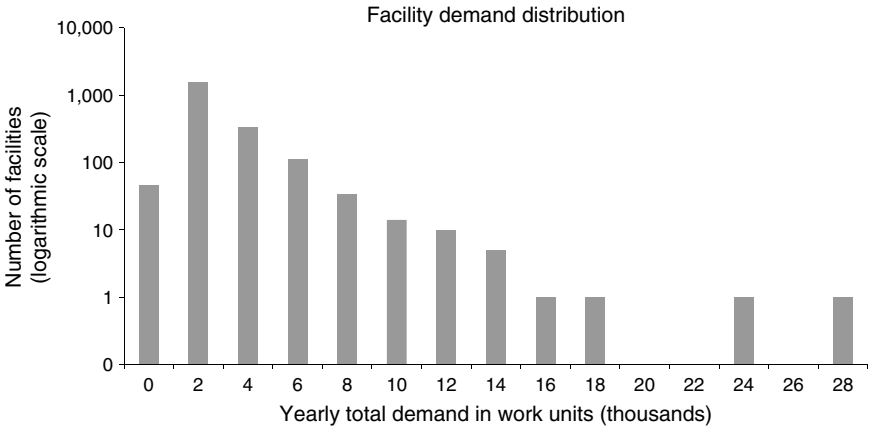

Figure 5: The yearly total demand distribution of facilities is highly skewed. The majority of the facilities (1,600 of 2,700 vRad clients) demand an average of five to six jobs per day.

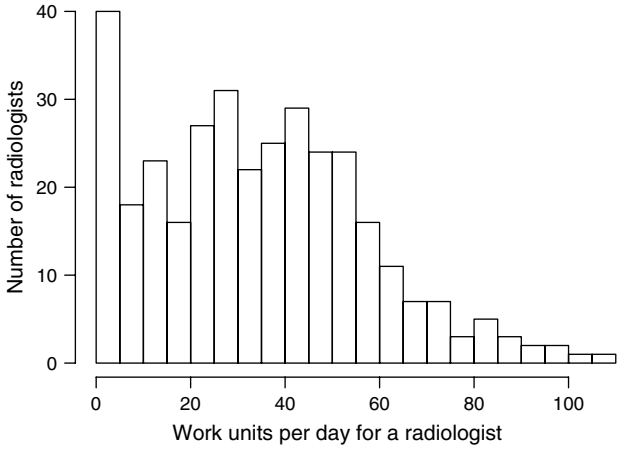

Figure 6: Some radiologists process many more jobs than others. The histogram for 2012 shows that 40 radiologists averaged fewer than five work units per day, whereas two radiologists averaged more than 100 work units per day.

and customer satisfaction. vRad has four job-priority classes, depending on the urgency of a job. For high-priority jobs, the completion time is critical. Emergency jobs, which must be processed as quickly as possible, constitute the highest-priority class; vRad's policy is to serve emergency jobs within half an hour.

2. Radiologist utilization: Radiologists are the company's most valuable resource. vRad's radiologists consistently rank among the top five highest-paid radiologists, with national averages of approximately $400,000 per year. vRad uses work units as a proxy for the amount of radiologist time needed to process a job (a read request), and radiologist pay is based on the total number of work units completed. Thus, it must effectively utilize its on-shift radiologists. More important than high utilization, vRad seeks a fair distribution of utilization among its radiologists; yet, Figure 6, which shows the distribution of work units processed by the radiologists, clearly indicates a large spread.

Problem Diagnosis

Initial analytics of supply and demand confirm that the successful daily operation of vRad requires an efficient way to match qualified radiologists with jobs. State licenses constitute a large cost item in vRad's budget; therefore, the company consistently tries to carefully manage them in relation to the performance of schedules. Thus, accurately predicting demand to assess the number and types of each license to be maintained is critical. Scaling the capacity of each radiologist with respect to the relative demand (see Figure 3) makes sense for aggregate capacity-planning purposes (i.e., whether the licenses are distributed in proportion to demand); however, this measure is not very useful for understanding the overlicensing and (or) underlicensing phenomena. To help us gain a better understanding of this issue, Figure 3 also plots, for each state and for the same day in 2012, its total demand against its total available capacity, which is calculated as the maximum amount of service possible to a state s if all the radiologists licensed in state s serve only state s. Although this capacity is not reflective of the operational capacity to serve state s, it provides some measure of the total number of radiologists licensed in that state. As the figure depicts, vRad takes a conservative approach to the licensing problem, leaving ample excess capacity for each state. Also, given the time required to obtain a license, it enables the company to quickly start serving a large customer base in any state. Yet, with better demand forecasting and ways to assess risk, we argue that vRad can responsibly reduce its overall spending on radiologist costs and state licenses without sacrificing its quality of service (i.e., without increasing average TATs).

Legacy Process

Previously, vRad employed a two-stage forecasting system conducted by two separate departments. Its finance

team used a labor-intensive process of generating quarterly demand forecasts using Microsoft Excel, with forecasts generated primarily at an aggregate level. The team infrequently updated the long-term forecasts, and the scheduling department frequently made short-term forecasts using daily and hourly average volumes from the prior two weeks to capture more recent trends. This decentralized forecasting approach left room for improvement on several fronts. First, the finance team made its long-term forecasts only quarterly; thus, it tended to miss short-term trends. Second, short-term scheduling forecasts looked only at short-term trends and did not build on long-term fundamentals. Third, forecasts did not capture variability to allow buffer capacity-type decisions. Finally, secondary to legacy information technology (IT) tools, forecasts were generated for the aggregate demand at the national level, missing much-needed facility-level and state-level granularity for capacity planning.

On the supply side, the scheduling team constructed biweekly shift schedules for the radiologists to reflect the short-term demand forecasts. This process was mostly an art, relying largely on the intuition and experience of vRad personnel that was hardcoded as a rule-based scheduling system. Historically, these manual techniques were developed in the early days of the company, when the operations volume was much smaller. With vRad's acquisition of NightHawk, in 2011, the operations volume increased significantly, raising operational complexity to a substantially higher level. Although the vRad staff is experienced and qualified, the labor-intensive planning approach inevitably resulted in an inability to handle all this complexity and caused inefficiencies. In particular, ongoing capacity shortfalls led to the need for the routine use of emergency calls to off-shift radiologists to ensure the timely processing of jobs.

Solution Approach

To improve vRad's forecasting and capacity planning, we proposed an automated forecasting and planning tool; Figure 7 illustrates our approach. As the figure depicts, the forecasting model creates short-term and long-term forecasts. Subsequently, licensing decisions are based on the analysis provided by the loading model. Given state licenses and short-term demand forecasts, vRad's daily capacity is planned using the loading model. The capacity plan determines the set of radiologists that will be on shift. The output of our forecasting and planning tool is then entered into vRad's study assigner. Finally, real demand is realized and jobs are allocated to available radiologists via vRad's study assigner. Note that vRad's study assigner is beyond the scope of this project.

