Submitted to *Interfaces*
manuscript (Please, provide the manuscript number!)

Business Analytics Assists Transitioning Traditional Medicine to Telemedicine at Virtual Radiologic

Ersin Körpeoğlu, Zachary Kurtz, Fatma Kılınç-Karzan, Sunder Kekre
Carnegie Mellon University, 5000 Forbes Ave. Pittsburgh PA 15213,
{ekorpeog, zkurtz, fkilinc, sk0a}@andrew.cmu.edu,

Pat A. Basu, Chief Medical Officer, MD, MBA
Virtual Radiologic, 11995 Singletree Lane Eden Prairie, MN 55344,
pat.basu@vrad.com

Virtual Radiologic (vRad), the largest teleradiology company in the United States, faces the difficult problem of matching its more than 400 radiologists with its seasonal demand. In addition to the constraints that traditional medicine 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 assesses demand more accurately and plans capacity in the system to provide better service to its customers. We discuss the underlying reasons for improvement and quantify the impact on vRad's entire system. We also explain managerial insights that will help not only vRad, but also other teleradiology and telemedicine companies.

Key words: Capacity Planning, Forecasting, Optimization, Healthcare Industry
History:

Introduction

With more than 400 radiologists, 2700 client facilities, 8 million radiology images processed annually, and an annual revenue of $300 million approximately, Virtual Radiologic (vRad) is the largest national provider of radiology in the United States. By integrating the on-site presence with the power of cloud computing, vRad's value proposition is, simply, to deliver the best clinical result for every patient by selecting the radiologist with the most relevant specialized training, in the least "turn-around-time," anywhere, anytime.

On a 24/7 basis, through customized high technology equipment located at client facilities, i.e., clinics and hospitals, medical images (hereafter jobs) are sent to vRad's cloud and then distributed to be read by vRad radiologists spread all over the US. The client and radiologist spread of vRad is displayed in Figure 1. Radiologists are the most expensive resources of vRad but are the cornerstone of its business and its most important assets in delivering the vRad's commitment to quality. Like many other medical practices, in order to achieve the best care outcomes, it is essential to assign each medical image to the right doctor. To provide the highest quality of

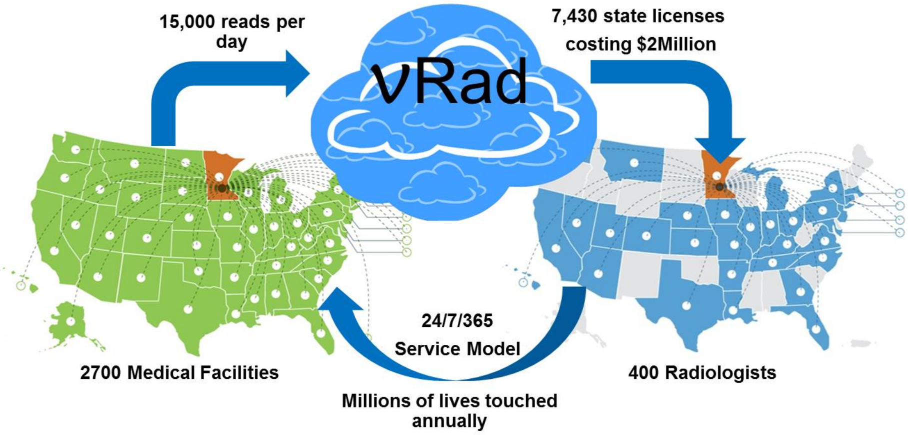

Figure 1 **The client and radiologist spread of vRad. The company is based in Minnesota.**

patient care, vRad radiologists are highly specialized, and 75% of them are trained in an array of sub-specialties, including, but not limited to, neuroradiology, musculoskeletal radiology, and pediatric radiology. For many traditional radiology departments in hospitals, having such an array of radiologists with special training at all times is simply impossible. On the other hand, as opposed to the traditional radiology departments in hospitals, vRad radiologists work flexible hours in the comfort of their homes via their own specialized work center equipment. vRad optimizes capacity utilization by pooling demand and supply 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 trying to staff itself with on-site radiologists cannot. Enabled by smart technology solutions and a large client base; better capacity utilization via resource pooling underlie vRad's ambitious business model.

In fact, this high technology orientation and large customer base allow vRad radiologists to achieve significantly higher productivity rates without sacrificing quality. An average vRad radiologist reads more than 25,000 studies (jobs) annually, significantly higher than the industry average while maintaining an exceptional 99.7% accuracy rate (Nicodemus (2010)). Moreover, considering the current supply and demand trends, and accounting for 32 million newly insured patients, Locumtenens.com (2010) estimates a growing need for higher radiologist productivity due to an estimated 16% gap in radiology supply versus demand by 2020. The projected gap between supply and demand points to the growing strategic importance of utilizing the high productivity rates of teleradiology.

With the advances in technology, the telemedicine industry, with an estimated $3.6 billion value (Monegain (2009)), offers flexibility and significant savings in the associated medical costs yet it comes with new operational challenges that were unknown to the classical medical practices. In an environment where a student in rural Georgia is diagnosed by a pediatrician in a far away state, the operational complexity reaches to a whole new dimension leading to four brand new challenges. The first challenge is to manage capacity effectively at the national level. Telemedicine requires a highly complex licensing process, e.g., in order for a doctor in one state to read in another, the doctor should be licensed by that other state. State licensing usually takes 3 to 8 months and is quite costly. Moreover hospital networks require hospital privileges (credentials) in addition to the state licenses for the doctors in order to diagnose medical images. Secondly, at the operational level, daily personnel management is no longer a small optimization problem faced at a local hospital level, but instead involves matching hundreds of specialized doctors licensed at various states with thousands of patients. Moreover, hospitals downsize their workforce capacity and rely on telemedicine at nights and weekends which causes huge variation in demand pattern for telemedicine companies. Thirdly, the success of the business is strongly tied to its promise on an incomparable efficiency gain, i.e., the productivity rates of doctors increase extremely due to higher flow of patients and thus 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, faster service, and lower costs. In the medical industry not only the service level is critical but also the service time is of utmost importance, which requires crucial analysis of the system at the level of minutes instead of hours or days which is traditionally used in manufacturing or other service environments. Therefore, there is a constant pressure on improving business operations, in particular, accurate demand forecasting at the level of minutes and comprehensive models for resource management play a vital role in these time-sensitive environments. The second challenge is to optimize the staffing structure to deal with huge temporal variation in daily telemedicine demand.

To assist vRad to overcome the challenges we detailed above, 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 the contracted radiologists to design a robust plan to handle various scenarios of demand. The real-world problem is 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. Also, the data have been masked to protect the company confidentiality and HIPAA privacy guidelines.

The remainder of the paper is organized as follows. We first describe the managerial problem in the next section. In particular, we summarize the operational challenges and the legacy process. Then, in the following section, we explain our solution approach. There are two phases of our approach: (1) data collection and (2) strategic/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 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 in vRad's business as follows:

- Cost of radiologists on shift
- Costs related to over- or under-utilization of radiologists
- Cost of licenses and credentials needed to run the business
- Opportunity cost of poor service (cost of losing business with clients)

One can easily notice some conceptual overlap between the items on the list above. For example, reducing the number of radiologists is likely to go hand-in-hand with increased 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 licensing requirements and sub-specialty expertise, etc.

In the following subsection, we provide an exploratory analysis of the 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 hours (see Figure 2 for a distribution of the number of radiologist working in a given hour, i.e., in a "typical" hour, there are between 40 and 120 radiologists reading at least one job for vRad). While vRad needs to have the appropriate radiologists available on shift in order to be able to process the images promptly, the schedules for the radiologists are created in advance. Therefore, at any given week, the list of radiologists that are contracted to work are fixed. In case of emergencies, i.e., when the composition and/or number of radiologists is not proper to cover the demand at the moment, vRad requests some of its off-shift radiologists to work additional shifts. Due to 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 for every state for which she/he can read jobs from. In addition, radiologists must be separately privileged at the hospitals sending jobs to vRad. This results in an extremely complex and expensive licensing and credentialing process, involving a total of 7500 state licenses (for all radiologists) and tens of thousands of facility credentials. Figure 3 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 4 displays the actual state level demand against the scaled available capacity per state, where the capacity of each

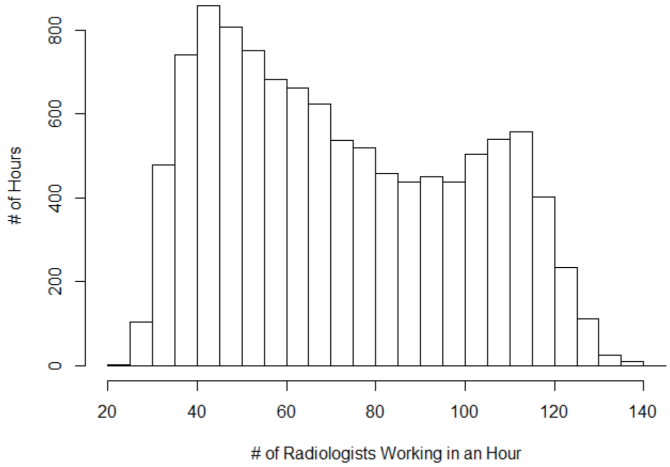

Figure 2 **Histogram of # of radiologist working in each hour. In a "typical" hour, between 40 and 120 radiologists will read at least one job for vRad.**

