Proceedings of the International Conference on Industrial Engineering and Operations Management
Paris, France, July 26-27, 2018

Improve an Emergency Medical System Capacity with Theory of Constraints

Bernardo Villarreal, Jose Arturo Garza Reyes*, Jenny Diaz Ramirez, Aracely Carranza, Gabriela Morales & Arturo Quezada
Universidad de Monterrey
San Pedro Garza García, N.L., México 66238
bernardo.villarreal@udem.edu Jenny.diaz@ude.edu arturo.quezada@udem.edu,
gabriela.morales@udem.edu, aracely.carranza@udem.edu

***Derby Business School**
The University of Derby
Derby, DE22 1GB, UK
j.reyes@derby.ac.uk

Abstract

The measure of ambulance cycle time is of particular interest in the performance of Emergency Medical Systems (EMS). Its value determines the ambulance capacity of the operations system of an EMS institution. For a given number of ambulances in operation, as the value of the average ambulance cycle time increases, total ambulance capacity for the system is reduced. Additional impacts resulting from this situation are; an increase on the average ambulance response time and patient´s health risk due to the unavailability of ambulances; and a greater need of ambulance replacement investment requirements and operating cost. This work suggests an improvement approach based on Theory of Constraints (TOC) and elimination of waste for reducing ambulance cycle time. The approach is applied to a Mexican EMS institution based on metropolitan Monterrey. Results of the application are provided.

Keywords
Ambulance cycle time; bottleneck, waste reduction; emergency operations

1. Introduction

The fundamental responsibilities of Emergency Medical Service (EMS) systems are to provide urgent medical care, such as pre-hospital care, and to transport the patient to the hospital if needed. The efficiency of EMS systems is a major public concern.

According to Fitch, et al., (2015), EMS systems are to provide urgent medical care, such as pre-hospital care, and to transport the patient to the hospital if needed. The activities involved are:

- Receive emergency call and an ambulance is assigned.
- Ambulance preparation.
- Transporting the ambulance to the emergency scene.
- Serving the injured or sick person until he is stabilized.
- Transfer the customer to a health institution.
- Delivering the customer to the health institution.
- Transportation back to ambulance base.

The activities previously described are part of the ambulance cycle. According to Blackwell, et al., (2009), the provision of optimal emergency medical services care in the pre-hospital environment requires a high level of coordination and integration of multiple operational and clinical resources utilized by many people located at different places. Activities such as call taking and dispatching, scene response, on-scene patient care, triage and hospital destination decisions, continuing care during transport, and transfer to definitive care are all subject to online and off-line medical direction and guidance. The level of performance of this process is determined by the adequate management of all these elements.

Two of the most important performance indicators for EMS institutions; the agility required to execute the process; and its efficiency (operations cost). The level of agility is measured by various time indicators; Paramedic response time; ambulance

turnaround time; ambulance cycle time; patient stabilizing time at scene among others. Paramedic response time to the scene of a call for emergency medical assistance has become a benchmark measure of the quality of the service provided by EMS operations (Pons et al., 2005). As suggested by Pons et al., (2005) a target response time of ≤ 8 minutes for at least 90% of emergent responses has evolved into a guideline that has been incorporated into operating agreements for many EMS providers. The International Guidelines 2000 Conference on Cardiopulmonary Resucitation and Emergency Cardiovascular Care recommended a response time of 8 to 10 minutes to insure a successful cardiac and cerebral resuscitation (Blackwell, et al., 2009). The time taken by paramedics to stabilize the patient at the scene is critical for achieving a service of quality. A standard established for a high quality service is a time less than or equal to ten minutes. This has been called as the Platinum Ten (Watson 2001). Other important performance indicator is related to the first 60 minutes after traumatic injury. This has been called the golden hour (Rogers et al., 204; Newgard et al., 2010).This concept states that definitive trauma care must be initiated within this 60-minute time window. The belief is that injury outcomes improve with a reduction in time to definitive care, and it is a basic premise of trauma systems and emergency medical services (EMS) systems. Ambulance Turnaround time is defined as the time taken by the ambulance starting from its arrival to a hospital until it is ready again and available to respond to other emergency call. Finally, the ambulance cycle time represents the total time taken by the ambulance from responding to an emergency call until it becomes available to respond to a new emergency call. All the previous time performance indicators, with the exception of cycle time, are important at the operational level. Ambulance cycle time could be related at the operational and the strategic level. This indicator is very well related to the ambulance installed capacity of the process and to the unit hour utilization indicator. Lower cycle times imply higher unit hour utilization and ambulance capacity. Furthermore, the level of these indicators impacts the level of the organization´s efficiency (operating cost).

Over the past three decades, a significant amount of research studies have been conducted to improve the performance of EMS systems. The major focus of these models is to reduce response time (the time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene) by placing the ambulances in optimal locations. The focus has been on response time because EMS systems are designed to rapidly provide advance medical care to critical patients such as cardiac arrest or trauma. However, the measure of ambulance cycle time is of particular interest in this work. Its value determines the ambulance capacity of the operations system of an EMS institution. For a given number of ambulances in operation, as the value of the average ambulance cycle time increases, total ambulance capacity for the system is reduced. Additional impacts resulting from this situation are; an increase on the average ambulance response time and patient´s health risk due to the unavailability of ambulances; and a greater need of ambulance replacement investment requirements and operating cost.

This work presents a scheme based on Theory of Constraints (Goldratt 2012) for improving the value of the average Cycle Time. Other works that have used similar procedures based on TOC are Villarreal et al., (2016), Siller et al., (2013) and Villarreal et al., (2014). The suggested approach is applied to the Mexican Red Cross operations of the Monterrey metropolitan area. The organization felt that the current level of its ambulance cycle time could be reduced to relieve the pressure of its high level of operating cost.

This paper consists of five sections. Section one offers an introduction to the problem and context around it. Section two provides a brief review of the literature on TOC. Section three gives a description of the recommended scheme. The application of this scheme is undertaken in section four and section five where conclusions and future recommendations are presented.

2. Review of Relevant Literature

Waste elimination is a fundamental aspect in Lean literature (Schonberger 1982; Ohno 1988). A process can be separated into value adding and non-value adding steps, also called waste, according to market's needs. Toyota was the first to contribute in the waste identification process. Toyota defined seven major types of waste in manufacturing and business processes (Ohno 1988). These include overproduction, waiting, unnecessary transport, incorrect processing, excess inventory, unnecessary movement and defects. It has been shown by Keyte et al., (2004) that a great deal of waste has yet to be identified and eliminated in the administrative processes that support shop floor operations. In order to facilitate it, they adapt the seven wastes previously described for manufacturing operations to administrative processes, adding a new waste of underutilized people. As the focus of the value stream includes the complete value adding (and non-value adding) process, from conception of customer requirements to the consumer's receipt of product, there is a clear need to extend this internal waste removal to the complete supply chain. The seven wastes previously mentioned required an adaptation to the supply chain environment. A process mapping tool called Value Stream Map (VSM) was developed by Womack et al., (1996) for the extended

enterprise, looking to identify waste between facilities and installations in a supply chain. Mapping at the supply chain level, unnecessary inventories and transportation become important wastes to identify and eliminate. At this level, transportation waste is related to location decisions that seek to optimize performance at individual points of the supply chain. Therefore, the solutions suggested for its elimination are concerned with the relocation and consolidation of facilities, a change of transportation mode or the implementation of milk runs.