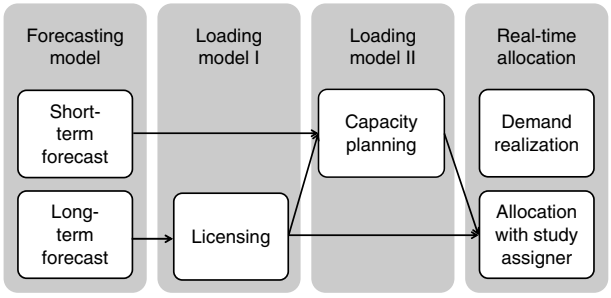

Figure 7: Our solution methods are used in three planning stages and support vRad's study assigner.

Forecasting Model

Literature and modeling choice: The drastic change in demand structure after the acquisition of NightHawk in mid-2011 motivated us to fit our forecasting model using only data after July 12, 2011. The data extend into October 2012, so sufficient data exist for a one-year training set and a 60-day forecasting test set. We observed a mild annual seasonal effect, with elevated demand in late summer and reduced demand in early winter. Our regression model incorporates a term for this seasonality. In addition, a coefficient for time captures the long-term growth trend.

Regular daily oscillation in demand provides the vast majority of the explanatory power of our hourly forecasting model. The daily pattern is clear in Figure 8 for both the observed half-hourly demand (dots) and for the forecast (continuous curve), which track the daily pattern closely. A relatively subtle weekly pattern is visible upon close examination of the forecast curve in Figure 8; Saturdays and Sundays have an average level of demand that is about 18 percent higher than weekdays. Moreover, Figure 4 shows a clear interaction between the type of read request (emergency versus nonemergency), the weekly variation in demand, and the daily variation in demand, because the shape of

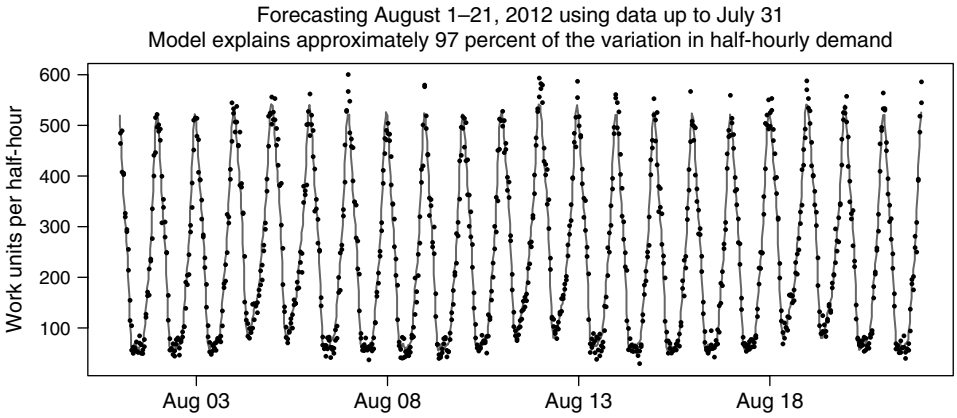

Figure 8: National aggregate demand follows a clear daily pattern. Each dot represents the number of work units demanded in a specific half-hour period. Thus, each day has 48 dots. The curve represents the forecast that was generated using only the data prior to August 1, 2012.

the demand over the hours of the day changes with the time of the week. Thus, we include a term for this interaction in our regression model.

In summary, our half-hourly forecasting model accounts for a linear time trend, the annual seasonality, the weekly pattern, the daily pattern, and the interaction between the weekly and daily patterns. Each of these effects is represented by a term in an additive modelthat is a modern regression framework that allows us to replace ordinary regression coefficients with general smooth functions (Hastie and Tibshirani 1990). We used the `mgcv` package in `R` (R Core Team 2012) to nonparametrically estimate the smooth functions as penalized regression splines. The purpose of introducing this extra machinery is to capture nonlinear effects. For example, the daily pattern is not simply a straight line; nor is it a step function, as would be implied by including a separate coefficient for each hour of the day. Instead, the daily pattern that dominates Figure 8 is more like a smooth sine-like wave, and any such smooth pattern can be represented accurately in an additive model (Wood 2006).

The residuals from our forecasting model tend to be autocorrelated, accumulating at the positive or negative side for brief periods. We build a real-time forecast update based on the local deviation pattern. At any point in time, we use known residuals to predict the residual in the next period, and we subtract this prediction from the main regression model forecast to obtain a forecast that we adjust for the local deviation pattern. We produce this real-time forecast adjustment by fitting an autoregressive moving average (ARMA) model that includes a linear term, two autoregressive lags, and two moving average lags using only the previous three weeks of observed residuals (Ripley 2002). We chose the number of lag terms by experimenting with several similar models on a small test set in search of the model with the minimum mean square error.

The Model. The foundation of our demand forecast is an additive regression model that regresses demand on the trend and seasonality effects of demand that are measured in work units. Appendix A provides the underlying mathematical formulation and details of the ARMA model. Our forecasting model can explain as much as 97 percent of the variation in half-hourly demand at the aggregate national level. To be precise, the variance of the residuals is approximately three percent as large as the variance of the demand. This error rate translates to a mean absolute percentage error of approximately nine percent for out-of-sample testing. Local autocorrelation trends are often evident in the residuals of the regression model; we use these trends to build a real-time adjustment for the next-period forecast, leading to a moderate accuracy improvement for short-term forecasting. The improvement in forecast accuracy as a result of the ARMA adjustment is modest at the national hourly level, reducing the mean absolute percentage error from 9.0 percent to approximately 8.7 percent.

We have described our forecasting approach for the half-hourly demand aggregated over all of the vRad

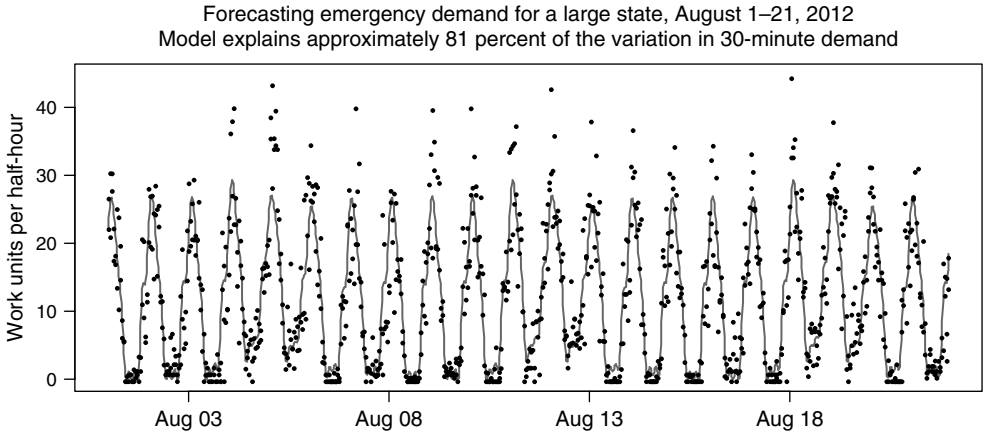

Figure 9: The basic interpretation of the dots and curve is the same as in Figure 8. The difference here is that the data concern demand from a large state instead of the national aggregate demand.

facilities in the United States, and we also presented vRad with an option to further aggregate the data for an hourly, eight-hourly, or daily forecast, which may be more relevant for some capacity-planning purposes. Alternatively, our framework is flexible for disaggregating the demand series to produce a localized forecast at the facility level, which is much more important from an operational point of view. However, forecasting demand separately for each state, facility, or specialty raises some issues. Our forecasting model may be applied based on data at any level of disaggregation, but the quality of the forecast changes with disaggregation.

