No more waiting around

Lean Six Sigma accelerates response at medical call centers

By Nilav Ranpara and Dr. Neeraj Tayal

Each month about 20,000 people call five general internal medicine (GIM) clinic centers at The Ohio State University Wexner Medical Center. Approximately 25 percent of the callers wait for longer than five minutes, well below the ultimate goal of answering all calls within 30 seconds. Additionally, the 75th percentile of wait time was in excess of five minutes, contributing greatly to patient dissatisfaction and creating significant delays in patient care. To help with the strategic aim of improving patient access, GIM sponsored a lean Six Sigma project to address this problem.

Reducing call center wait time at five clinics

Wexner Medical Center is a multidisciplinary academic medical center on the main campus of The Ohio State University. In 2010, the center was ranked as one of "America's Best" in 11 specialties. General Internal Medicine (GIM) is a division within the Department of Internal Medicine.
Patients calling GIM experienced unacceptable wait times, sometimes as high as 21 minutes. This resulted in patient dissatisfaction and often delays in patient care. Standardized customer satisfaction surveys are playing a more important role in assessing health system performance, which is having more of an impact on reimbursement levels from third-party payers.

To fix this problem, GIM put together a core project team that included a senior from Ohio State's Integrated Systems Engineering Department, GIM's director and clinic managers. The project aimed to improve wait time by 15 percent and increase service level by 10 percent. The team decided to look at historical demand and identify choke periods – times in 15-minute intervals where demand on the system exceeded capacity. The team also sought to understand how workers spent their time by performing a work sampling study.

Queuing theory was used to determine staffing requirements during the peak periods at three of the five centers. The team determined the number by which each call center was short-staffed during peak periods. Based on the work sampling study during peak periods, the team found that workers spent about 30 percent of their time doing activities outside of answering phone calls, thus presenting an opportunity for 15 percent to 20 percent improvement.

The team implemented the staffing and the standardization of work during peak periods and decided to pilot it at two of the five clinics to measure impact. At the pilot site, wait time reduced by 50 percent and service level increased by 25 percent. Additionally, calls with wait times greater than five minutes declined from 16 percent to 2 percent, greatly enhancing patient satisfaction at the pilot clinic. In the current state, the project is going to be rolled out to other GIM clinics.

**Define**

This lean Six Sigma effort was a second generational project. The primary defect (Y) was the time that patients wait when they call general internal medicine. The wait time, defined as "response time delay" in the system, was a metric measured on a daily basis by clinic managers. Response time delays were unacceptable to management and had significant swings on a day-to-day basis. Defect for this project was defined as any customer who waited for more than one minute.

This project focused on three of the five GIM clinics – Martha Morehouse, CarePoint East and Stoneridge – the ones with the worst historical response time delays. Martha Morehouse had the worst delays, with the 75th percentile being as high as five minutes in October and November of 2012.
The core project team included the GIM division director, GIM director of clinical operations, clinic managers and call center employees. The initial kickoff meeting laid down the expectation that the project would improve response time delays by 15 percent and increase service levels by at least 10 percent. Additionally, the core team expected the project to determine optimal staffing levels to meet service level goals and devise a strategy to remove calls from the system to reduce queue lengths.

The voice of the business was loud and clear: Reduce response time delays so that patients can get their queries answered as soon as possible.

**Measure**
Thirty-five percent of the callers to general internal medicine waited longer than one minute.

**FIGURE 1**

**MEASURE**

<table>
<thead>
<tr>
<th>Day of the week</th>
<th>D</th>
<th>M</th>
<th>A</th>
<th>I</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>9:00</td>
<td>9:15</td>
<td>9:30</td>
<td>9:45</td>
<td>10:00</td>
</tr>
<tr>
<td></td>
<td>10:15</td>
<td>10:30</td>
<td>10:45</td>
<td>11:00</td>
<td>11:15</td>
</tr>
<tr>
<td></td>
<td>11:30</td>
<td>11:45</td>
<td>12:00</td>
<td>12:15</td>
<td>12:30</td>
</tr>
<tr>
<td>75th percentile of response time delay</td>
<td>was charted in 15-minute intervals from 8 a.m. to 5 p.m.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demand was categorized as peak (red regions), middle peak (yellow regions) and nonpeak (green regions) based on historical call volumes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Response time delay analysis**
- Choke period analysis revealed red regions where demand on the system exceeded capacity
- Monday all day and Tuesday-Friday mornings were times during which the system was constrained
- Demand gradually tapered off as the week progressed
- The periods of high demand matched with periods of high response time delays
- Work sampling study showed that choke periods had 15-to-20 percent scope for improvement in terms of answering calls

**Activity breakdown during peak times at CarePoint East**

- Outbound call 2%
- Telephone communication 25%
- Letters/other paperwork 5%
- Cater to misguided call 9%
- Answering phone 75%
Measure

The measure phase aimed to understand "choke periods" in 15-minute intervals where the demand on the system exceeded capacity. It included a work sampling study to measure how much time call center employees spent on various tasks.