Lean healthcare seems to be an effective way of improving healthcare organizations and the growing number of implementations and reports found in the literature reinforce this view. An initial consideration about the use of lean concepts is offered by Heinbuch (1995), in the particular case of just-in-time. Some later publications that describe further applications are Chalice (2005), Chalice (2007), Zidel (2006) and Brandao (2009).

As it was the case for the manufacturing and administrative processes, it is necessary to define what is considered waste or value. Bentley et al., (2008) suggests a framework of three types of waste—administrative, operational, and clinical. Both administrative and operational wastes are components of inefficient production, and clinical waste is a form of allocative waste. Administrative waste is the excess administrative overhead that stems primarily from the complexity of the U.S. insurance and provider payment systems, and operational waste refers to other aspects of inefficient production processes. Clinical waste is waste created by the production of low-value outputs. A proposed extension of the seven Toyota manufacturing wastes to the healthcare area is given by Graban (2016) and it is illustrated in Table 1. This waste classification is more suitable for process improvement inside the health institution. They are best associated to hospital operations.

Table 1 Description of the Eight Healthcare Extended Wastes

Type of Waste	*Brief Description*	*Hospital Examples*
Defects	Time spent doing something incorrectly, inspecting for errors, or fixing errors	Surgical case cart missing an item; wrong medicine or wrong dose administered to patient
Overproduction	Doing more than what is needed by the customer or doing it sooner than needed	Doing unnecessary diagnostic procedures
Transportation	Unnecessary movement of the "product" (patients, specimens, materials) in a system	Poor layout, such as the catheter lab being located a long distance from the ED
Waiting	Waiting for the next event to occur or next work activity	Employees waiting because workloads are not level; patients waiting for an appointment
Inventory	Excess inventory cost through financial costs, storage and movement costs, spoilage, wastage	Expired supplies that must be disposed of, such as out-of-date medications
Motion	Unnecessary movement by employees in the system	Lab employees walking miles per day due to poor layout
Overprocessing	Doing work that is not valued by the customer or caused by definitions of quality that are not aligned with patient needs	Time/date stamps put onto forms, but the data are never used
Human potential	Waste and loss due to not engaging employees, listening to their ideas, or supporting their careers	Employees get burned out and quit giving suggestions for improvement

It is important to point out that the EMS process described earlier can be considered as the basic transportation process described in Villarreal (2012). Furthermore, According to Simmons et al., (2004), improving transport operations performance can also be achieved increasing its efficiency through waste elimination. Transport efficiency was originally suggested by Simmons et al., (2004). They made the measurement with the Overall Vehicle Effectiveness (OVE). Similar to the estimation of OEE, it is required to calculate the availability, performance and quality efficiency factors. The product of the three efficiency factors would yield an overall OVE percentage rate. This measure converted the OEE losses from manufacturing to transport operations. The result was the definition of five transport losses or wastes. These are driver breaks, excess load time, fill loss, speed loss and quality delays. The previous measure has also been modified by Villarreal (2012). In this case, the OVE measure is adapted to consider total calendar time as suggested by Simmons et al., (2004).

Figure 1 illustrates the concepts and losses involved in the proposed measure that is called Total Operational Vehicle Effectiveness and represented by the term TOVE. In summary, four components for the new efficiency measure are suggested; Administrative or strategic availability, operating availability, performance and quality. The new measure would be obtained from the product of administrative availability, operating availability, performance and quality efficiency factors.

The TOVE index and related wastes is adapted to the EMS operations in this work. The new index is called the Ambulance TOVE and will be represented by A-TOVE hereafter (Villarreal et al., 2017). The associated wastes are illustrated in Figure 1. The wastes related to the Administrative Availability efficiency factor are similar to those of the Transportation Value Stream Map (TVSM) described in Villarreal (2012). The wastes considered in the operating availability efficiency factor are ambulance waiting to be assisted by a resource, ambulance time taken in excess to execute operation procedures, and corrective maintenance.

These could happen before the ambulance departs to the point where it is required according to the call, during transportation of patients, triage and delivery of the patient to the hospital. The wastes considered in the performance efficiency are; speed loss; fill loss; and distance traveled in excess. Distance traveled in excess is a result of a deficient ambulance transport planning; wrong ambulance site definition; deficient route planning; and inadequate ambulance assignment and dispatching policies. Fill loss is related to the number of injured or sick persons that do not use the ambulance to transport them to the medical institution, either because it is not required or they use other means of transportation. Finally, speed loss is an important waste because it determines the time at which the ambulance will reach the emergency scene or the hospital at which the patient will be delivered.

Quality efficiency wastes are related to the percentage of times that international standard times are not met putting the patient´s health in risk; time in excess of response time; golden hour and platinum ten are considered. The Value Stream Map (VSM) utilized in this work is a modified version of the one provided by Villarreal (2012) and will be denoted as the Ambulance VSM (A-VSM) hereafter. This A-VSM is obtained from following the Ambulance.

3. Description of the Waste Reduction Scheme

As reference Goldratt (2012) suggest, the productivity of the manufacturing system is determined by a bottleneck or the most constrained capacity resource. According to Huang et al., (2003) this type of resource is the one with the highest Operational Equipment Effectiveness (OEE) value. Thus, the procedure they develop is fit into the five-step improvement cycle designed by Goldratt (2012). The first two phases are designed to estimate the OEE values for each production resource. The third phase consists of the identification of the bottleneck or most constrained resource. This is done by identifying the resource with the highest OEE value. Once this is carried out, the identification of wastes or losses is the next task. These are associated with the availability, performance and quality efficiency factors of the bottleneck. The final phase of the procedure includes the definition of projects or actions with the purpose of eliminating the wastes found. This is done until the constraint is broken. The process is repeated if a new constraint is identified and it is desirable to continue improving the productivity of the manufacturing system. The previous procedure is constructed such that it is required to have all the OEE values for each capacity resource of the operations system. This is so because the identification of the bottleneck is done with this information. Other works that have used similar procedures based on TOC are Villarreal et al., (2016), Siller et al., (2013) and Villarreal et al., (2014).