For example, Figure 9 illustrates the forecast that results from fitting the regression model for emergency demand data for a large state. The ratio of noise to signal is clearly much greater for this cut of the data than for the national aggregate (see Figure 8). We can perhaps best explain this difference in accuracy between the state forecast and the national forecast by realizing that the variance of a mean grows as the sample size shrinks.

Loading Model

The loading model is a linear programming (LP)-based diagnostic and planning tool that lets vRad analyze, based on a given planning horizon, whether it has the right kind and number of radiologists scheduled to meet the expected demand, and whether the current set of licenses suffices to efficiently serve all demanding facilities. The loading model takes the demand forecast as a primary input.

Literature and Modeling Choice. The loading model has three streams of related literature. The first stream is scheduling and capacity planning. Graves (1981) provides a comprehensive review of scheduling literature, and Pinedo (1995) is a useful resource on scheduling models and well-known algorithms. We use the well-established total-weighted completion time (total-weighted TAT in the context of this paper) as our objective in the loading model. In vRad's environment, each radiologist has a different hourly reading capacity, and jobs have different release dates and due dates. Unfortunately, vRad's scheduling problem is not tractable; nor, to the best of our knowledge, does a state-of-the-art scheduling algorithm exist that would fit such an environment. The loading model has similarities to flexible workforce scheduling (e.g., Hung 1994) in terms of flexible work hours of radiologists and healthcare or emergency room (Beaulieu et al. 2000) and operating room (Batun et al. 2011) scheduling problems. Ernst et al. (2004) provide a comprehensive survey of this literature and argue that, because of tractability issues, healthcare and flexible workforce scheduling studies focus on specific aspects of healthcare planning problems. vRad, however, must consider both the supply and demand sides of the problem.

A second stream of literature pertains to call center capacity management; for examples, see Whitt (1999) and Armony and Maglaras (2004). Gans et al. (2003)

provide a comprehensive survey. In particular, vRad's problem can be modeled as a skill-based call center problem (e.g., Wallace and Whitt 2005, Visschers et al. 2012); however, several issues render such models intractable for vRad. First, vRad's problem translates into a skill-based call center problem that contains hundreds of server types as a result of (1) the distinct skills and licenses each radiologist possesses and (2) the thousands of customer types. Second, vRad's demand is nonstationary and exhibits seasonality at the monthly, daily, and hourly levels. This does not make the problem unsolvable, because several papers have been written on nonstationary call center problems; however, it adds more complexity. Third, vRad's radiologists work flexible hours and are on duty only during a certain portion of a shift. In a call center context, this requires some servers to become operational and nonoperational within the planning horizon. We are not aware of any work in the call center literature that tackles such a large-scale and complex problem.

A third stream of related literature is on capacity rationing (e.g., Swenson 1992, Lee and Zenios 2009, Chan et al. 2012) and capacity pooling (e.g., Ata and Van Mieghem 2009, Mahar et al. 2011, Vanberkel et al. 2012, Best et al. 2012) problems in healthcare. Although the models from the capacity-rationing and pooling literature provide useful insights, they offer no guidance for tackling large-scale real-life capacity-planning problems.

We made the following modeling choices in conjunction with vRad to keep the model tractable and focused. Our model builds on the demand forecast; because demand is variable, we discretized the planning horizon. In a short-term analysis, we broke down a day into 30-minute time periods (48 periods per day). We chose the 30-minute interval to reflect vRad's policy to serve emergency jobs within 30 minutes. For mid- or long-term analysis, we also provided the flexibility to adjust the period length to a day or a week. To simplify the heterogeneity of jobs, we built the model in work units. For example, instead of keeping track of the time and job types (e.g., MRI, CT), the model tracks the work units, which are proportional to the time it takes to read various job types. Hence, we eliminated this added complexity resulting from job types by converting everything into work units; all periodic capacities and bounds are in work units. The inherent inaccuracies involved in demand forecasting would naturally lead to a stochastic model; however, our short time interval of 30 minutes and the large dimension of the resulting model limited us to a deterministic model. As a mitigation strategy, we equipped users with automated tools to make proper sensitivity analyses and to check the performance of the system under several scenarios.

The Model. The loading model uses a linear program that matches demand with the supply of available radiologists (see Appendix B for the complete mathematical formulation). The objective is to minimize the weighted-average TAT with weights assigned to different priority classes to move higher-priority jobs forward in the queue. In conjunction with vRad, we assigned the weights so that the higher-priority jobs are served first, and the remaining capacity is used for the lower-priority jobs. For example, the weight of the first-priority jobs is assigned so that the model always chooses first-priority jobs over other jobs for allocation. We chose to assign priority weights instead of solving nested LPs for each priority class for two reasons. First, in the first of several nested linear programs, the capacity of radiologists usually exceeds the demand (although the capacity is tight when all priority classes are considered). This results in alternative optimal solutions for the highest-level nested linear program; yet some of these alternative solutions are suboptimal and (or) lead to infeasible solutions for the general problem at the later stages. Second, using weights to adjust the importance of the jobs has led to a more flexible system than nested linear programs, because weights can easily be updated in the case of any change in the priority structure or company policy. vRad's other critical factor (i.e., having close-to-uniform levels in terms of the utilization of each radiologist on shift) is partially incorporated into the constraint set via the addition of a constraint that sets the lower bound for the workload of each radiologist. We handle increasing the radiologist utilization mainly using sensitivity analysis, as we discuss in more detail in the *Impact* section.

Each radiologist has a different periodic reading capacity, which we embed in the model via the first constraint. The second constraint sets a lower and an upper bound on the total number of work units that a radiologist can have within a planning horizon.

The lower bound aims to balance the workloads of radiologists; the upper bound ensures that the total workload of each radiologist stays within a reasonable level. (Note that the lower bound can cause the model to be infeasible. In the implemented version, the user is alerted to decrease the lower bound in case of infeasibility.) The third constraint balances the number of jobs demanded, processed, and backlogged between periods. For a radiologist to be able to read a job in a given period, he (she) should be available and should have the necessary licenses, credentials, and subspecialty expertise. We embed these restrictions in the last constraint.