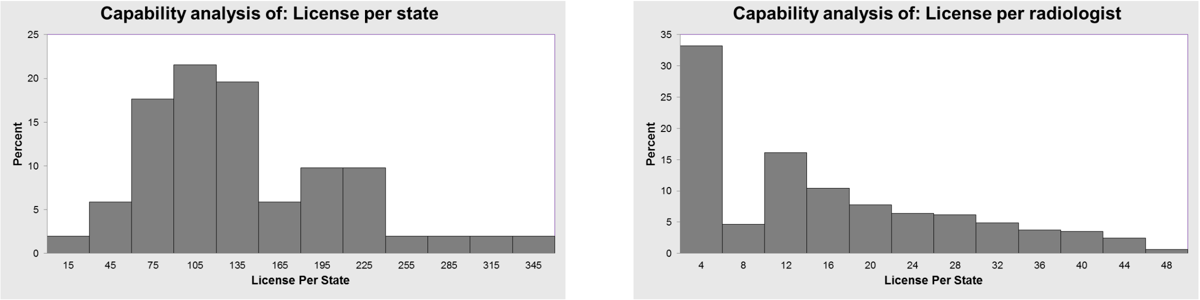

Figure 3 **The distributions for (i) number of radiologists licensed per each state and (ii) the number of state licenses per radiologist. For the graph on the left hand side, the vertical axis represents the percentage of states with the given number of licensed radiologists. Most states are covered by many radiologists, i.e., for only 2% of the states, there are 30 or less radiologists licensed in. For the remainder of states, there are significantly more licensed radiologists, i.e., for 4% of the states there are more than 300 radiologists. For the graph on the right hand side, the vertical axis represents the percentage of radiologists with the given number of state licenses, e.g., 33% of the radiologists have 6 or less state licenses. As the graph depicts, there is a high variance of the number of state licenses per radiologists.**

radiologist is allocated proportionally to the states in which she/he is licensed to based on the total demand from those states. Such an apportioning is not truly reflective of the actual assignment of capacity used to cover state level demand. Nonetheless, it still shows signs of the "mismatch" of demand versus capacity in certain states (such as states 12 and 39).

While some jobs may be read by nearly all radiologists, other jobs can only be read by radiologists that are licensed for relatively uncommon sub-specialties. This complicates the process of matching capacity with demand.

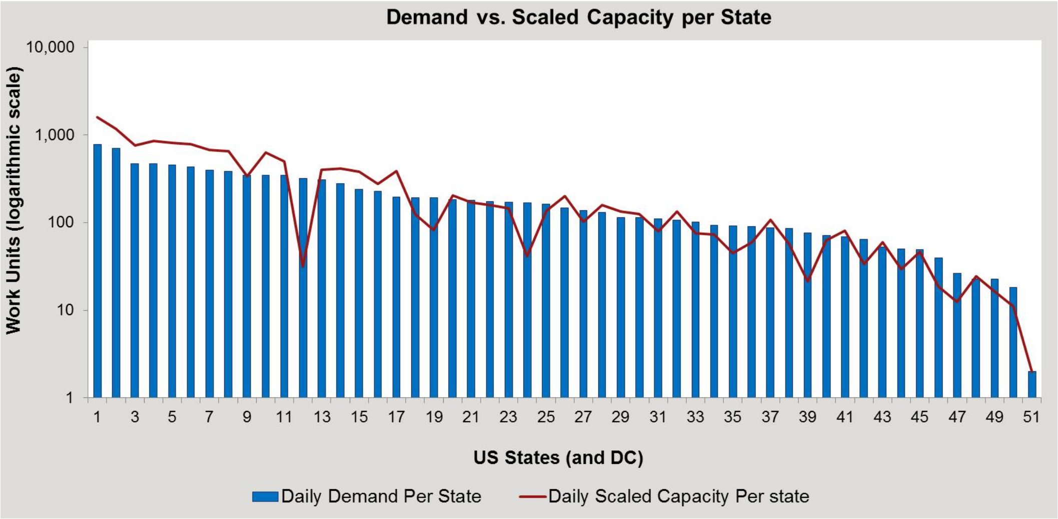

Figure 4 The demand versus total scaled service capacity for each state. The capacity is calculated as follows: For each state, the capacity of each licensed radiologist is normalized based on the total demand of the states that the radiologist serves. Then the normalized capacities are summed to get the scaled capacities for each state. The scaled capacity metric measures how fairly licenses are distributed across states based on the actual demand. However, the metric does not indicate whether there a state is over or under licensed.

Analytics of Demand

The time required for a radiologist to process a job depends on the job type (for example, MRI scan versus Xray). 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 of a prediction of the number of work units in each future time period. Designing a forecast involves significant challenges, as the demand in work units is noisy and non-stationary, and shows seasonal, weekly, and daily cyclic structure. Figure 5 shows the hourly demand (in terms of work units) as well as a drastic change in demand structure after the acquisition of NightHawk in early-2011. Disregarding the acquisition in March 2011, a mild annual seasonal affect is visually evident in Figure 5, with elevated demand in late summer, and reduced demand in early winter.

In addition to this, there are weekly patterns, with Saturdays and Sundays having about 18% higher demand on average as compared to weekdays. Figure 6 explores 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 includes jobs that are needed in especially time-sensitive medical conditions and comprise approximately 94% of the work units. Non-emergency jobs account for the remainder of the work units. It is clear from Figure 6 that the daily pattern depends both on the time of the week (weekend versus weekday) and on the type of read request (emergency versus non-emergency).

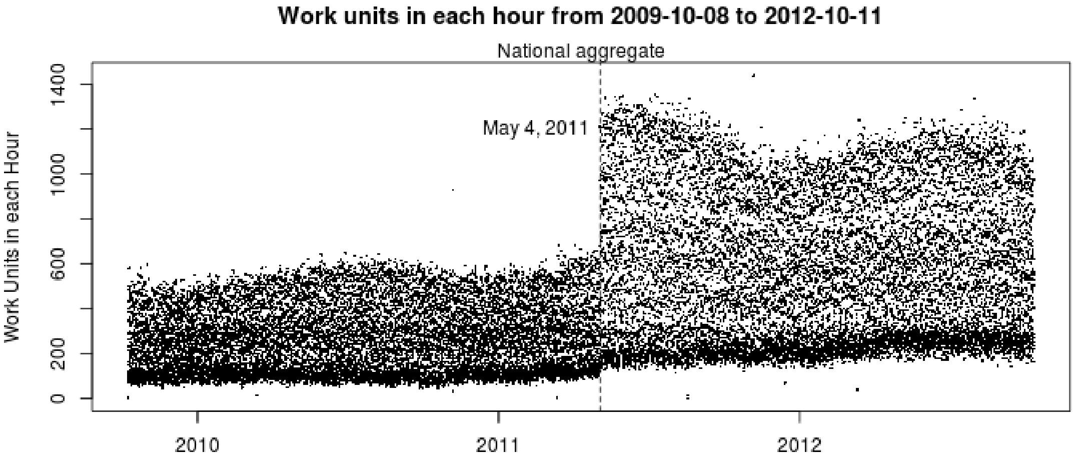

Figure 5 **National hourly demand. Each point indicates the aggregate demand for a specific hour. The pattern displays a huge jump after the acquisition of NightHawk. Moreover, there is an apparent seasonality in demand.**

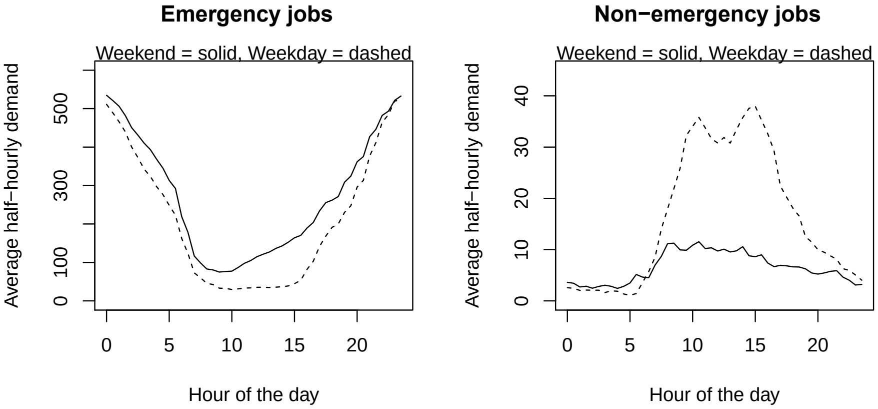

Figure 6 **Daily pattern averaged for weekdays and weekends.**

Figure 6 shows a clear interaction between the weekly variation in demand and the daily variation in demand, since the shape of the demand over the hours of the day is different for each day of the week. Furthermore, the distribution of demand across facilities (See Figure 7) is highly skewed, with a large number of very small facilities that send only a few jobs per day. Generating an accurate forecast at the individual facility level for shorter time intervals such as 30 minutes is exceptionally challenging, as demand may be intermittent.