The procedure suggested in this paper is very similar. However, it does not require all the OEE values for each capacity resource. The determination of the bottleneck or capacity constrained resource is carried out through a load analysis and the elaboration of a value stream map. In addition, the waste classification recommended by Graban (2016) is used for the phase of waste identification and elimination. This alternate procedure is described as follows:

(1) Elaborate the Ambulance Value Stream Map and a capacity load analysis for the process of interest.
(2) Identify the bottleneck or more restrictive resource. Identify and prioritize important wastes.
(3) Exploit and elevate the bottleneck by implementing waste elimination initiatives.
(4) The previous step continues until a new bottleneck is found or management decides to stop. If a new bottleneck is found, continue to step 2. Otherwise, the process ends.

4. Implementation and Results

This work proposes a scheme to improve the level of the Ambulance cycle time of the Mexican Red Cross´ operations in the Monterrey metropolitan area. The operations count with seven (7) fixed locations and three (3) mobile locations from which ambulances are sent to service pre-hospital events. The organization has 34 ambulances but the financial resources to operate

50% of them during any day. The number of emergency calls considered in the analysis carried out in this work total 30,600 and occurred during the period of November of 2016 and June of 2017. Three types of emergency calls accounted for 93% of all calls; those related to people with a sickness total 40%; vehicular accident related calls account for 31%; calls of other type of accidents are 22%; and finally 7% of the calls are due to various other causes. The organization felt that the current level of Turnaround time needed to be reduced significantly to satisfy international standards and satisfy the new upsurge of 16% in demand due to the project of centralization of emergency calls implemented by the government of the state of Nuevo Leon. This new situation put a great pressure on the level of operations cost and the requirements for more ambulance investment.

Figure 1 illustrates the behaviour of daily call demand per hour and day of the week. Two characteristics are to point out; every day has the same pattern, with the exception of Sunday at night and; two demand levels are present every day, a high level of emergency calls of about six calls per hour occurring from 7 to 22 hrs and a low level of demand of three calls per hour during the night shift.

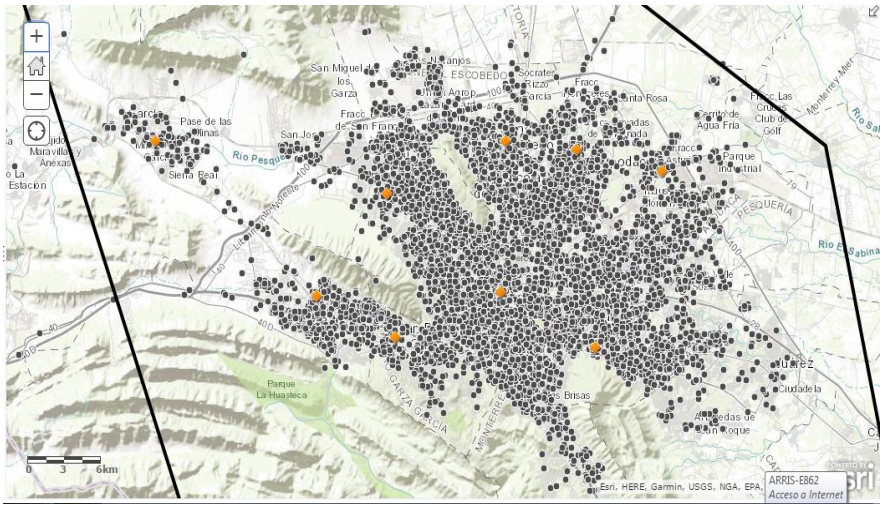

Figure 1 Geographical location of the sample of emergency calls

4.1. Mapping the ambulance cycle process

The first step of the methodology is the mapping of the operations. In this case, an A - VSM for the cycle process of interest is elaborated. Figure 2 presents the Ambulance VSM of the Monterrey metro area operations. According to the A - VSM, the average ambulance response time is estimated in 24.7 minutes, Also, the average time recorded to stabilize the patient´s health is 34% over the Platinum Ten. Finally, taken into account the time from the emergency call until the time at which the patient is delivered into a health institution, the average estimated time is 84% above the golden hour. Turnaround time is estimated in 49.1 minutes and total Ambulance cycle time averaged 124.9 minutes with a standard deviation of 117.8 minutes. The value of ambulance cycle time implies a patient throughput per ambulance of 0.48 per hour. Ambulance turnaround time is 39.3% of the cycle time and it is the greatest element of it. The second greatest component of the cycle time is the ambulance response time which accounts for 19.8%. The bottleneck of the process would be the one related to the turnaround time.