Implementation Issues

vRad implemented our planning software, which personnel in its operations and planning department use primarily, in December 2012. Ongoing communication between the teams at vRad and Carnegie Mellon University (CMU) was critical for the success of this project. In weekly conference calls and meetings, both teams shared updates to keep the project focused and exchanged ideas to form tractable, easy-to-apply models that are well suited to the vRad environment. From the beginning of the project, an understanding of what vRad needed and what the CMU team could deliver in terms of modeling restrictions was essential. Thus, we initially spent a substantial amount of time to understand vRad's system and to form a mutual understanding on what to expect from the project. Next, we summarize the major implementation issues both teams faced.

Data Integrality and Validity. Prior to the implementation of this project, a number of data challenges had to be resolved. In particular, the company's operations and planning department had to establish and validate a centralized SQL database containing all the pertinent records. This process had several steps to ensure data integrity by cross-validating the data from different sources at the company while ensuring that the required granularity of the data was maintained.

Handling Scale and Complexity. After careful analysis of the demand structure and in conjunction with vRad, we decided to cluster all facilities, except the 200 largest facilities, into state clusters for short-term forecasting and capacity-planning purposes. The reason for this decision was the difficulty of obtaining accurate demand forecasts at the half-hourly level for small facilities that send five to six jobs a day; as Figure 5 depicts, approximately 60 percent of the facilities fall within this category. After grouping the number of facilities (or facility state aggregates) by state, we entered 254 facilities into the loading model; these included the 200 largest facilities based on demand volume, 50 U.S. states, Puerto Rico, the District of Columbia, pro bono jobs, and government categories.

Typically, forecasts are generated for specific cuts of the data (e.g., all emergency reads from California) by querying the database and running the forecasting model from within `R` (R Core Team 2012). Although queries typically took seconds to run and the prediction routine was also able to fit the main regression model and produce a forecast in under five seconds, the ARMA adjustment for correlated residuals required approximately an hour to run for a single forecast at a half-hourly frequency. Therefore, we provided users with the option to select the forecasts, with or without the ARMA adjustment, depending on their need for forecast accuracy and speed.

The IT implementation of the loading model was relatively challenging; it required a user-friendly input or output module and also had to be powerful enough to tackle vRad's large-scale planning problems. We coded the model using AIMMS software. The inputs and outputs of the model are communicated to the user via Microsoft Excel, and IBM ILOG CPLEX 12.4 is used as the underlying optimization engine. The loading model had to solve problem instances containing hundreds of thousands of constraints and variables for a small-scale problem even after preprocessing. For example, a problem with 300 facilities, 250 radiologists, one subspecialty class, and 48 periods translated to an LP model (after preprocessing) with 1,327,933 variables and 1,364,550 constraints, and it was solved in 90 seconds. For larger-scale instances, the size of the problem extended to a level where it was impossible to construct and load the model into the computer memory (see Appendix B for details). Thus, we proposed a decomposition algorithm that exploits the specific sparse structure of vRad's problem instances. This algorithm (see Appendix B) iteratively solves the loading model for different smaller instances and is especially useful in medium- and long-term analyses where more computing power is necessary.

Organizational Inertia. vRad's existing system had been in use for almost a decade, and vRad executives and personnel had several concerns about completely replacing it. In addition, frequently adjusting the contracts and licenses of radiologists on a large scale was undesirable and impractical. Both of these factors posed several challenges. First, vRad feared harming customer satisfaction, as it tried to improve its operational costs. Second, vRad was concerned about upsetting its radiologists who could potentially harm the company in the long run. Third, although the simulation results we used to gauge the project's impact were very promising, the planning personnel were concerned about their role after the project's completion and were reluctant to trust a black-box type of planning tool. To overcome the organizational inertia, vRad decided to implement an incremental analysis approach. In this approach, in the short-term planning, planners would take the existing radiologist schedules and currently available licenses and use the new tools to assess the effect and value of various incremental changes (e.g., extending a radiologist's shift or adding extra radiologists to a shift). This type of incremental analysis not only gave the users an opportunity to adjust to the new tools but also helped them leverage their expertise. Thus, our tools are now being used continually by the operations and planning department to make incremental updates to its plans for shorter-term decisions. Moreover, at the strategic level, these tools are used to evaluate different scenarios prior to making major changes, such as decisions about hiring, contract renewal, and license expansion.

We elaborate on the impact of our planning tools in the next section.

Impact

First, we describe the experiments we made to estimate the impact of our planning tools, and we illustrate some of the analysis that vRad is now able to make with our planning tools. Second, we detail the project's quantifiable and qualitative impact after implementation.

Gauging the Impact of Implementation

The methodology to estimate the impact of the loading model involved many pilot implementation experiments using real data from vRad. Our experimental studies with the forecasting and loading models were quite promising. The forecasting model provided national-level forecasts that can explain 97 percent of the demand variability. Moreover, the forecasting model offered state-level and facility-level forecasts, which had not previously been incorporated into vRad's legacy process. Figure 10 depicts, at the state level for half-hourly forecasts, how well our forecasts fit the data. For states with large demand volumes, our forecast fit the data especially well. With the implementation of the project, vRad can now plot similar figures and gain insights on how good forecasts will be on a state-by-state level. In the long run, this information will help the company improve its forecasting quality by clustering small states. Another important benefit of the forecasting tool is the ability it provides to run queries. vRad can pull the demand structure of certain states or facilities, enabling it to analyze how its business is progressing. The tool also provides the users with confidence intervals that were not available in the legacy process.

In our experiments, we compared the schedules our planning tools created with the actual allocations

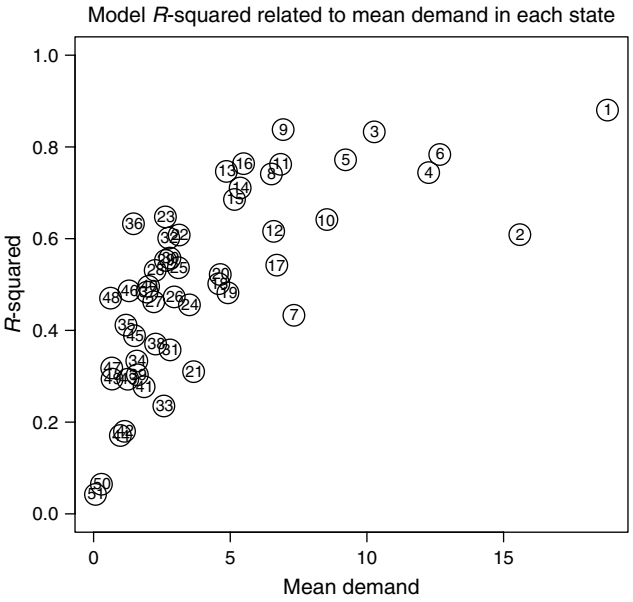

Figure 10: The ratio of signal to noise in the demand time series depends strongly on the level of aggregation. This plot shows how the demand from states with high average demand tends to be more predictable than demand from states with low average demand. For example, states 50 and 51 (lower left) have both low demand and low *R*-squared (a measure of predictability); state 1 (upper right) has both high demand and high *R*-squared.