Key Performance Metrics

In addition to these operational constraints, vRad has two key performance measures contributing to the operational complexity:

1. *Turn-Around-Times (TATs):* TAT is the time from a job's arrival to 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 and customer satisfaction.

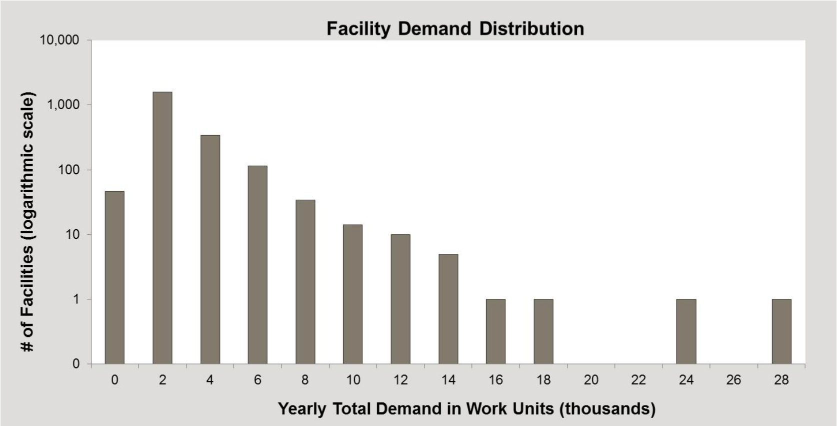

Figure 7 The yearly total demand distribution of facilities that vRad serves in logarithmic scale. The majority of the facilities (1600 out of 2700 vRad clients), on average, demand 5-6 jobs per day.

vRad has four job priority classes depending on the urgency of a job. For high priority jobs, the completion time of a job is critical. Emergency jobs constitute the highest priority class – it is imperative that these jobs are processed as quickly as possible[1].

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 pay is based on the total number of work units completed. Thus, it is critical for vRad to effectively utilize its radiologists that are on-shift. On top of high utilization, vRad seeks a fair distribution of utilization among its radiologists for their happiness, yet Figure 8, i.e., the distribution of number of jobs processed by the radiologists, clearly shows a big spread.

Problem Diagnosis

Initial analytics of supply and demand confirmed that the successful daily operation of vRad requires a way to efficiently match qualified radiologists with the jobs. In fact, state licenses constitute a large cost item in vRad's budget and therefore vRad consistently tries to manage them carefully in relation with the performance of schedules. Thus, it is important to accurately predict demand in order to assess the number and types of each license to be maintained. Scaling the capacity of each radiologist with respect to the relative demand (as displayed in Figure 4) makes sense for aggregate capacity planning purposes, i.e., whether the licenses are distributed in proportion to demand, yet this measure is not very useful for understanding the over- and/or under-licensing phenomena. To have a better understanding of the over- and/or under- licensing issue, for the same day in 2012, in Figure 9, for each state, we plot its total demand against its

[1] vRad has a company policy of serving emergency jobs within half an hour.

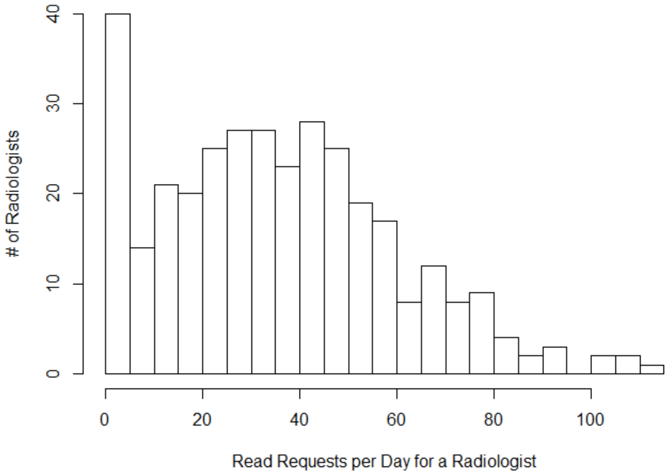

Figure 8 **Histogram of # of read requests processed per day. Some radiologists process a lot more jobs than others. In 2012, while 40 radiologists processed less than 5 jobs per day, on average, there were five radiologists averaging more than 100 jobs per day.**

total available capacity, which is calculated as the maximum amount of service possible to a state s if all the radiologists licensed to state s serve only state s. While this capacity is not reflective of the operational capacity to serve state s, it provides an idea about the total licensed radiologists to that state. As the figure depicts, vRad takes a conservative approach to the licensing problem, leaving ample excess capacity for each state. This conservative approach is meant to minimize turnaround times by making sure that almost all of its radiologists can serve demand from every state. Also given the time it takes to obtain a license, it enables them to quick start serving a large customer in any state. Yet, with better demand forecasting and ways to assess risk, we will argue that vRad can responsibly reduce its overall spending on radiologist costs and state licenses without sacrificing from its quality of service, i.e., without increasing average turnaround times.

Legacy Process

Previously, vRad had a two stage forecasting system conducted by two separate departments. vRad's finance team had a labor intensive process of generating quarterly demand forecasts using Excel, with forecasts generated primarily at an aggregate level. Long-term forecasts were updated infrequently, and the scheduling department made relatively frequent 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 long term forecasts made by the finance team were made only quarterly, and thus tended to miss short term trends. Second, short term scheduling forecasts only looked at short term trends and did not build on long term fundamentals. Third, forecasts did not capture variability to allow safety stock type decisions. Finally, secondary to legacy IT tools, forecasts were generated for the aggregate

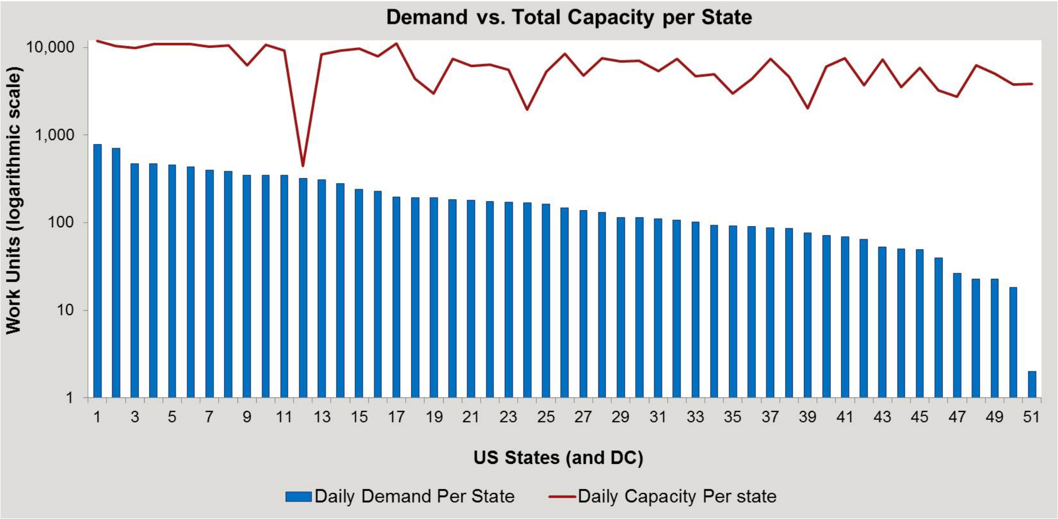

Figure 9 **The demand and total service capacity for each state. The capacity is calculated as the maximum possible service to a state. The figure depicts that there is a substantial gap between the capacity and demand. For instance in state 48, the realized demand is less than 1% of the allotted capacity.**

demand at the national level, missing much-needed facility- and state-level granularity for capacity planning.

On the supply side, the scheduling team constructed bi-weekly 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, which is hard coded 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 its largest competitor, NightHawk, in 2011, the operations volume almost doubled, raising operational complexity to a new level. Although the vRad staff is experienced and qualified, the labor intensive planning approach inevitably lacked the ability to consider all of vRad's complexity, creating inefficiencies. In particular, regular capacity shortfalls had 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 forecasting and capacity planning, we proposed an automated forecasting and planning tool (see Figure 10 for an illustration of our approach). As the figure depicts, the Forecasting Model creates the short term and long term forecasts. Subsequently, licensing decisions are based on the analysis provided by the Loading Model. Finally, given state licenses and short term demand forecasts, vRad's daily capacity is planned using the Loading Model. As a result of the capacity plan, the set of radiologists that will be on shift are determined. The output of our forecasting and planning tool is inputted to vRad's study assigner. Finally, real demand is realized and jobs are allocated to available radiologists via vRad's study assigner.