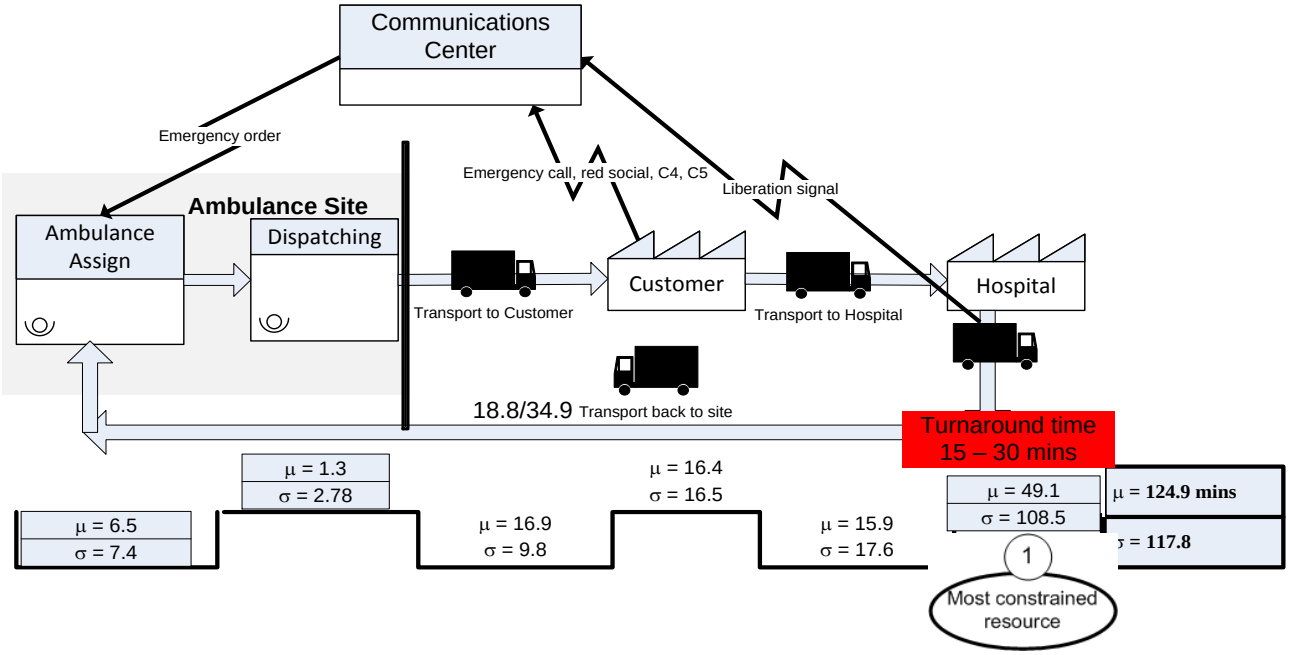

Figure 2 Description of the A-VSM for the Monterrey Metro operations

4.2 Breaking the first bottleneck

The main bottleneck of the ambulance cycle process corresponds to the activities involved with the patient handover to the corresponding health institution. The time required for executing them is called turnaround time. Five institutions account for about 77% of the emergency arrivals of the Cruz Roja ambulances. These are Hospital de Zona 21 (HZ) with 26%, Hospital Metropolitano (HM) with 23%, Hospital Universitario (HU) with 18%, Clínica 6 del IMSS (IMSS 6) with 6% and Clinica 17 del IMSS (IMSS 17) with 4%.

The previous demand pattern of emergency calls was serviced by the fleet of ambulances available. However, not all of them required the transportation of the patient to a health institution. On average, the hourly level of calls that needed transferring the patient to a health institution decreased significantly. The daily behavior of average arrivals to the health institutions is presented in Figure 3. As shown, three health institutions receive most of these services; Hospital de Zona (HZ), Hospital Universitario (HU) and Hospital Metropolitano (HM). Two general patterns are identified; a high level of emergency arrivals occurring from 7 to 22 hrs and; a low level load pattern during the rest of the day (at night).

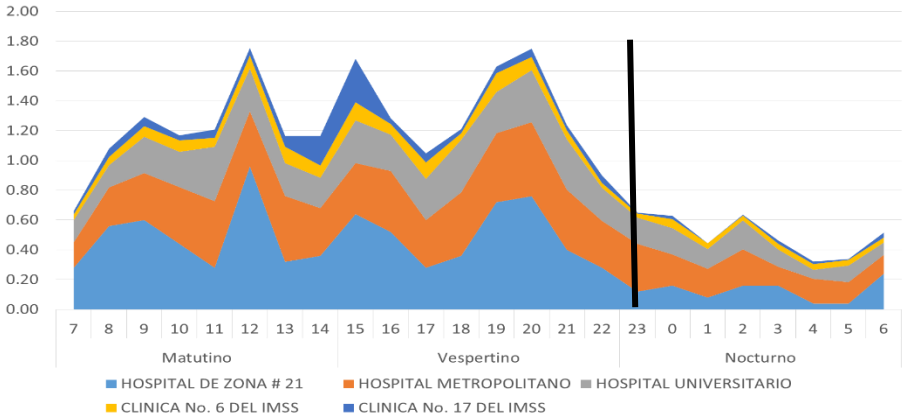

Figure 3 Behavior of the emergency arrivals to each health institution

In order to find out the most important causes for long turnaround times, a sample of 850 observations were taken during the period of August 15 to September 10 of 2017. These were gathered in the first five health institutions mentioned previously; HU, HM, HZ, IMSS 6 and IMSS 17. As shown in Figure 4, total average turnaround time is estimated in 59.36 minutes. The highest element of the turnaround corresponds to stretch liberation with 55.9%. This value is significantly higher than the recommended standard range of 15 – 30 minutes. HU and IMSS 17 have the highest turnaround times with 90.4 and 83.5 minutes respectively.

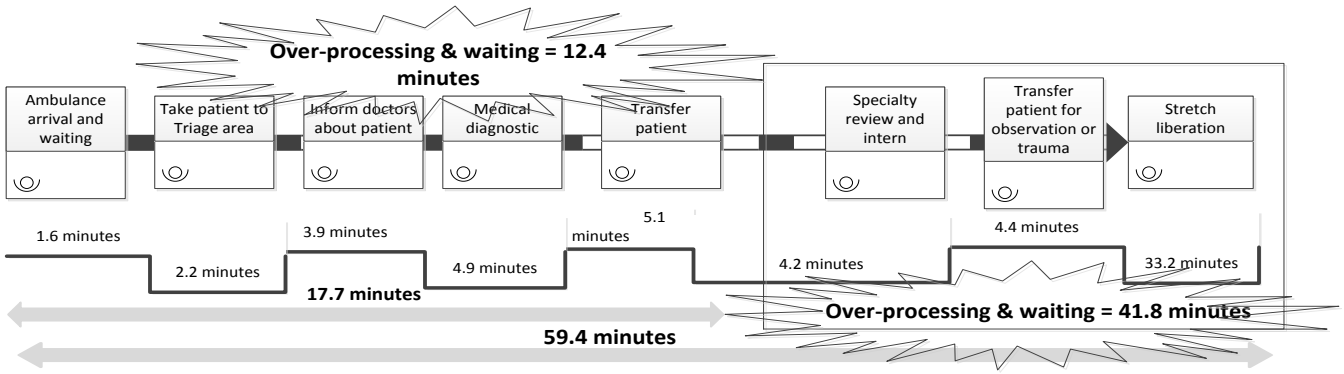

Figure 4 VSM for the EMS operations of the Red Cross

Comparing the patient handover process for the Red Cross with the ideal process suggested by NHS (2012) , one can realize that it differs significantly in its activities. Ideally, the process should terminate after the triage with the transfer of the patient to the health institution responsibility. However, Red Cross paramedics rarely finish their participation at the ideal suggested stage. Due to the lack of stretches available at each health institution, they are forced to continue "their participation" in the process until there is a way to release the patient safely. Hence, what should be an average of 17.7 minutes of participation in the process, ends up being an average of 59.4 minutes of collaboration.

Description of areas of opportunity

The main wastes are due to waiting for doctors and stretch mainly, excess of movement and over-processing. Figure 4 illustrates the location of occurrence of these wastes. In summary, paramedics take an average of 41.8 minutes working for the health institution performing tasks that are not their duty. In addition, it was observed that the rest of the activities contain a 70 % of waste time due to over-processing information, waiting for doctors, nurses and medical students, and excess movements of the patients.

Defining and implementing the improvement strategy

The strategy for improving performance consists of three initiatives: incorporate an ambulance decoupling team; define internet based tools to communicate patient information during his transfer to the hospital and the status of stretches in the hospital and; negotiate with the hospitals the Red Cross responsibility in the patient handover process. It is important to point out that if the last initiative is implemented, it will automatically imply the liberation of the ambulance immediately after the patient triage.

The initiative of the ambulance decoupling team consists of putting together a crew formed by a paramedic and an assistant per shift and hospital. This team will be responsible of receiving the patient from the ambulance crew and proceed with the rest of the activities. The potential decoupling points can be located right after the arrival of the ambulance to the hospital or after the execution of the patient triage. The impact of implementing this initiative may be very important if it is implemented. Total average turnaround time could be reduced to a range of 2 to 18 minutes approximately.