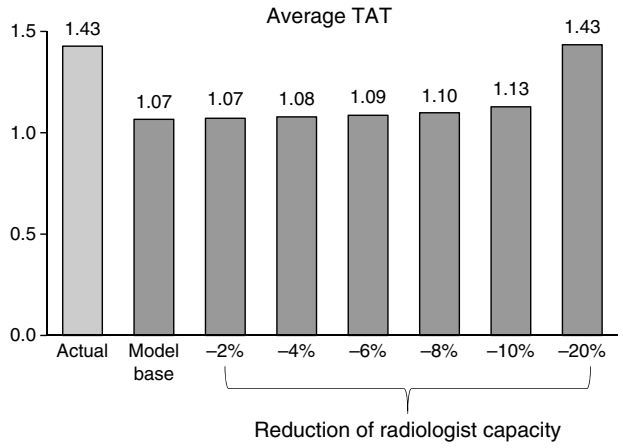

Figure 11: The figure compares the average TATs of past allocations with the schedules provided by our planning tools using different capacity levels. (The numbers are representative of the real ratios, but we scrambled them for confidentiality purposes.)

made by the vRad study assigner before the project implementation. As Figure 11 depicts, the loading model schedules resulted in a 25 percent improvement in average TATs and a promising reduction in TAT variance. Although the comparison of our schedule with the actual allocations of vRad's study assigner is not an apples-to-apples comparison, and real-time allocation has many additional challenges, the 25 percent margin was indicative of the potential improvement our planning tools could provide to vRad. To measure the robustness of our planning tools, we used a feature embedded in the implemented loading model user interface: we tested the TATs that the planning tools created under different levels of radiologist capacity. As Figure 11 demonstrates, the average TATs of our solution remained better than the TATs under the legacy process up to a 20 percent capacity buffer.

In addition to the sensitivity analysis detailed above, in the loading model, we provided a large variety of scenario and sensitivity analyses tools to the user. For example, for a given day, Figure 12 depicts the total job demand and the total number of jobs read (supplied). When radiologist capacity is decreased by 20 percent, keeping up with peak demand is no longer possible, and some jobs are postponed to future off-peak periods. In conjunction with temporary understaffing of radiologists, this hurts the TATs. This type of analysis helps the company to balance the system load and TATs by making small changes to the radiologist schedules in advance.

Similar to changes in radiologist capacity, the users can now analyze, in both the short and long term, the system's sensitivity to changes in demand, the set of radiologists, and the set of licenses. For example, by inflating the demand by a certain percentage, they observe the resulting estimated effects on TAT distribution at each state and identify the states that require special attention; for those states, they could call in backup radiologists in the short term and focus on adding new licenses in the long term. The operational personnel also solve the loading model by incrementally updating the set of licenses or radiologists to improve TATs. As a result, they estimate the

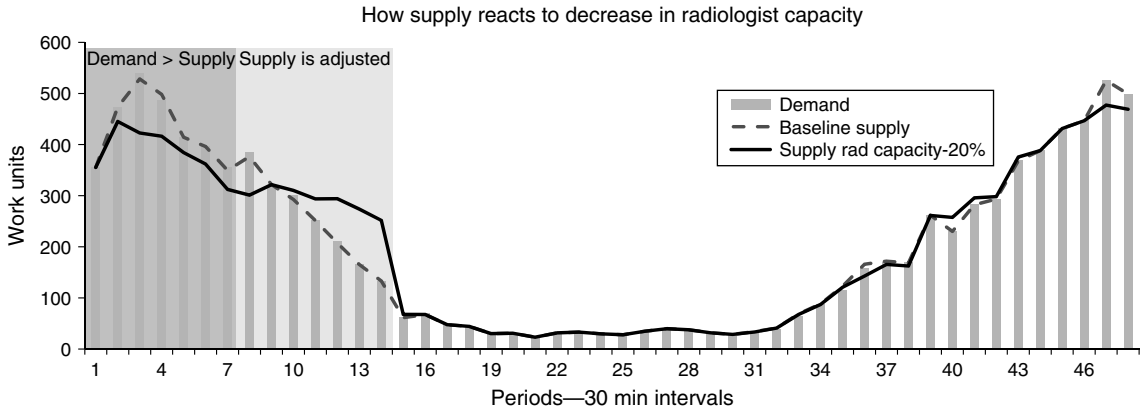

Figure 12: When capacity is reduced, some jobs are postponed and read during off-peak hours.

benefit (cost) of adding (eliminating) a set of licenses or radiologists to (from) the system in terms of TATs.

Realized Impact

Since the beginning of the implementation, we have continually enhanced our planning tools to meet vRad's needs. Following the initial implementation of our planning tools, the company typically used the forecasting tool biweekly and the capacity planning tool monthly to review existing licenses and make new licensing decisions. The improved accuracy of forecasting and better capacity planning has had a direct impact on vRad's radiologist selection, licensing, and credentialing decisions. In particular, vRad has observed a reduction in TATS ranging from 12 to 15 percent, despite its growing business and tighter capacity planning. Moreover, its operational costs have decreased in the aggregate by four to five percent.

In addition to the quantifiable gains listed above, the project has provided many qualitative benefits to the company. Rick Jennings, vRad's chief technology officer, describes some of these benefits:

> The automated forecasting system has simplified the tasks of the operations planning personnel, and enabled the early detection of outlier behavior at the state and large-facility level, an ability that was not easily accessible to us before. The capacity-planning optimization tool enabled us to easily study multiple scenarios and determine a robust capacity plan in terms of hiring and licensing decisions for radiologists. As a result, the overall need for last-minute adjustments to call in backup capacity from radiologists for handling sudden shifts in volume and mixes of demand has been reduced, simplifying the task of the operations department each day. These tools in totality have transformed the related analysis and decision-making process at vRad to be more transparent and rational, removing much of the individual biases and preferences.

As Figure 3 illustrates, vRad was conservative in the licensing process, leaving ample excess capacity for each state. For example, for some states the demand is less than one percent of the capacity. The conservatism in licensing is understandable, given vRad's desire to keep TATs as low as possible and to quickly serve new customers without waiting for new state-granted licenses. However, with the improved demand forecast and better planning, vRad can now use the loading model to make long-term planning decisions and can see the impact of reducing licenses. This allows it to quantify the risk of reducing licenses, saving the company a significant amount in licensing costs.