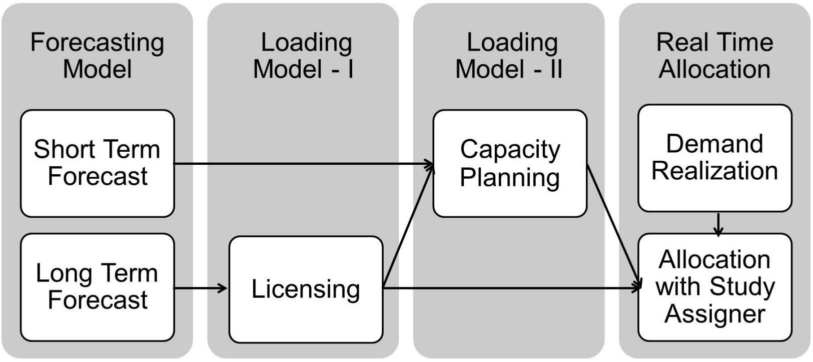

Figure 10 **Our solution methods are used in three different planning stages and support vRad's study assigner.**

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 extends into October 2012, so there is enough data for a 1-year training set and a 60-day forecasting test set. Disregarding the NightHawk acquisition, a mild annual seasonal affect is visually evident in Figure 5, with elevated demand in late summer, and reduced demand in early winter. Our regression model incorporates a term for this annual seasonality. In addition, a coefficient for time captures the long-term growth trend.

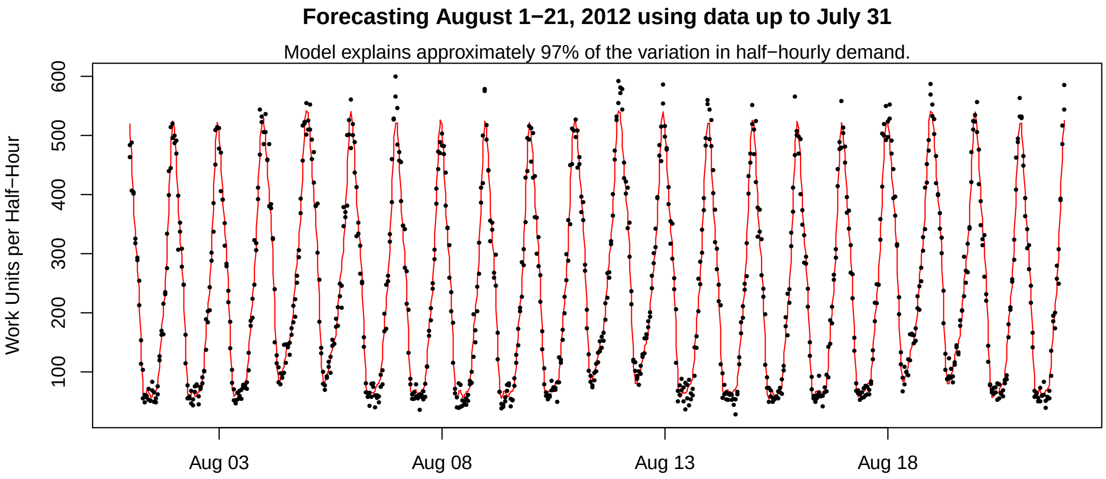

Figure 11 **Forecast (red curve) and actuals (black dots) over the first three weeks of the test set.**

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 11, both for the observed half-hourly demand (black points) and for the forecast (red curve), which tracks the daily pattern closely. A relatively subtle weekly pattern is visible under close examination of the forecast curve in Figure 11, with Saturdays and Sundays having an average level of demand about 18% higher than for weekdays. Moreover, Figure 6 shows a clear interaction between the type of read request (emergency versus non-emergency), the weekly variation in demand and the daily variation in demand, since the shape of 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 pattern. Each of these effects is represented by a term in an additive model. An additive model is a modern regression framework that allows us to replace ordinary regression coefficients with rather general smooth functions (Hastie and Tibshirani 1990). We used the `mgcv` package in `R` (R Core Team 2012) to estimate the smooth functions nonparametrically as penalized regression splines. The purpose of introducing this extra machinery is to capture nonlinear effects in the model. 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 11 is more like a smooth sine-like wave, and any such smooth pattern can be represented very accurately in an additive model (Wood 2006).

The residuals from our forecasting model tend to be autocorrelated, accumulating to the positive or negative side for brief periods. Figure 12 shows the regression model forecast residuals for the 60-day test set as the black dots. 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 pattern. 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 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 minimum mean square error. Figure 12 displays the ARMA next-period forecast (red line) of residuals for the 60-day test set.

The Model: The foundation of our demand forecast is an additive regression model that regresses demand on the trend and seasonality effects of demand, which is measured in work units. The underlying mathematical formulation is provided in Appendix A. Our forecasting model can explain as much as 97% of the variation in half-hourly demand at the aggregate national level. To be precise, the variance of the residuals is approximately 3% as large as the variance of the demand. This error rate translates to a mean absolute percentage error of approximately 9% for out-of-sample testing. Local autocorrelation trends are often evident in the residuals of the regression model, and we use these trends to build a real time adjustment for the next-period forecast,

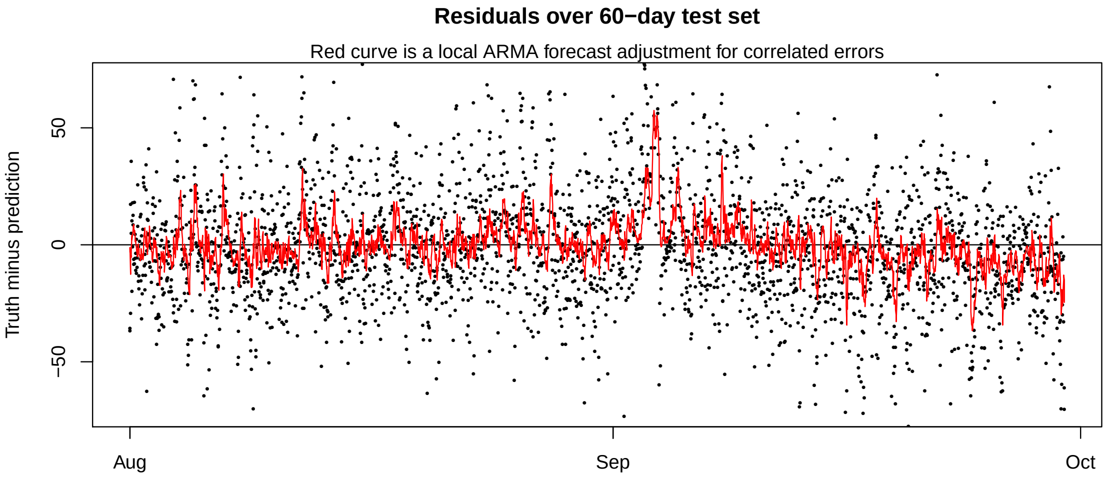

Figure 12 **ARMA residual forecast for correlated errors (red curve) and regression model residuals (black dots) over the 60-day test set.**

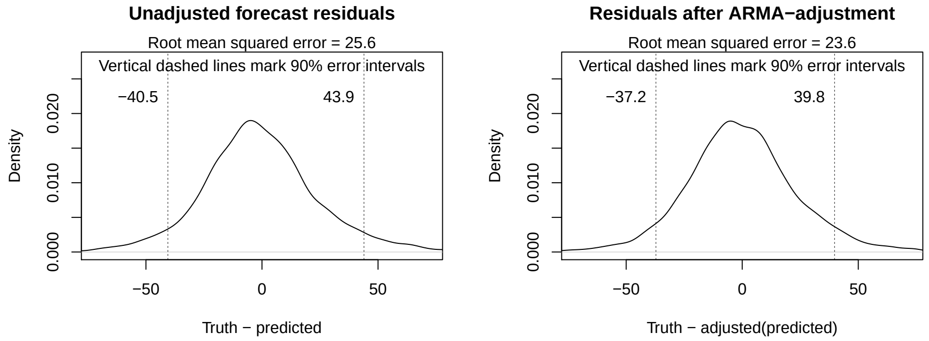

Figure 13 **The curve in each plot is a smoothed histogram of residuals for the national hourly forecast over the 60-day test set. The residuals in the right panel are slightly smaller due to the adjustment based on the local ARMA model.**

leading to a moderate accuracy improvement for short-term forecasting. 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% to approximately 8.7%. Figure 13 shows density estimate for the residuals before and after the ARMA adjustments. In particular, the width of the 90% prediction interval shrinks noticeably, and the root mean square error falls from 25.6 work units to 23.6 work units.

While we have described our forecasting approach for the half-hourly demand aggregated over all of the facilities in the nation, we also present vRad with an option to further aggregate the data for an hourly, 8-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. On the other hand, forecasting demand separately for each state, facility, or specialty raises some issues. While our forecasting model may be applied based on data at any level of disaggregation, the quality of the forecast changes with disaggregation.