The determination of the crew size and the stretch inventory level at each hospital was facilitated by the use of simulation. Simulation models for the handover processes in the HU and HM were developed using the ProModel software suite considering a decoupling point located after the patient triage. The database of emergency calls taken during the period of August 15 to September 10 of 2017 was used to obtain the probability distribution functions for each activity of the handover processes with the StatFit tool provided by ProModel.

The statistical robustness of the models was enhanced with the determination of the warming and total simulation periods of the models and the definition of the number of replicates. According to the simulation, the minimum turnaround time was

achieved with a crew size of two paramedics and a stock of two stretches in both hospitals. The average turnaround time obtained from the simulation models was 12.9 minutes for the HM and 17.3 minutes for the HU.

The initiative of concern was tested in a two-stage pilot project at HU. The initial stage will consider the location of the decoupling point immediately after executing the patient triage. Then the team will be responsible to continue performing the rest of the activities including the stretch recuperation. The second stage of the project would be implemented simultaneously with an internet based tool that will enable to transfer patient information from the ambulance to the team while transporting the patient. This information will make it possible for the team to receive the patient immediately after arriving to the hospital. The team will then be in charge of taking the patient through the triage and the rest of the activities of the handover process.

The implementation of the pilot project, in its first stage, of the ambulance decoupling point team was undertaken starting October 15 and at the HU. Thus, the results described hereafter are based on information gathered during about 15 days. The total number of observations is 108. The real average turnaround time for HU, during the pilot project period, is estimated in 19 minutes, a 79% reduction from the original.

Negotiation with hospitals of responsibility in handover process
This initiative considers the definition of the responsibility boundaries of the Red Cross, and the associated hospital, with respect to their participation in the patient handover process. Ideally, the collaboration of the Red Cross must be until the patient passes the triage activity, leaving the patient under the hospital´s responsibility. If this is carried out, the ambulance turnaround time would result in a significant reduction.

The implementation of this initiative was undertaken for HU and HM with different results up to date. The negotiation with HM was successfully completed and the one with HU was dependent on the results obtained from the implementation of the ambulance liberation team. The project with HM started on October 17. Therefore, the results so far of about 13 days with about 67 observations are important. Ambulance average turnaround time is estimated in 23.6 minutes during the pilot project period. This new level represents a 54% decrease versus the original value.

Implementing internet based tools
The internet based tools considered are destined to improve the communication between the ambulance and the hospital during its transport to the institution and to increase the level of control of stretches inside the hospitals. The impact of the first tool is expected to be at the triage activity reducing the time taken in about four minutes. The tools identified to monitor the stretches status would improve their recovery on time. The implementation of this initiative was actually postponed for the first semester of year 2018 once the operating budget for the institution is approved.

The following steps of the implementation of the first two initiatives in the following two months, will be to consider; the implementation of the ambulance decoupling team after the arrival of the ambulance at each hospital HU and HM and; to finish the negotiations to reconsider the responsibility of the Red Cross in the patient handover process in the HU.

4.3 Breaking the next bottleneck
The following bottleneck identified corresponds to the process involved in the ambulance response time (see Figure 5). This is estimated in 24.7 minutes on average and a standard deviation of 12.5 minutes. About 31.5% of this time is associated with the execution of the activities required to prepare the trip of the ambulance to the place where the patient is located. At this point, the new average ambulance cycle time is projected to 75.8 minutes based on the results of the pilot tests. The new estimated average turnaround time is 3.9 minutes.

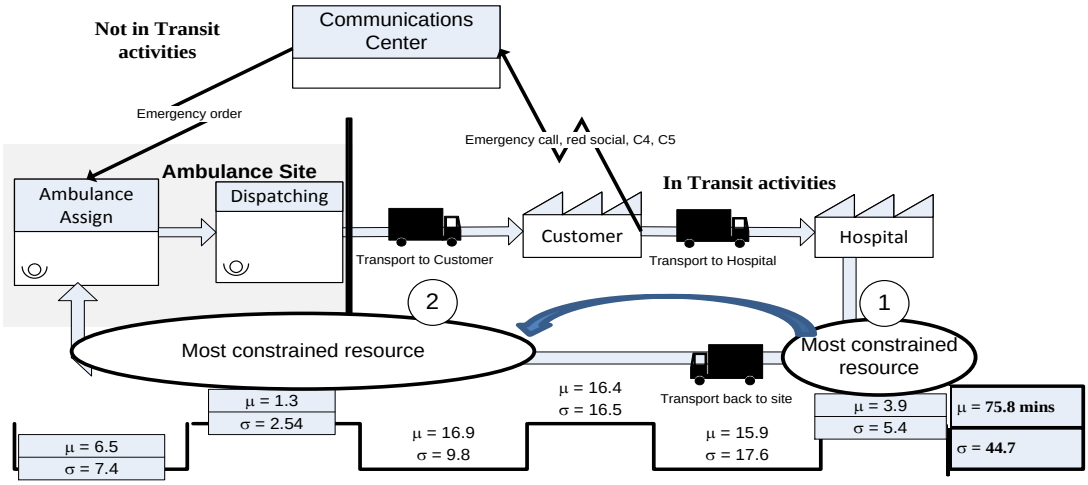

Figure 5 Illustration of the new bottleneck resource

The daily behavior of emergency calls received by the Red Cross is similar to the daily behavior of emergency arrivals to the health institutions as illustrated in Figure 6. Again, two patterns are recognized; A low emergency call level per hour at night from 22 to 7 hrs and; a high emergency call level per hour during the rest of the day.

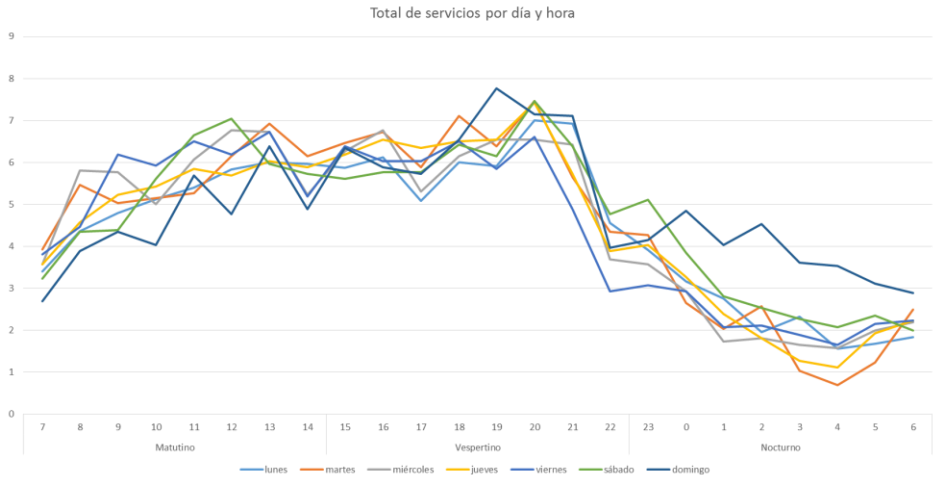

Figure 6 Daily behavior of emergency calls to the Cruz Roja

Figure 7 illustrates the (kernel) density map of the emergency calls. The areas with highest call density correspond to Monterrey downtown followed by General Escobedo downtown. The rest of the metropolitan Monterrey region has the same density of emergency calls.