Lessons Learned

For the academia members on the research team, the project underscored some vital aspects of operations research (OR) implementations:

Early stakeholder acceptance: From the start of the project, vRad's upper-level management team was convinced of the need for advanced analytical tools. However, our team had to ensure that the major stakeholders were also aligned. The owners were primarily interested in efforts that boosted the value of the company, whereas the chief medical officer, who is also a coauthor, was keen on capturing critical variables that likely would both improve patient care and address the needs of the service providers (i.e., radiologists). The chief technology officer wanted us to ensure that the proposed software would function within the company's existing IT infrastructure, whereas the operations and planning personnel stressed that the applications had to be user friendly and the analytical tools simple enough to help them make better day-to-day operational decisions. Balancing all these trade-offs and priorities and designing tools that were consistent with past policies, while capturing essential operational constraints and leaving room for possible future modifications, proved to be a delicate balancing act. We learned that modeling skills alone are not sufficient to ensure project success. The human dimension in terms of stakeholder acceptance of OR implementations is equally critical.

OR modeling perspective: First, our practical experience from this project revealed the importance of using different forecasting methodologies at various levels and their significant impact on the efficiency of a telemedicine company. In particular, it highlighted the lack of literature and tools in intermittent demand forecasting for this type of medical operation. Second, in large-scale, time-critical environments, such as the one in which vRad operates, we realized the importance of decomposing the system into meaningful and tractable pieces that could be optimized. This trade-off between (1) optimality and speed and (2) shaping practical, easy-to-use models seems to be present in many business environments. We hope that our findings will guide

managers in other service environments who are facing similar challenges and seeking ways to harness OR techniques and tools.

Appendix A. Details of the Forecasting Model

The data used to fit each forecast model come from the set of all read requests in the time interval starting in the first hour of July 13, 2011 and ending in the last hour of July 31, 2012. We refer to this time interval as the *training period*, because we use these data to train each model to forecast into August 2012 and beyond.

A separate model can be fitted for any specific kind of read request. For example, we can model the read requests from an individual client facility, the read requests from West Virginia, the CAT scans from California, or simply *all* read requests in vRad's network. Thus, the first step in the modeling process is to choose a specific kind of read request. Let $\mathcal{R}$ denote the set of all read requests that meet this criterion.

Let $i = 1, \ldots, n$ index the hours of the training period. For example, $i = 1$ corresponds to the hour from midnight to 1 AM on July 13, 2011, $i = 2$ corresponds to the hour from 1 AM to 2 AM, and so on. Let $|\mathcal{R}|$ denote the number of read requests in $\mathcal{R}$. If $\mathcal{R}$ corresponds to a single small facility, we may have $|\mathcal{R}| < n$, which would imply that the average number of reads per hour was less than 1. If $\mathcal{R}$ corresponds to the set of all reads in the network, then $|\mathcal{R}|/n$ is on the order of 600 (see Figure 8).

Let y_i denote the number of read requests in the ith hour ($i = 1, \ldots, n$). Let h_i denote the hour of the day of the ith hour. Hence, $h_1 = 0$, $h_2 = 1, \ldots, h_{23} = 23$, $h_{24} = 0$, $h_{25} = 1$, and so on. Let w_i denote the day of the week, so that $w_i = 0$ if the ith hour is on a Sunday, $w_i = 1$ if the ith hour is on a Monday, and so on. Similarly, let m_i denote the month of the ith hour. Finally, let t_i denote the number of hours since the first hour. Our most general model takes the form

$$E(y_i) = b_0 + b_t t_i + \beta(w_i) + f_h(h_i) + I(w_i, h_i) + f_m(m_i). \quad \text{(A1)}$$

Here, b_0 is an intercept term; b_t is a linear coefficient of time; $\beta(w_i)$ is shorthand for $\sum_{j=0}^{6} \beta_j I(w_i = j)$, a sum of indicator variables for the day of the week with linear coefficients; $f_h(h_i)$ is a penalized regression spline; $I(w_i, h_i)$ is a thin-plate regression spline approximation that represents the interaction between the time-of-day and time-of-week effects; and $f_m(m_i)$ is a quadratically penalized regression spline.

Each regression spline is a nonlinear function that approximates the partial effect of the argument on the response. One way to understand a spline is by contrasting it with a traditional linear model. In a linear model, the regression spline $f_h(h_i)$ might appear simply as bh_i, where b is a regression coefficient. The term bh_i would represent the partial effect of the time of day on the dependent variable, implying that demand increases (or decreases) linearly over the time of day, with a discontinuity at midnight. Such a linear term is obviously inappropriate in our application; a nonlinear function is needed. Regression splines work much like moving averages and can take on highly arbitrary nonlinear shapes that fit the data much better than a straight line can. For a theoretical introduction involving the precise definition and estimation of the regression splines and underlying mathematical machinery, see Wood (2003); for direct guidance on software and implementation, see Wood (2006).

The residuals from our forecasting model tend to be autocorrelated, accumulating to the positive or negative side for brief periods. Autocorrelation is present (but difficult to detect visually), and some weak patterns are noticeable, with residuals tending to be positive in early September and negative in late September. We build a real-time forecast update based on these kinds of local deviation patterns. Given any point in time, we use known residuals to predict the residual in the next period, and we subtract this prediction from the main regression model forecast to obtain a forecast that is adjusted for the local deviation pattern. We produce this real-time forecast adjustment by fitting an ARMA model that includes a linear term, two autoregressive lags, and two moving average lags using only the previous three weeks of observed residuals (Ripley 2002). We choose the number of lag terms by experimenting with several similar models on a small test set to find the model with the minimum mean square error.

The improvement in forecast accuracy that results from using the ARMA adjustment is modest at the national hourly level, reducing the mean absolute percentage error from 9.0 percent to approximately 8.7 percent. In particular, the width of the 90 percent prediction interval shrinks noticeably, and the root mean square error falls from 25.6 work units to 23.6 work units.

Appendix B. Details of the Loading Model

This section details the sets, parameters, decisions, assumptions, objective, and mathematical formulations of the loading model.

The loading model uses a set of radiologists $\mathcal{R} = \{1, \ldots, R\}$, a set of facilities $\mathcal{F} = \{1, \ldots, F\}$, a set of different priorities $\mathcal{P} = \{1, \ldots, P\}$, and a set of subspecialties $\mathcal{S} = \{1, \ldots, S\}$. The time horizon (planning horizon) of the model consists of T periods. We refer to a job with subspecialty $s \in \mathcal{S}$, priority $p \in \mathcal{P}$, and coming from facility $f \in \mathcal{F}$ as a job of type (f, s, p).

The parameters of the model are summarized in Table B.1 and can be grouped into three sets. The first set of parameters determines whether a radiologist is available and able to read a job. In particular, the parameter $b_{fspt,r}$ takes a value 1 if radiologist $r \in \mathcal{R}$ can read a job coming from facility $f \in \mathcal{F}$, which possesses subspecialty $s \in \mathcal{S}$ and priority $p \in \mathcal{P}$ at period $t \in \{1, \ldots, T\}$. This parameter takes into account three different restrictions. First, the radiologist r should have the necessary licenses to serve facility f. Second, r should have the sufficient expertise to serve subspecialty s and priority p. Finally, r should be available at time t to be able to read a job.