For example, Figure 14 illustrates the forecast that results from fitting the regression model for emergency demand data for a large state. It is clear that the ratio of noise to signal is much greater for this cut of the data than for the national aggregate (Figure 11). This difference in accuracy between the state forecast and the national forecast is perhaps best explained by the fact that the variance of a mean grows as the sample size shrinks.

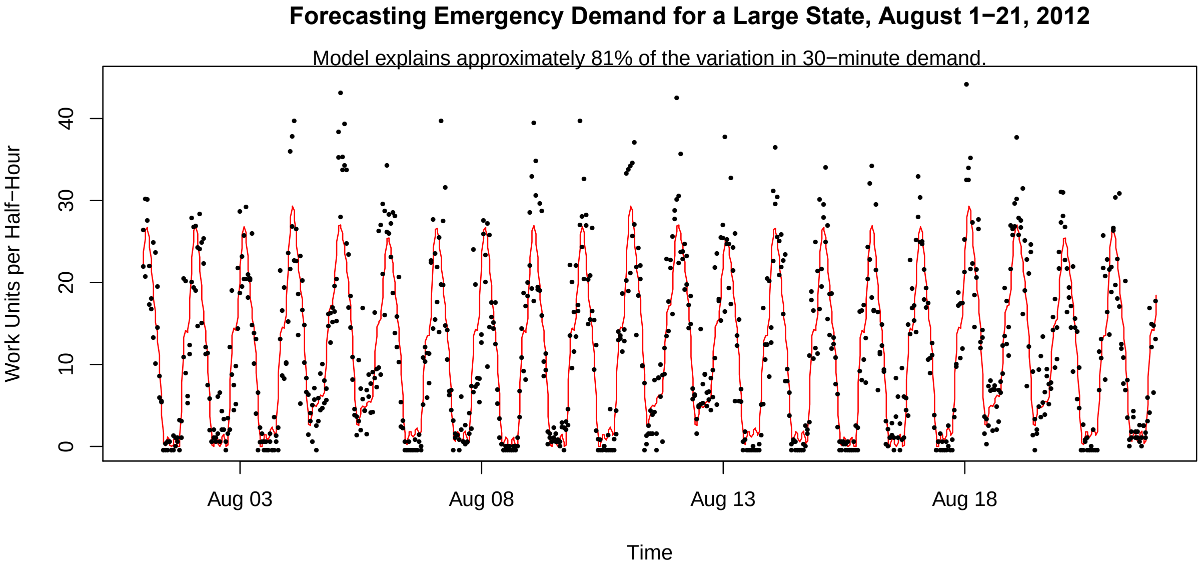

Figure 14 **Forecast (red curve) of emergency demand in a large state (black dots) over the first 21 days of the test set.**

Loading Model

The Loading model is a Linear Programming (LP) based diagnostic and planning tool that lets vRad analyze on a given planning horizon, whether they have the right kind and number of radiologists scheduled to meet the expected demand and whether the current set of licenses suffice to serve all demanding facilities efficiently. The Loading Model takes the demand forecast as a primary input.

Literature and Modeling Choice: For the loading model, there are three streams of related literature. The first stream is scheduling and capacity planning. Graves (1981) provides a comprehensive review of scheduling literature and Pinedo (2008) 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 neither tractable nor is there a state of the art scheduling algorithm, to the best of our knowledge, 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 health care or emergency room (e.g. Beaulieu et al. (2000)) and OR room (Batun et al. (2011)) scheduling problems. Ernst et al. (2004) provides a comprehensive survey of this literature and argues that, due to tractability issues, health care and flexible workforce scheduling studies focus on specific aspects of health care planning problems. In vRad's case, however, considering both supply and demand side of the problem is essential.

A second stream of literature pertains to call center capacity management (for examples, see Whitt (1999), Armony and Maglaras (2004), and for a comprehensive survey, see Gans et al. (2003)). Call center studies do not apply to vRad's environment due to the following three properties. First, vRad has highly variable demand, with daily, weekly, and annual seasonality. Second, radiologists can only serve a subset of jobs due to licensing, credentialing and sub-specialty restrictions which makes the corresponding queuing model extremely complicated. Finally, the number of call centers and job types along with the priority classes renders such a model intractable.

A third stream of related literature is on capacity rationing (e.g. Swenson (1992)) and capacity pooling (e.g. Ata and Van Mieghem (2009)) problems in Health Care. Capacity rationing papers investigate efficient distribution of scarce capacity in clinical facilities and intensive care units. They focus on how to fairly select patients to serve (Swenson (1992)), when to discharge patients (Chan et al. (2012)) and overbooking patients in clinics in order to be able to serve more patients (Lee and Zenios (2009)). Capacity pooling literature investigates the benefit of pooling resources such as hospital beds (Best et al. (2013)), rooms (Vanberkel et al. (2012)) or machinery (Mahar et al. (2011)) to achieve economies of scale or, on the other extreme, focusing resources on different patient groups to improve patient care and achieve economies of focus. This question may be translated to general licensing strategy of vRad and could be used to assess whether it is more beneficial for vRad to dedicate radiologists to states rather than licensing them to more than one state. Although the models from 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, in order to keep the model tractable and focused, while still capable of solving very large instances in a short amount of time. Our model builds upon the demand forecast. Since demand is variable, we discretize the planning horizon. In a short term analysis, we break 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 provide 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.

Thus, all periodic capacities and limits are in work units. The inherit 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 use a deterministic model instead. On the other hand, as a mitigation strategy, we equip the users with automated tools for making proper sensitivity analyses and checking the performance of the system under several scenarios.

After careful analysis of demand structure and in conjunction with vRad, for short term forecasting and capacity planning purposes, we decided to cluster all the facilities (except 200 largest facilities) by state. This decision was mainly due to the difficulty of accurately forecasting the demand for small sized facilities that send 5-6 jobs a day (As Figure 7 depicts, approximately 60% of the facilities fall within this category.) After grouping by state, the number of facilities (or facility state-aggregates) that would be inputted to the loading model became 254 including 200 largest (in daily demand volume) facilities, 50 US states, Puerto Rico, District of Columbia, pro bono jobs, and government categories.

The Model: Loading model uses an LP that matches the demand with the supply of available radiologists (see Appendix B for the complete mathematical formulation).

The objective tries to minimize the weighted average turnaround time (TAT), with weights assigned to different priority classes to move higher priority jobs forward in the queue. In conjunction with the company, the weights are assigned so that the higher priority jobs are first served as soon as possible and then the remaining capacity is used for the lower priority jobs. The other critical factor for vRad, 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. Increasing the radiologist utilization is mainly handled using sensitivity analysis, as we discuss further in the Impact Section.

Each radiologist has a different periodic reading capacity, which is embedded to the model via the first constraint. The second constraint sets a lower and an upper bound on the total amount of work units that a radiologist can have within a planning horizon. While the lower bound tries to ensure the workload of radiologists are balanced, the upper bound ensures that the total amount of workload of a radiologist in a shift is within a reasonable level. The third constraint balances the number of jobs demanded, processed, and backlogged between periods. In order for a radiologist to be able to read a job in a given period, the radiologist should be available and should have the necessary licenses, credentials, and sub-specialty expertise. These restrictions are encoded in the last constraint.

The model is coded in 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. A challenging issue while implementing the loading model was the extreme size of the problem instances, easily leading to hundreds of thousands of constraints and variables for a small scale problem even after preprocessing. For the large scale instances, the sizes of the problem instances extend to a level where it was impossible to construct and load the model into the computer memory. Therefore we proposed a decomposition algorithm, referred to as an "iterative backlog algorithm" that exploits the specific sparsity structure of vRad's problem instances. The algorithm, as presented in Appendix B, solves the loading model for different smaller instances iteratively, and is especially useful in medium and long term analyses where more computing power is necessary.

Impact

Implementation Experiment and Scenario Analyses

The methodology to estimate the impact of the loading model involved many pilot implementation experiments using real data from vRad. As we discussed in the previous section, while the model included millions of variables and constraints, we were able to solve it within minutes. This efficiency of the model allowed us to make a variety of sensitivity analyses. Figure 15 demonstrates the results of one such analysis. In the figure, on a given day, the total demand of jobs and the total amount of jobs read (supplied) are illustrated. When radiologist capacity is decreased by 20%, it is impossible to keep up with peak demand, and some jobs are backlogged to future off-peak periods. This use of backlogging in conjunction with temporary understaffing of radiologists hurts the turnaround times. On the other hand, understaffing improves average radiologist utilization. Figure 16 depicts the gradual change in utilization distribution of radiologists as the capacity levels decrease. At full capacity, only 34% of radiologists work with almost full (90-100%) utilization. By contrast,