One major challenge was procuring data from the information warehouse. The reports generated from the information warehouse used operational definitions inconsistently. Moreover, management used reports from a different database, OneSource, to track metrics. The project needed data on response time delays in 15-minute intervals.

The approach of this project is outlined using CarePoint East clinic as an example.

An X-Bar and R chart of response time delays at CarePoint East showed the process to be out of control, as illustrated in Figure 1. Information from the information warehouse showed that 35 percent of the callers waited more than one minute, with 8.5 percent waiting for more than 5 minutes.

The project conducted a demand analysis to understand variation on an hour-by-hour, day-by-day and month-by-month basis. A study of hour-by-hour variation in 15-minute intervals predictably showed that the highest demand happened between 9:15 a.m. and noon. Demand gradually dropped off as the day progressed, with few calls after 3:15 p.m. A study of day-by-day variation also showed Mondays to be the busiest day, with demand gradually decreasing as the week progressed.

This is typical of a healthcare setting, where Mondays are the busiest day. A one-way analysis of variance on demand showed that Monday was significantly different from Tuesday and Wednesday, which differed quite a bit from Thursday and Friday.

The project team needed to assess the severity of the problem at a weekly level across five days a week and eight hours each day. The team created a matrix for response time delays by 15-minute intervals to identify periods where response time delays went above the specified goal. This matrix spanned each day of the week from 8 a.m. to 4:45 p.m.
This matrix is akin to a heat map. Regions that had response time delays above the 75th percentile spec were categorized as "pain point" regions. Regions that had response time delays between 30 seconds and one minute were categorized as "manageable." Response time delays that were less than 30 seconds were categorized as acceptable. These distinctions came about after discussions with the project's core team.

The results of this matrix showed that Mondays had the maximum number of pain points by 15-minute intervals, an outcome consistent with demand analysis over the same period. Tuesday and Wednesday mornings had a mixture of pain points and manageable response times, but the delays were long enough that management deemed them unacceptable.

After this choke period analysis, a work sampling study was conducted at all three clinics to understand how staff members spend time during the peak periods of demand and choke periods of response times. The power of sample used in the study was 85 percent, thus the data collected was representative of the population.

The sampling study revealed that during peak periods, call center employees spend only 70 percent of their time on the phone. Other activities included writing up nonurgent telephone encounters, doing paperwork-related activities, catering to misguided calls and making outbound calls. Standardizing the work flow during these peak periods offered an opportunity to spend 15 percent to 20 percent more time on the phone. The presence of misguided calls showed an opportunity to reduce queue lengths, a deliverable agreed upon in the define phase of the project.

In addition, calls from new patients took 20 to 30 minutes because staffers had to collect myriad demographic and insurance information. Therefore, in the past year management created a centralized call center just for new patient appointments. The phone trees were changed for all clinics to direct new patient calls to this centralized resource, which was staffed full time with two to three call center employees.
The focus on this phase was to determine optimal staffing levels and standardize work flow during peak hours.
Queuing theory was used in this phase to understand the staffing scenario. The M/M/c variant of the queuing theory model, where the underlying assumption was that inter-arrivals were exponentially distributed and the calls had an exponential service time distribution, was used.

After checking for all assumptions, queuing theory formulas were applied to determine staffing levels.

The service level chosen to calculate the staffing requirements was 75/60 – that is, 75 percent of callers wait less than 60 seconds. Call volumes, talk time and service levels were the inputs.

The output was the average response time delay, the 75th percentile of response time delay and the number of servers required to meet the service level goal.

A matrix was created to show staffing requirements in 15-minute intervals for all days of the week. Time periods during which CarePoint East was understaffed versus overstaffed were determined and are shown in Figure 2. Because the calculation was made using a theoretical approach, the results were checked with the clinic manager, who validated them as consistent with the day-to-day experience of being understaffed.

The team used the staffing model to build a user-friendly calculator that would help managers determine optimal staffing solutions on a quarterly basis. The calculator will use an Excel database to refresh data and feed it in the calculator to perform all queuing theory calculations and give optimal staffing level outputs.

An engineered forecast predicted that following the staffing strategy swung the response time metric favorably, sometimes as much as 15 percent during peak periods.