Parameter	Definition
$b_{fspt,r}$	1 if r can read a job of type (f,s,p) at time t, and 0 otherwise.
d_{fspt}	Total work units of jobs of type (f,s,p) arriving at time t.
w_p	The weight in the objective associated with reading a job of priority p.
$C^{\#}_{t,r}$	Maximum amount of work units that can be read by r at time t.
C^{w}_{r}	Maximum amount of work units that r can read within the planning horizon.
L^{w}_{r}	Minimum amount of work units that r can read within the planning horizon.

Table B.1: The table shows the parameters of the loading model ($f \in \mathcal{F}$, $s \in \mathcal{S}, p \in \mathcal{P}, r \in \mathcal{R}, t \in \{1,\ldots,T\}$).

The second set of parameters pertains to the arrival of jobs and the service time and work units of the jobs. Specifically, d_{fspt} defines the total work units of jobs of type (f,s,p) that arrive to the system at time t. The final set of parameters handles the capacity of a radiologist r and the maximum and minimum amount of work units that r can have.

The loading module has two main decisions to make. The first decision is the number of work units of each type (f,s,p) that each radiologist r reads in period t. This decision is denoted by $y_{fspt,r}$, which can take any nonnegative real value. The second decision is the amount of work units backlogged from a period $t-1$ to the next period t; this decision is denoted by the variable I_{fspt}. This decision variable is also nonnegative and real valued. The formulation of the model is as follows:

$$\max \sum_{f,s,p,t,r} (T-t+1) w_p y_{fspt,r} \tag{B1}$$

$$\text{s.t.} \sum_{f,s,p} y_{fspt,r} \le C^{\#}_{t,r} \;\; \forall\, t \in \{1,\ldots,T\},\; r \in \mathcal{R}, \tag{B2}$$

$$L^{w}_{r} \le \sum_{f,s,p} y_{fspt,r} \le C^{w}_{r} \quad \forall\, r \in \mathcal{R}, \tag{B3}$$

$$\sum_{r\in\mathcal{R}} y_{fspt,r} + I_{fsp(t+1)} = d_{fspt} + I_{fspt} \;\; \forall\, f,s,p,t, \tag{B4}$$

$$y_{fspt,r} \le b_{fspt,r} \sum_{\tau=1}^{t} d_{fsp\tau} \quad \forall\, f,s,p,t,r, \tag{B5}$$

$$y_{fspt,r} \ge 0 \quad \forall\, f,s,p,t,r, \tag{B6}$$

$$I_{fspt} \ge 0 \quad \forall\, f,s,p,t. \tag{B7}$$

The constraints of the loading model are as follows. Constraint (B2) sets the maximum amount of work units that a radiologist can read in a period. Constraint (B3) defines lower and upper bounds in terms of the total work units of jobs that a radiologist can read across the whole planning horizon. Constraint (B4) is a backlogging balance constraint that ensures that for each job type and for each period t, the total amount of jobs read plus the amount of jobs backlogged to period $t+1$ (i.e., left to be read in a future period) should be equal to the amount of demand plus the amount of jobs backlogged from period $t-1$. Finally, constraint (B5) implements the condition that a job can be read by a radiologist only if the radiologist is available and able to read the job. The demand term is used to strengthen the constraint using the fact that if there are no jobs to read, then the variable y should take the value 0.

The loading model uses LP instead of integer programming (IP), which would be more realistic. The main reason is the complexity of the problem. In our experiments with the loading model with both past data from vRad and randomly generated instances, the maximum problem instance that could be solved (because of memory restrictions) was for 300 facilities, 250 radiologists, one subspecialty class, and 48 periods. This instance translated to an LP model (after preprocessing) with 1,327,933 variables and 1,364,550 constraints, which was solved in 90 seconds. For similar instances, an IP version of the problem could not even be loaded into the computer memory, let alone be solved. Another drawback of an IP model is that work units can no longer be used as the base unit. Hence, the job type should be added as an index to the variable y in the IP formulation; this in turn enlarges the problem size to an intractable level, even for an LP relaxation of the problem. Therefore, we model the loading model as a linear program.

To be able to solve larger instances than the loading model can solve, we proposed a periodic decomposition algorithm, called the iterative backlog algorithm, which iteratively solves the loading model for subsets of periods. In the algorithm, we first divide the set of periods $\mathcal{T} = \{1,\ldots,T\}$ into K pieces, where the kth piece starts at the beginning of period T^{k-1} and ends at T^k (for notational consistency, let $T^{-1}=1$). Let $\mathcal{T}^k$ denote the kth subset (i.e., $\mathcal{T}^k = \{T^{k-1}+1,\ldots,T^k\}$ for all $k \in \{1,\ldots,K\}$). The algorithm can be summarized as follows: For each $k=1,\ldots,K$, the loading model is solved iteratively, and at each step, $y^*_{fspt,r}$ and I^*_{fspt} are obtained for each $t \in \mathcal{T}$. Then, for the subsequent problem, the demand is updated as

$$d_{fsp(T^k+1)} \leftarrow d_{fsp(T^k+1)} + I^*_{fsp} \quad \text{for all } k \in \{1,\ldots,K\},\, f,s,p,$$

and the loading model is solved again. At each step, all y and I variables that are 0 because of the absence of demand are eliminated. This way, the algorithm exploits the sparseness of the demand array d_{fspt}. Using the iterative backlog algorithm, we were able to solve larger instances. For example, we were able to solve an instance with 300 facilities, 500 radiologists, 10 subspecialty classes, and 48 periods within five minutes.

Acknowledgments

We thank CMU MBA students Cornelia Berger, Brett Cottle, Sudi Gummi, Som Sreedhar, and Erik Thompson, and CMU student Liren Peng for their contributions to the project. We also owe many thanks to Rick Jennings, Bronwyn Lepper, Robert Ribciuc, and other vRad personnel for their continuous support.

References

Armony M, Maglaras C (2004) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52(2):271–292.

Ata B, Van Mieghem JA (2009) The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Sci.* 55(1):115–131.

Batun S, Denton BT, Huschka TR, Schaefer AJ (2011) Operating room pooling and parallel surgery processing under uncertainty. *INFORMS J. Comput.* 23(2):220–237.

Beaulieu H, Jacques AF, Gendron B, Michelon P (2000) A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Management Sci.* 3(3):193–200.

Best TJ, Sandikci B, Eisenstein DD (2012) Managing hospital bed capacity through partitioning care into focused wings. Working paper, University of Chicago, Chicago.

Chan CW, Farias VF, Bambos N, Escobar GJ (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* 60(6):1323–1341.

Ernst AT, Jiang H, Krishnamoorthy M (2004) Staff scheduling and rostering: A review of applications, methods, and models. *Eur. J. Oper. Res.* 153(1):3–27.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Graves SC (1981) A review of production scheduling. *Oper. Res.* 29(4):646–675.

Hastie TJ, Tibshirani RJ (1990) *Generalized Additive Models* (Chapman & Hall/CRC, New York).

Hung R (1994) Multiple-shift workforce scheduling under the 3–4 workweek with different weekday and weekend labor requirements. *Management Sci.* 40(2):280–284.