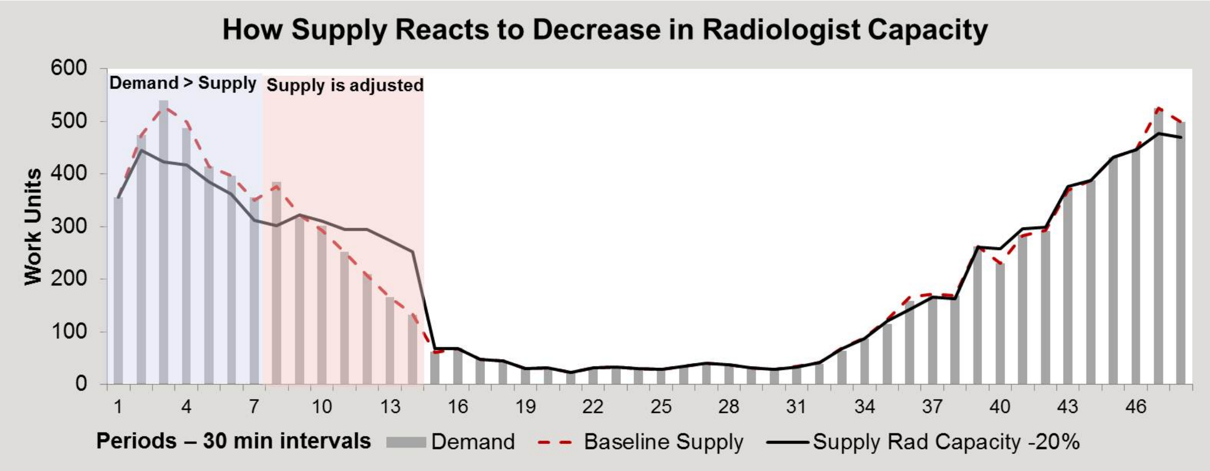

Figure 15 **Loading model, the reaction of supply a reduced capacity.**

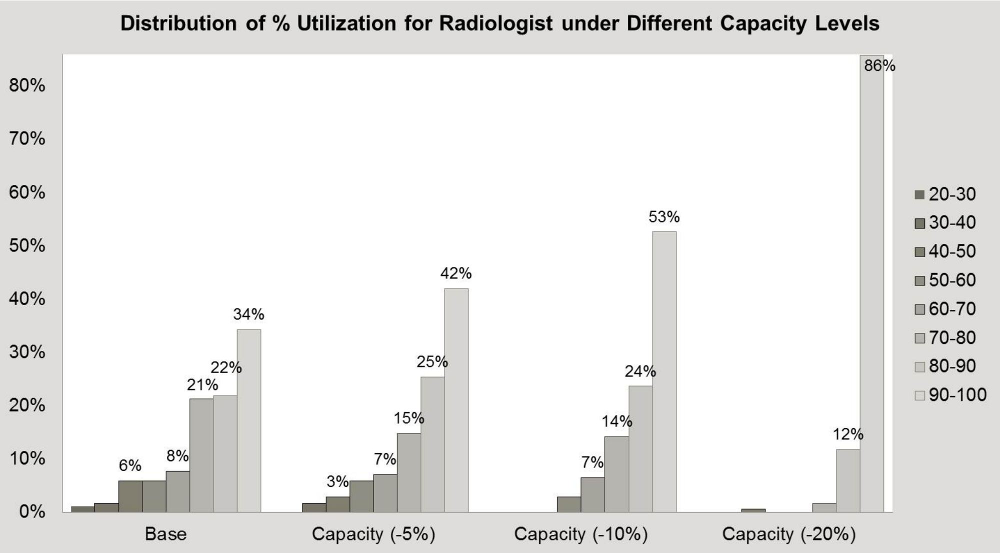

Figure 16 **The distribution of radiologist utilization ($100 \times \frac{\text{Total Service}}{\text{Total Total Capacity}}$) given different levels of radiologist capacities. The figure depicts the percentages of radiologists with utilizations in different levels such as 20-30% or 90-100%. According to the figure, the total percentage of radiologists with almost full utilization (90-100%) jumps from 34% to 86% when there is a 20% decrease in radiologist capacity (or service rate).**

the percentage of radiologists with almost full utilization increases to 86% when the radiologist capacity is reduced by 20% albeit at the expense of a 33% increase in the average turnaround times (See Figure 21 for the average TAT under different capacity levels). This type of an analysis allows vRad to quantify the trade-off between TATs and utilizations. Another fact demonstrated by Figure 16 is that a small percentage of the radiologists have low utilization levels, which prompts us to ask how efficient the radiologist schedules were for the given day. In particular, one could ask whether it is suboptimal to have a low utilization radiologist be on shift. Figure 17 depicts the change in supply as the radiologists with low utilizations are removed from being on shift. As the graph illustrates, the turnaround times worsen as the gap between demand and supply increases. However, this level of change may be acceptable (only 0.5% and 1.5% respectively), given the fact that the utilization rates are improved significantly. The figure also demonstrates that the supply of all cases (with usual and reduced number of radiologists) tends to overlap after period 10, i.e. after the peak period is realized. This analysis suggests that vRad can reduce the number of radiologists on shift by 3% with only a 0.5% increase in the average turnaround times.

Insights and Observations

While conducting the project with vRad, we made several key observations that would serve as useful managerial insights.

a) Eliminating Over-licensing: As Figure 9 depicts, vRad was conservative in licensing process leaving ample excess capacity for each state. For instance, for some states the demand is

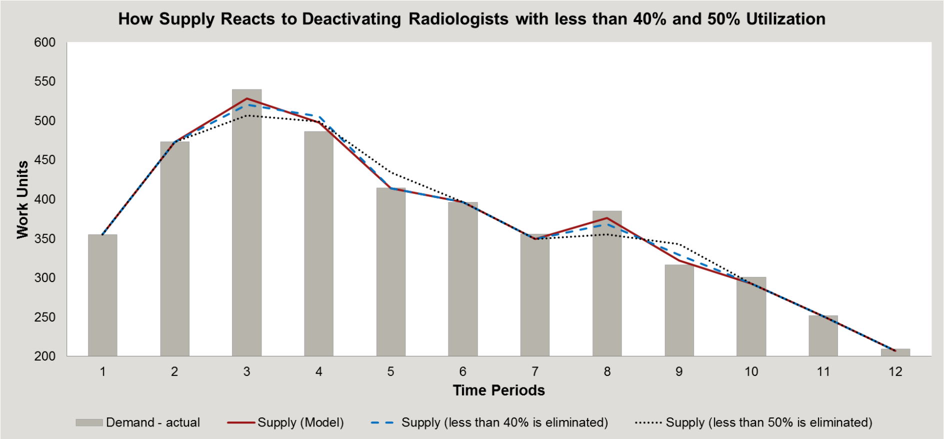

Figure 17 **Supply in loading model reacts to reducing the number of radiologist on shift. Specifically, the supply level with the original set of radiologist is higher than the supply level when the radiologists with less than 40% and 50% utilization are eliminated. The figure includes only time periods 1 to 12 (among 48 periods) because supply has the same structure for all cases through periods 12 to 48.**

less that 1% of the capacity. The conservatism in licensing is understandable given vRad's desire to keep turnaround times as low as possible and to quickly serve new customers without waiting for new state granted licences. However, with the improved demand forecast and better planning, vRad now is able to use the loading model to make long term planning and see the impact of reducing licenses even with a conservative demand estimation. This allows vRad to quantify the risk of reducing licenses, saving the company from a significant amount of licensing costs.

b) Utilizing State-wise Aggregation: A state-wise aggregation of facilities would significantly improve the effectiveness of planning. The need for aggregation is evident from the difficulty of forecasting the intermittent demand typical of individual facilities. Figure 18 shows the forecast and demand series for a typical facility in vRad's network. The residuals for this forecast are much smaller than the residuals for Figure 11 due to the fact that overall volume at a single facility is much smaller than the national or state-wise aggregate. However, the forecast for demand at this facility explains only about 23% of the variation in demand, down from 97% for the national aggregate demand forecast. Due to the inaccuracy of facility level forecasts, vRad can significantly improve its planning by clustering the facilities in the same state. This type of a clustering also reduces problem instances for loading model allowing more scenario and what if analyses. Although state-wise clustering is useful in many dimensions, it requires alignment of the state licenses with the facility credentials. Therefore, it is advisable to credential a radiologist to all facilities in all states she/he is licensed to. For example, if a radiologist is licensed in Pennsylvania, she/he should be credentialed by all facilities in Pennsylvania. Given the large number of facility credentials, this kind of a change would not mean increasing the number of credentials, but simply reshuffling the