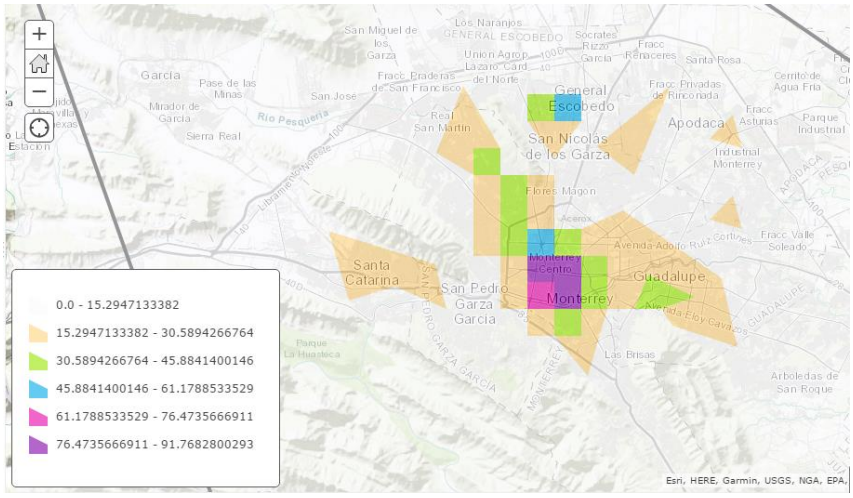

Figure 7 Density map of emergency calls

Improving ambulance response time

Ambulance response time includes the ambulance assigning and dispatching times and the one required to get to the scene of the emergency call. From Figure 2, the average time observed before the ambulance departs to service an emergency call is 7.8 minutes. Therefore, the required international standard of 10 minutes is practically reached without the ambulance being used at all. In addition, the average time taken by the ambulance to reach the call scene is 16.9 minutes.

The location of ambulance depots depends upon the behavior of service density and its dynamics throughout the day and the ambulance desired response time. Considering daily service demand requirements behavior, described in Figures 6 and 7, two different patterns are identified; a low demand level in the range of two to three services per hour occurring from the 23:01 hrs of a day to 7: 59 A.M. of the following day and; a high demand level with a range of four to five services per hour occurring the rest of the following day. Therefore, two daily ambulance deployment strategies were developed for each day.; A high-demand and a low-demand strategy for all days of the week.

The previously described information sets the general context required to guide the determination of ambulance capacity and location. The ambulance location problem has been exhaustively treated in the Operations Research area. An excellent review of ambulance location and relocation models is provided by Brotcorne et al., (2003). Also, Leigh et al., (2011) illustrate a scheme in which a variation of the double standard model used for ambulance dispatching by Gendreau et al., (1997). However, in this work, a similar scheme to the ones suggested by Ong et al,. (2010) and Peleg et al,. (2004) is used to derive such strategies. An ambulance deployment scheme with the support of the geospatial analyses and the use of the ESRI Software System was performed during this study.

The application of the previous scheme is carried out in two stages; the first stage consists of defining the optimal location for the current ambulance fleet in operations. The second stage includes the determination of the optimal fleet size required to meet the international standard for ambulance response time. The current location of the ambulances satisfy 37% of the emergency calls with a response time of less than or equal to 10 minutes. Therefore, the first stage involved the definition of the optimal location of the 21 ambulances operating from the 23:01 hrs of a day to 7: 59 A.M. of the following day and the eleven ambulances operating at night. The optimal new locations are determined by using the ESRI Software System considering an ambulance capacity of four services per shift. The expected percentage of emergency calls covered with a response time less than or equal to ten minutes is 87% for the high-demand period. For the low-demand emergency call period the expected percentage of demand with a response time less than or equal to ten minutes is 75%. Figure 8 illustrates the high-demand recommendation as an example. The expected average response time estimated for this scenario is 5.34 minutes. This increases to 6.63 minutes for the case of low-demand case.

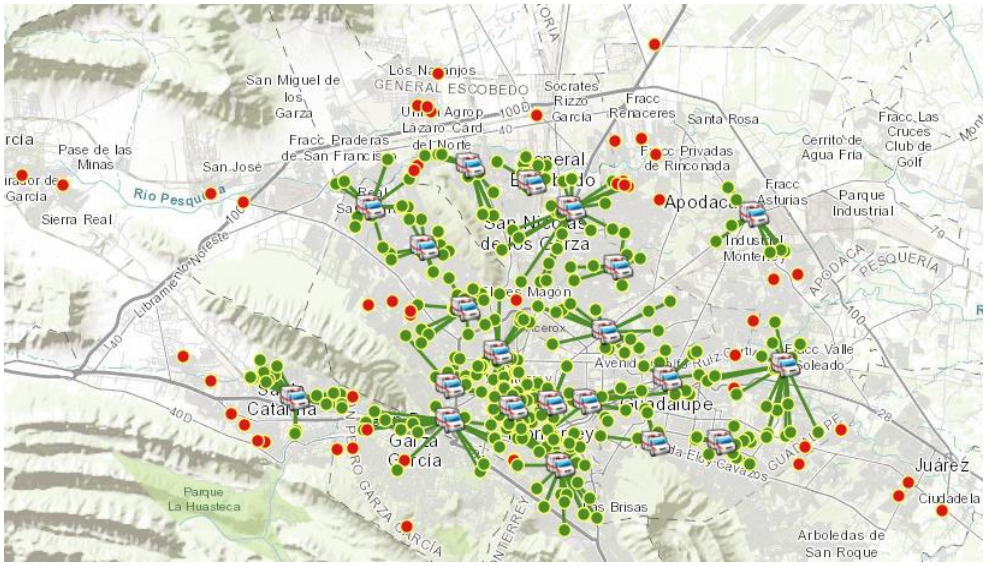

Figure 8 illustration of the optimal new ambulance location during the high-demand emergency call period

The impact of reducing ambulance turnaround time is an increase on ambulance capacity. Figure 9 illustrates how the ambulance transport time from its dispatching at the base until the arrival to the patient scene and the percentage of calls with service time less than ten minutes behave with respect to the number of ambulances in operation. Considering an increase in ambulance capacity from 4 to 5 services per day per ambulance, the number of ambulances operating can be reduced to 9 during the night shift without impacting the service level. Similarly, for 19 ambulances operating the rest of the day, the percentage of calls with time less than 10 minutes is maintained. However, the response time is increased about half a minute. Decreasing two ambulances per day from operations has an impact on the cost. Savings on labor cost equivalent to six ambulance crew shifts per day are obtained. Additionally, the cost on fuel and maintenance associated to these change will be achieved.