Standard work considerations led management to require that activities other than answering calls happen during nonpeak periods. The "nonurgent" activities were defined later, in the improve phase. And to increase the likelihood of reaching their target, employees decided to make outbound calls to patients between 5 p.m. to 7 p.m.

In addition, a parallel 5S project consolidated all needed information for the call center staff onto a Web-based platform and gave call center workers an additional computer monitor. When managers update this system with information like physician codes, common phone numbers, resources and vacation schedules, the data are immediately available to all call center staff within a few clicks on their second monitor.
Improve

CarePoint East was chosen to pilot the staffing and standard work solution elements during peak hours. GIM managers hired temporary associates during the pilot phase to evaluate the solutions' effect on response time delays.

The solution elements encompassed three unique strategies. The first was to staff the call centers based on the queuing model output mentioned in the analyze phase. This was expected to have the biggest impact. The second solution element entailed deferring nonurgent work to nonpeak hours, categorizing work as high, medium and low priority.

High priority or urgent work entailed any patient request that put health at risk, like shortness of breath, fever, suicidal thoughts, etc. All other calls, such as patient requests for paperwork involving faxes and scans, were considered nonurgent work that could either be fulfilled by the end of the day (medium priority) or within 48 hours (low priority). The lead physician decided what was medium and low priority, and these activities were supposed to be performed when call volume was low. The same rules of urgent versus nonurgent were applied to outbound calls.

The last solution was increasing the awareness of MyChart, a Web-based portal that lets patients schedule appointments with doctors, make requests for refilling prescriptions or ask questions based on their requirement. It is particularly useful for nonpressing matters. A simple regression analysis showed that increasing the adoption rate of MyChart by 25 percent would reduce calls per patient by 50 percent on an annual basis, a tremendous potential for reducing call volume. To increase awareness, physicians promoted MyChart to their patients, and information about the system was incorporated into the phone tree.

These solutions were piloted for two months. Based on an initial evaluation, peak times at CarePoint East were adjusted for on Mondays, Tuesdays and Wednesdays to reflect choke periods better. It also was discovered that wait times increased during the lunch hour (11:30 a.m. to 12:30 p.m.). Therefore, management staggered lunch breaks to ensure that at least three people staffed the call center at all times.

Figure 3 details how the primary metric, response time delay, had a 51 percent improvement. The 75th percentile of response time delay improved from 110 seconds to 52 seconds. The service level, or percentage of calls answered within 30 seconds, increased from an average of 50 percent to an average of 80 percent, a historical high. Average response time delay hit a historic low of 21 seconds.
The pilot's success led upper management to implement the solutions at the other four GIM clinics.
Control

Any good lean Six Sigma projects make provisions to sustain the improvements. The team put into place three key control elements to ensure continuity: a sustainable measurement system that created transparency for the performance of the primary defect of call waiting time; a staffing calculator that enabled management to refresh staffing on a quarterly basis; and a project transition action plan that highlights key items that GIM management should continue after the project finishes.

After validating the shift in process capability through hypothesis testing at CarePoint East, GIM management was ready for a systemwide rollout not only to the GIM clinics but to other clinics under the Wexner Medical Center umbrella.

After an evaluation period of 16 months at four other sites, a two-sample t-test concluded that response time delay significantly improved after the solutions were implemented, with a confidence level of 85 percent. Figure 4 details those improvement metrics.
The staffing calculator was the key for the systemwide rollout of the broader solution. It enabled clinic managers and GIM management to assess their staffing needs periodically. A survey revealed that managers "didn't rely on the staffing calculator as a one-time fix. Rather, they used it as a regular and fluid assessment of their staffing needs. They depended on it to guide their hiring/staffing and business decisions."

In addition to the staffing calculator, GIM management uses an executive dashboard to gauge call center performance on a weekly basis. This dashboard leverages response time delays, service level and a new metric – percentage of calls with a response time of more than five minutes. Managers have been given easy access to the dashboards and the staffing calculator to assess their performance continually.

Medical assistants said that their experiences when phone calls are at their peak are "considerably less stressful" and that they can "address the needs of their patients in a much more timely and organized fashion." In addition, MyChart adoption increased by 15 percent across the system, representing more avenues for patients to have access to care.

Ensuring this project's sustainability is critical to maintaining high patient satisfaction, which leads to enhanced access to care for Wexner Medical Center patients and increased referrals.

Nilav Ranpara is a graduate of the Department of Integrated Systems Engineering at The Ohio State University. The certified green belt is a business technology analyst with Deloitte Consulting.

Dr. Neeraj Tayal is a general internal physician at the Ohio State University Wexner Medical Center and has served as the director for the division of general internal medicine and geriatrics since 2001.