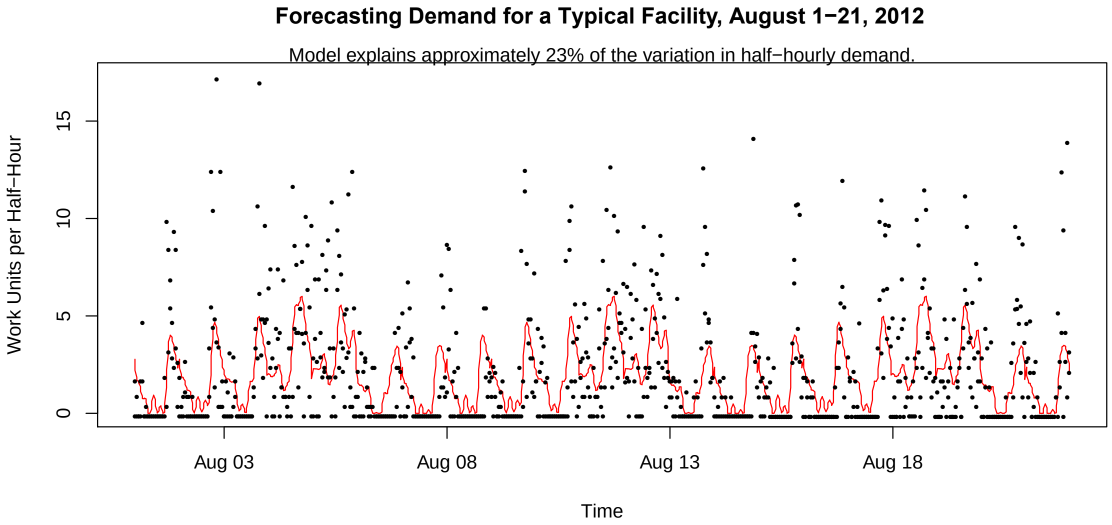

Figure 18 **Forecast (red curve) and actuals (black dots) over the first 21 days of the test set for a typical facility in vRad's network.**

current set of credentials. Yet we also recognize the limitation of this recommendation as many hospital facilities are slow or reluctant to grant or change physician privileges.

c) Handling Intermittent Demand: From Figure 18, we notice that a large fraction of the time periods have zero demand. More generally, the distribution of residuals is asymmetric (and therefore not exactly normal). This asymmetry is even more pronounced for smaller facilities, where hourly demand begins to look more like an over-dispersed Poisson distribution rather than a normal distribution. At some low level of demand, a regression model forecast delivers little increase in accuracy relative to simply using the historical mean demand level as the forecast. For such intermittent demand series, it may be appropriate to estimate the full distribution of demand levels rather than producing a point estimate for the expected demand level (Snyder et al. 2012). Several additional forecasting approaches for intermittent demand series appear in Leonard et al. (2008). These intermittent demand forecasting models all have in common that they reduce prediction variance by averaging (at least implicitly) across a large number of adjacent time periods. While any of these approaches could yield interesting results in our application, we have not implemented them. We suspect that applying our forecasting approach for the time-aggregated series – such as a daily series instead of a half-hourly series – will yield comparable results.

Experience and Organizational Impact

The partnership with vRad and CMU not only lead to the creation of loading and forecasting models, but also opened up new horizons for vRad. With a more accurate planning process in effect, vRad is now able to focus on other related operational components. For example, vRad has now the opportunity to enhance its study assigner tool using the inputs of loading and forecasting models. This in turn can allow vRad to improve key performance metrics such as TAT distributions and license utilizations. For instance, Figure 19 demonstrates the license utilization for the actual allocation vRad study assigner provided and the license utilization for the loading model allocation given our implementation experiment data. As the figure depicts, an allocation based on the loading model provides a license distribution is with less variance. With less variance on license distribution, vRad is less likely to delay jobs due to lack of state licenses. Moreover, there is a reduction in excessively underutilized licenses.

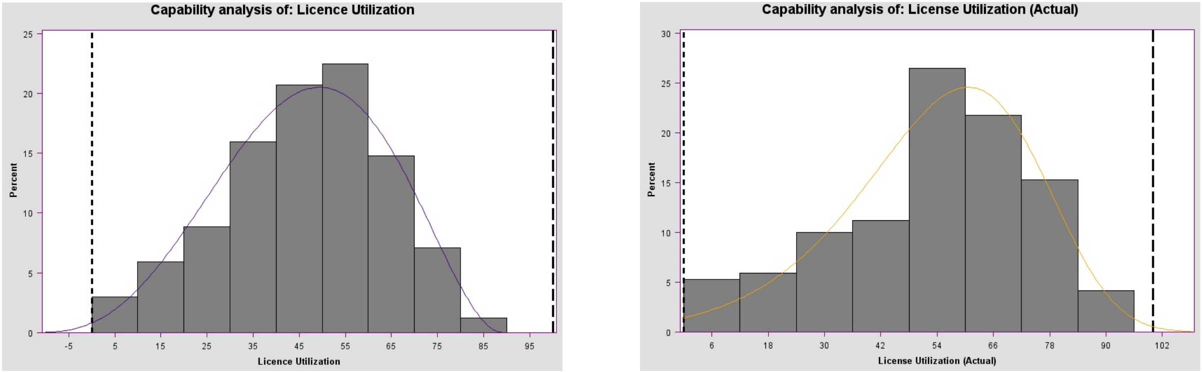

Figure 19 **License utilization histogram per radiologist for loading model output and actual allocation of vRad from historical data. The graph on the left hand side depicts the distribution of license utilization for the loading model. Loading model, as desired, creates a license distribution with less variance.**

Our experimental studies with the forecasting and loading models was quite promising. The forecasting model provided national level forecasts that can explain 97% of the variability of demand. Moreover, forecasting model offered state and facility level forecasts, which was not previously incorporated in vRad's legacy process. Figure 20 depicts at the state level, for half-hourly forecasts, the percentage improvements achieved via our forecasting over standard estimates used at vRad. For states with large volume of demand, our forecast fit the data especially well.

In our experiments, we compared the schedules our planning tools generated with the actual allocations made by vRad study assigner. As depicted in Figure 21, the loading model schedules provided a 25% improvement in average turnaround times and a promising reduction in the variance of turnaround times.

Although the comparison of our schedule with the actual allocations of vRad's study assigner is not an apple-to-apples comparison, and real time allocation has many additional challenges compared to a schedule, the 25% margin is indicative of the potential improvement our planning tools provide to vRad. To measure the robustness of the improvement our planning tools provide, we tested turnaround times resulting from our schedule leaving some radiologist capacity as buffer for additional complexity real time allocation may create. As Figure 21 demonstrates, the average TATs of our solution approach remained better than the TATs under the legacy process up to a 20% capacity buffer.

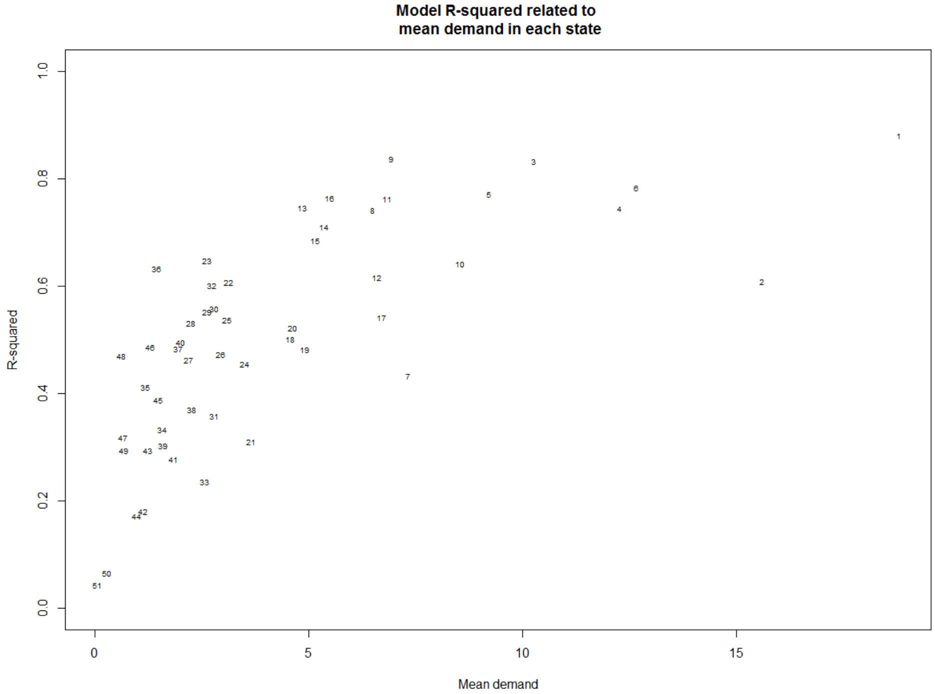

Figure 20 **% of improvements achieved via our forecasting over standard estimates used at vRad. Mean coefficient of determination values related to the mean demand at each state. For larger states, R-squared values are close to 1, indicating a good fit.**

Cultural Impact

This project has had some cultural impact at vRad, a company that values both human intuition and experience. We found the employees affiliated to be extremely experienced and good at their jobs. On the other hand, as a result of this project, we believe that vRad's executives and employees realized the potential of applying business analytics and investing in more sophisticated IT and new automated operational tools as catalysts for the effective use of intuition and experience. This improvement, in turn, can lead to better performance and more scalable operations. Because of our findings from this project, vRad started an ongoing collaboration with Carnegie Mellon University.

Acknowledgments

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

Appendix A: Details of the Forecasting Model

The data used to fit each forecast model comes 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 this time interval as the *training period*, because we use this data to train each model to forecast into August 2012 and beyond.