Lee DKK, Zenios SA (2009) Optimal capacity overbooking for the regular treatment of chronic conditions. *Oper. Res.* 57(4):852–865.

Mahar S, Bretthauer KM, Salzarulo PA (2011) Locating specialized service capacity in a multi-hospital network. *Eur. J. Oper. Res.* 212(3):596–605.

Monegain B (2009) New research forecasts swelling telemedicine market. *Healthcare IT News* (October 8), http://www.healthcareitnews.com/news/new-research-projects-swelling-telemedicine-market.

Nicodemus A (2010) Hospitals farm out radiology diagnostics. *Worcester Telegram Gazette* (July 6), http://www.telegram.com/article/20100706/NEWS/7060382/1101.

Pinedo ML (1995) *Scheduling: Theory, Algorithms, and Systems* (Prentice Hall, Englewood Cliffs, NJ).

R Core Team (2012) R: A language and environment for statistical computing. Accessed November 20, 2012, http://web.mit.edu/r_v3.0.1/fullrefman.pdf.

Ripley BD (2002) Time series in R 1.5.0. *R News* 2(2):2–7.

Swenson MD (1992) Scarcity in the intensive care unit: Principles of justice for rationing ICU beds. *Amer. J. Medicine* 92(5):551–555.

Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, Litvak N (2012) Efficiency evaluation for pooling resources in health care. *OR Spectrum* 34(2):371–390.

Visschers J, Adan I, Weiss G (2012) A product form solution to a system with multi-type jobs and multi-type servers. *Queuing Systems* 70(3):269–298.

Wallace RB, Whitt W (2005) A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* 7(4):276–294.

Whitt W (1999) Using different response-time requirements to smooth time-varying demand for service. *Oper. Res. Lett.* 24(1):1–10.

Wood S (2006) *Generalized Additive Models: An Introduction with R* (Chapman & Hall/CRC, New York).

Wood SN (2003) Thin plate regression splines. *J. Roy. Statist. Soc. Ser. B* 65(1):95–114.

Verification Letter

Rick Jennings, Chief Technology Officer, Virtual Radiologic, 11995 Singletree Lane, Suite 500, Eden Prairie, MN 55344, writes:

"This letter is further to my earlier communication to you dated April 28, 2013 to express our continued support to the manuscript titled 'Business Analytics Assists Transitioning Traditional Medicine to Telemedicine at Virtual Radiologic' that has been submitted to *Interfaces* for review and publication. We partnered with the research team of Dr. Fatma Kilinc-Karzan and Dr. Sunder Kekre of Tepper School of Business, Carnegie Mellon University in August 2012 to develop a system for better forecasting and demand loading for our virtual network of radiologists around the country. The primary objective we set was to improve our turnaround time of reads (TAT) while minimizing system wide costs in our network.

"I am delighted to report that the analysis carried around within this project has provided good insights and has been critical for us in a number of different ways. The initial implementation of the tools presented in the paper has been completed. We are currently using the forecasting tools on a biweekly basis and the capacity planning system on a monthly basis for the review of existing licenses and the new licensing decisions. These tools are continually being enhanced depending on our needs. In particular, we have observed the following positive impact in our organization:

- The automated forecasting system has simplified the tasks of the operations planning personnel, and enabled the early detection of outlier behavior at the state and large facility level, which was not easily accessible to us before.
- Capacity planning optimization tool enabled us to easily study multiple scenarios and determine a robust capacity plan in terms hiring and licensing decisions of the radiologists.
- As a result, the overall need for the last minute adjustments to call in backup capacity from radiologists for handling sudden shifts in volume and mix of demand has been reduced, simplifying the task of the operations department on a daily basis.
- These tools in totality have been transforming the related analysis and decision-making process at vRad to a more transparent and rational basis removing much of the individual biases and preferences.

Overall, the improved accuracy of forecasting and better capacity planning has had a direct impact on our strategic decisions pertaining to radiologist selection, licensing and credentialing. The changes initiated by the CMU project has been done in parallel with other organizational projects and therefore it is difficult to isolate the benefits occurred at the project level. The TAT indicators have also shown a downward

trend (within the range of 12%–15%) in spite of our growing business and tighter capacity planning. System wide our operating costs after adjusting for demand growth has in the aggregate been reduced by 4% to 5%. The higher enterprise value (in millions of dollars) created as a result of these projects and policy changes has helped us enormously in convincing and bringing on board new owners. On the other hand, given the confidential nature of the new contracts and sale of the business, we are unable to disclose financial details."

Ersin Körpeoğlu is a PhD candidate of operations management at the Tepper School of Business at Carnegie Mellon University. His research interests lie in game-theoretical analysis of innovation and supply chain management problems, as well as business analytics and production scheduling applications. His work appears in several journals including *Interfaces*, the *European Journal of Operational Research*, and *Computers and Operations Research*.

Zachary Kurtz is a modeling analyst at PNC Financial Services Group. He obtained his PhD in statistics from Carnegie Mellon University (CMU). His thesis was on capture-recapture methodology. Prior to CMU, he was a research assistant for two years at the Federal Reserve Board in Washington, DC. Dr. Kurtz's interests include monetary policy, survival skills, prediction, and Appalachian music.

Fatma Kılınç-Karzan is an assistant professor of operations research at the Tepper School of Business at Carnegie Mellon University. She received her PhD in industrial engineering from the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. Her primary research interests lie in large-scale optimization methodologies especially related to handling uncertainty and their applications in business analytics and machine learning. Her work appears in several journals including *Annals of Statistics, Linear Algebra and Analysis, Mathematical Programming, Operations Research Letters*, *the SIAM Journal on Matrix Analysis and Algebra*, and the *International Journal of Production Research*.

Sunder Kekre, PhD, is the Bosch Professor of operations management and the director of PNC Center for Financial Services Innovation at the Tepper School of Business at Carnegie Mellon University. His research interests include interdisciplinary issues that examine the interface between operations and areas such as marketing, accounting, information systems, and engineering design. His work appears in several journals including *Interfaces, Management Science, Operations Research*, *Manufacturing and Service Operations Management*, and the *European Journal of Operations Research*.

Pat A. Basu, MD, MBA, is currently the Chief Medical Officer of Doctor on Demand. Previously, he served as Chief Operating Officer and Chief Medical Officer of Virtual Radiologic. Prior to this executive role, Dr. Basu served as a White House Fellow and worked at the highest levels of the federal government, helping to shape public policy. Throughout his career, Dr. Basu has advised Fortune 500 firms, venture capital firms, and major healthcare systems. He continues his commitment to service as an appointed member of the Council on Foreign Relations and of the Illinois State Board of Health. He also serves on the Board of Governors of the Stanford Medical Alumni Association and the Board of Directors for the University of Illinois Alumni Association. Dr. Basu formerly served as chief resident at Stanford University. He holds a medical doctorate, an MBA from the University of Chicago, and a mechanical engineering degree from the University of Illinois.