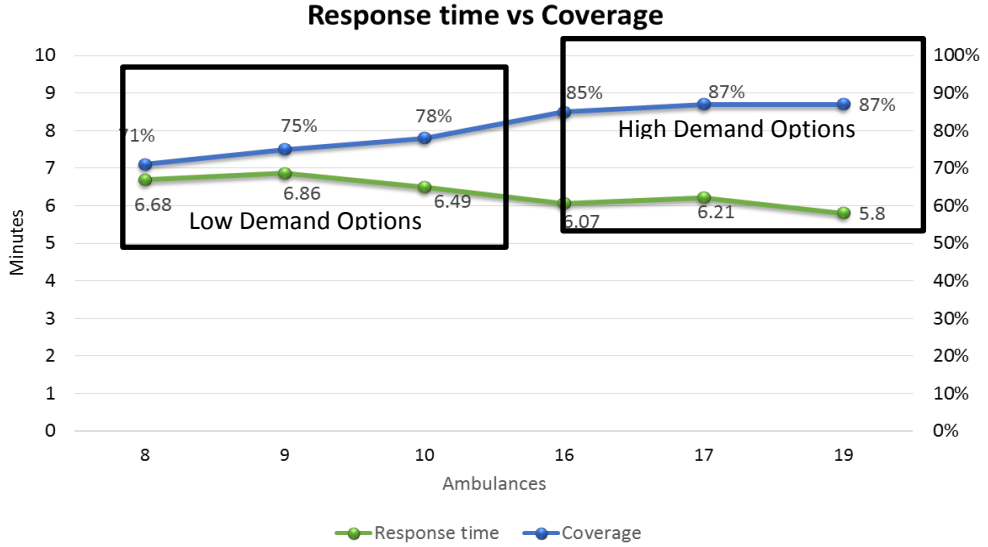

Figure 9 Description of behavior of response time and coverage percentage with respect to the number of ambulances

In summary, the increase in ambulance capacity due to the potential reduction in turnaround time will impact positively on the service level and operating costs of the Red Cross EMS operations. As shown in Figure 9, a better location of the ambulances with respect to the emergency call demand will decrease ambulance response time by about ten minutes on

average. This will improve ambulance cycle time and furthermore its capacity. The expected new cycle time is 65.8 minutes as illustrated in Figure 10.

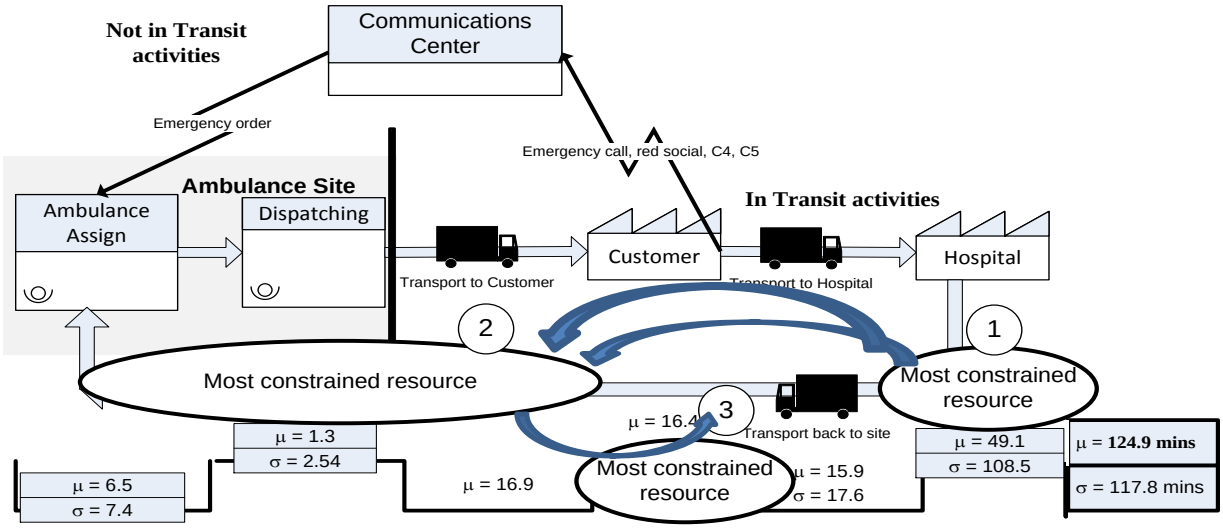

Figure 10 New improved A-VSM for the Red Cross EMS operations

The ambulance re-location initiative breaks the second bottleneck shifting the focus to a new most constrained activity; the patient stabilization activity. The institution decided to pursue the application of the scheme on this activity, and continue until all the modifications suggested so far are implemented in full and become mature and stable.

5. Conclusions

This work describes the improvement strategy of the Red Cross operations located in the metro area of Monterrey Mexico for increasing its ambulance fleet capacity. The strategy is based on the elimination of wastes (Graban 2016) guided by a Theory of Constraints scheme.

The identification of the most restricted resources/operations was facilitated by the elaboration of an Ambulance VSM used as a guiding framework. As shown in Table 2, The initial most restricted operation identified was the patient handover process. This restriction was broken through the definition of three initiatives; the creation of ambulance release teams; the re-negotiation of the responsibility of the Red Cross and each hospital in the handover process and; the use of internet-based technology to facilitate communication between the Red Cross ambulances and the Emergency departments of the hospitals. Pilot projects were undertaken for the first two initiatives in the HU and HM. The expected improvement with the previous initiatives is the reduction of ambulance cycle and turnaround time in about 50 minutes, increasing ambulance capacity from 4 to a maximum of 6 services per shift.

The next bottleneck identified corresponded to the response time of the ambulances. The project undertaken was the application of mathematical modeling for re-locating ambulances. The benefits expected with this project is an increase of the emergency call coverage from 37% 70 85% within 10 minutes. It is also expected that response time could be decreased 10 minutes on average and the maximum number of services per ambulance per shift would be 7. The analysis realized in this bottleneck supported the idea that decreasing the number of ambulances would not impact negatively the level of covering service. At the same time, it would enable management to achieve a reduction of the operating cost.

Finally, it is important to point out that the operations management officers had a very positive response to the use of the suggested scheme. It proved to be an excellent aid to focus and guide improvement efforts, and prioritize those that were more attractive.