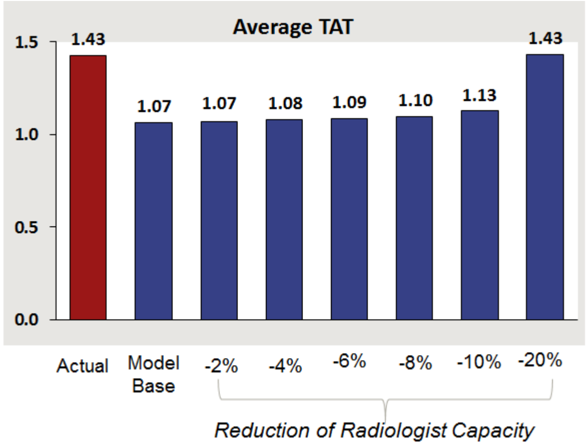

Figure 21 **Average TATs for the result of vRad study assigner and the schedules provided by our planning tools (The numbers are representative of the real ratios but are scrambled for confidentiality purposes). Additionally, the figure depicts the average TAT for different radiologist capacity reduction values. For instance, our planning tools generate 1.43 base units of turnaround time when the model is solved with 80% radiologist capacity. This average TAT value is equal to the realized TAT of the allocation provided by vRad study assigner. The remaining 20% capacity that is unassigned by the loading model may serve as a buffer to handle unexpected complications of real time allocation.**

A separate model can be fitted for any specific kind of read requests. 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 requests. Let $\mathcal{R}$ denote the set of all read requests that meet this desired criterion.

Let $i = 1, ..., n$ index the hours of the training period. For example, $i = 1$ corresponds to the hour from midnight to 1 a.m. on July 13, 2011, and $i = 2$ corresponds to the hour from 1 a.m. to 2 a.m., 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, than $|\mathcal{R}|/n$ is on the order of 600, as indicated by Figure 11.

Let y_i denote the number of read requests in the ith hour ($i = 1, ..., n$). Let h_i denote the hour of the day of the ith hour. Hence, $h_1 = 0$, $h_2 = 1$, ..., $h_{23} = 23$, $h_{24} = 0$, $h_{25} = 1$, 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

$$E(y_i) = b_0 + b_t t_i + \beta(w_i) + f_h(h_i) + I(w_i, h_i) + f_m(m_i). \tag{1}$$

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 of the regression splines 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 against 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. For a theoretical introduction involving the precise definition and estimation of the regression splines and underlying mathematical machinery, see Wood (2003), and for direct guidance on software and implementation, see Wood (2006).

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 $\mathscr{R} = \{1, ..., R\}$, a set of facilities $\mathscr{F} = \{1, ..., F\}$, a set of different priorities $\mathscr{P} = \{1, ..., P\}$ and a set of sub-specialties $\mathscr{S} = \{1, ..., S\}$. The time horizon (planning horizon) of the model consists of T periods. A job with sub-specialty $s \in \mathscr{S}$, priority $p \in \mathscr{P}$ coming from facility $f \in \mathscr{F}$ is referred as a job of type (f, s, p).

The parameters of the model are summarized in Table 1 and can be grouped into three sets. The first set of parameters determine whether a radiologist is available and able to read a job. In particular, the parameter $b_{fspt,r}$ takes value 1 if radiologist $r \in \mathscr{R}$ can read a job coming from facility $f \in \mathscr{F}$ that possesses sub-specialty $s \in \mathscr{S}$ and priority $p \in \mathscr{P}$ at time period $t \in \{1, ...T\}$. 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 sub-specialty s and priority p. Finally, r should be available at time t to be able to read a job. The second set of parameters pertain 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 handle the capacity of a radiologist r along with the maximum and minimum amount of work units that r can have.

Table 1 The parameters of the loading model ($f \in \mathscr{F}, s \in \mathscr{S}, p \in \mathscr{P}, r \in \mathscr{R}, t \in \{1, ...T\}$)

$b_{fspt,r}$	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.

There are two main decisions in the loading model. The first decision is the amount of work units of each type (f, s, p) that each radiologist r reads at time period t. This decision is denoted by $y_{fspt,r}$ which can take any non-negative real value. The second decision is the amount of work units backlogged from a period $t-1$ to the next period t and this decision is denoted by the variable $I_{fsp(t+1)}$. This decision variable is also non-negative real valued.

The formulation of the model is as follows:

$$\max \sum_{f,s,p,t,r} (T-t+1) w_p y_{fspt,r} \tag{2}$$

$$S.t. \quad \sum_{f,s,p} y_{fspt,r} \leq C^{\#}_{t,r} \text{ for all } t \in \{1,...,T\}, r \in \mathscr{R} \tag{3}$$

$$L^w_r \leq \sum_{f,s,p} y_{fspt,r} \leq C^w_r \text{ for all } r \in \mathscr{R} \tag{4}$$

$$\sum_{r \in \mathscr{R}} y_{fspt,r} + I_{fsp(t+1)} = d_{fspt} + I_{fspt} \text{ for all } f,s,p,t \tag{5}$$

$$y_{fspt,r} \leq b_{fspt,r} \sum_{\tau=1}^{t} d_{fsp\tau} \text{ for all } f,s,p,t,r \tag{6}$$

$$y_{fspt,r} \geq 0 \text{ for all } f,s,p,t,r \tag{7}$$

$$I_{fsp} \geq 0 \text{ for all } f,s,p \text{ and } h_r \text{ for all } r \in \mathscr{R}. \tag{8}$$

The constraints of the loading model are as follows: Constraint (3) sets the maximum amount of work units that a radiologist can read in a period. Constraint (4) 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 (5) is a backlogging balance constraint and it 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 $t+1$ (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 (6) implements the condition that a job can be read by a radiologist only if the radiologist is available, able to read the job. The demand term is used to strengthen the constraint using the fact that if there is no jobs to read, then the variable y should take the value 0.

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

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

and the loading model is solved again. At each step, all y and I variables that are 0 due to the absence of demand are eliminated. This way, the algorithm exploits the sparseness of the demand matrix d_{fspt}.

References

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

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

Batun, S., B. T. Denton, T. R. Huschka, A. J. Schaefer. 2011. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS journal on Computing* **23**(2) 220–237.

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

Best, T.J., B. Sandikci, D.D. Eisenstein. 2013. Managing hospital bed capacity through partitioning care into focused wings. *Working Paper* .

Chan, C. W., V. F. Farias, N. Bambos, G. J. Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research* **60**(6) 1323 – 1341.

Ernst, A.T., H. Jiang, M. Krishnamoorthy. 2004. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research* **153** 3–27.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: A tutorial and literature review. *Manufacturing Service Operations Management* **5**(2) 79–141.

Graves, S.C. 1981. A review of production scheduling. *Operations Research* **29**(4) 646–675.

Hastie, T.J., R.J. Tibshirani. 1990. *Generalized additive models*, vol. 43. Chapman & Hall/CRC.

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

Lee, D. K. K., S. A. Zenios. 2009. Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations Research* **57**(4) 852–865.

Leonard, M., B. Elsheimer, M. John, U. Sglavo. 2008. Small improvements causing substantial savings-forecasting intermittent demand data using SAS® forecast server. *SAS Institute Inc* .

Locumtenens.com. 2010. Locumtenens.com predicts possibility of eight to 16% shortage of radiologists by 2020. URL http://www.locumtenens.com/press-releases/2010-archive/locumtenenscom-predicts-possibility-of-eight-to-16-shortage-of-radiologists-by-2020.aspx.

Mahar, S., K.M. Bretthauer, P.A. Salzarulo. 2011. Locating specialized service capacity in a multi-hospital network. *European Journal of Operational Research* **212** 596–605.

Monegain, B. 2009. New research forecasts swelling telemedicine market. URL http://www.healthcareitnews.com/news/new-research-projects-swelling-telemedicine-market.

Nicodemus, A. 2010. Hospitals farm out radiology diagnostics. URL http://www.telegram.com/article/20100706/NEWS/7060382/1101.

Pinedo, M.L. 2008. *Scheduling Theory, Algorithms and Systems*. Springer.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Ripley, B.D. 2002. Time series in R 1.5.0. *R News* **2**(2) 2–7.

Snyder, R.D., J.K. Ord, A. Beaumont. 2012. Forecasting the intermittent demand for slow-moving inventories: A modeling approach. *International Journal of Forecasting* **28**(2) 485–496.

Swenson, M. D. 1992. Scarcity in the intensive care unit: Principles of justice for rationing icu beds. *The American Journal of Medicine* **92** 551–555.

Vanberkel, P.T., R.J. Boucherie, E.W. Hans, J.L. Hurink, N. Litvak. 2012. Efficiency evaluation for pooling resources in health care. *OR Spectrum* **34** 371–390.

Whitt, W. 1999. Using different response-time requirements to smooth time-varying demand for service. *Operations Research Letters* **24**(1) 1–10.

Wood, S. 2006. *Generalized additive models: an introduction with R*, vol. 66. Chapman & Hall/CRC.

Wood, S.N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1) 95–114.