Table 2 Summary of improvement sequence

Bottleneck number	Most restricted activity	Improvement initiatives	Cycle time (minutes)	Maximum number of services per shift per ambulance	Average response time (minutes)
1	Patient handover	- Ambulance decoupling - Negotiation of responsibility in handover process - Use of internet-based technology	75.8	6	24.2
2	Ambulance transport from base to patient	- Relocation of ambulance bases	65.8	7	14.2
3	Patient stabilization	- To be defined	To be defined	To be defined	To be defined

References

Bentley, T.G.K., Effros, R.M., Palar, K. and Keeler, E.B., Waste in the U.S. Health Care System: A Conceptual Framework, *The Milbank Quarterly***,** Vol. 86, No. 4, pp. 629 – 659, 2008.

Blackwell, T.H., Kline, J.A., Willis, J.J. and Hicks, G.M., Lack of Association Between Prehospital Response Times and Patient Outcomes, *Prehospital Emergency Care*, Vol. 13, pp. 444–450, 2009.

Brandao, D.S.L., Trends and approaches in lean healthcare, *Leadership in Health Services*, Vol. 22, No. 2, pp.121–139, 2009.

Brotcorne, L., Laporte, G. & Semet, F., Ambulance location and relocation models, *European Journal of Operational Research*, Vol. 147, pp. 451 – 463, 2003.

Chalice, R., *Stop rising healthcare costs using Toyota lean production methods: 38 steps for improvement*, Quality Press, Milwaukee, WI, 2005.

Chalice, R., *Improving Healthcare Quality using Toyota lean Production Methods: 46 Steps for Improvement*, 2nd ed., Quality Press, Milwaukee, WI, 2007.

Fitch, J.J., Knight, S., Griffiths, K. and Gerber, M., The new EMS imperative: Demonstrating value, *Infocus*, ICMA Publishing, Vol. 47, No. 1. pp. 1 – 19, 2015.

Gendreau, M., Laporte, G. and Semet, F., Solving an ambulance location model by tabu search, *Location Science*, Vol. 5, No. 2, pp. 75 – 88, 1997.

Goldratt, E.M., Cox, J. and Whitford, D., *The Goal: A Process of Ongoing Improvement*, Third Edition, North River Pr., 2012.

Graban, M., *Lean Hospitals: Improving Quality, Patient Safety and Employee Engagement*, 3rd Edition, CRC Press,2016.

Heinbuch, S.E., A case of successful technology transfer to health care: total quality materials managements and just-in-time, *Journal of Management in Medicine*, Vol. 9 No. 2, pp.48–56, 1995.

Huang, S.H., Dismukes, J.P., Shi, J. & Robinson, D.E., (2003), Manufacturing Productivity Improvement Using Effectiveness Metrics and Simulation Analysis, International Journal of Production Research, Vol. 41, No. 3, pp. 513-527, 2003.

Jeong, K. and Phillips, D.T., Operational efficiency and effectiveness measurement, *International Journal of Operations and Production Management*, Vol. 21, No. 11, pp.1404– 1416, 2001.

Keyte, B. and Locher, D., *The Complete Lean Enterprise: Value Stream Mapping for Administrative and Office Processes*, Productivity Press, 2004.

Leigh, J.M., Dunnett, S.J. and Jackson, L.M., Predictive policing using hotspot analysis, *Proceedings of the International Multiconference of Engineers and Computer Scientists*, Vol. II, IMECS 2016, March 16 – 18, Hong Kong, 2016.

Newgard, C.D., Schmicker, R.H., Hedges, J.R., Trickett, J.P., Davis, D.P., Bulger, E.M., Aufderheide, T.P., Minei, J.P., Hata, J.S., Gubler, K.D., Brown, T.B., Yelle, J., Bardarson, B., and Nichol, G., Emergency medical services

intervals and survival in trauma: Assessment of the "Golden Hour" in a north american prospective cohort, *Annals of Emergency Medicine*, Vol. 55, No. 3, pp.235 – 244, 2010.

NHS Confederation, *Zero Tolerance*, London, UK, 2012.

Ohno, T., *Toyota Production System: Beyond Large-Scale Production*, Productivity Press, Cambridge, MA., 1988.

Ong, M.E., Chiam, T.F., Ng, F.S., Sultana, P., Lim, S.H., Leong, B.S., Ong, V.Y., Tan, E.C., Tham, L.P., Yap, S., & Anantharaman, V., Reducing ambulance response times using geospatial-time analysis of ambulance deployment, *Academic Emergency Medicine*, Vol. 17, No. 9, pp. 951 – 957, 2010.

Peleg, K. and Pliskin, J.S., A geographic information system simulation model of EMS: Reducing ambulance response time, *American Journal of Emergency Medicine*, Vol. 22, No. 3, pp.164 – 170, 2004.

Pons, P.T., Jason S. Haukoos, J.S., MS, Whitney Bludworth, M.S., Thomas Cribley, T., A. Pons, A. & Markovchick, V, Paramedic response time: Does it affect patient survival? *Acad Emerg Med*, July 2005, Vol. 12, No. 7, 2005.

Rogers, F.B, and Rittenhouse, K., The golden hour in trauma: Dogma or medical folklore?, *The Journal of Lancaster General Hospital*, Vol. 9, No. 1, pp. 11 – 13, 2014.

Schonberger, R.J., *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*, Macmillan, New York, NY. 1982.

Siller, M, Villarreal, B. & Melissa Rosselly, M., A Theory of Constraints Approach to Achieve Agility: An Application, The 2nd Industrial and Systems Engineering World Conference, Las Vegas, NV, USA, November 5-7, 2013.

Simmons, D., Mason, R. and Gardner, B., Overall Vehicle Effectiveness, *International journal of Logistics: Research and* Applications, Vol. 7, No. 2, pp. 119-34, 2004.

Villarreal, B., Granda-Gutierrez, E.A., Lankenau, S., Bastidas, A.C. & Montalvo, A., Decreasing Ambulance Response Time Through an Optimal Base Location, Proceedings of the 2017 International Symposium on Industrial Engineering and Operations Management (IEOM), Bristol, UK, July 24-25, 2017.

Villarreal, B., Elizondo, M.M., Morales, B. & Turrubiates, M., A TOC-OEE Driven Approach to Increase Order On-Time Delivery, Proceedings of the 2016 Industrial and Systems Engineering Research Conference, H. Yang, Z. Kong & M.D. Sarder, eds, Anaheim, CA, May 21-28, 2016.

Villarreal, B. & López, N., A TOC-OEE Based Scheme to Improve Productivity, Proceedings of the 2014 International Conference on Industrial Engineering and Operations Management, Bali, Indonesia, 2014.

Villarreal, B., The Transportation Value Stream Map (TVSM), *European Journal of Industrial Engineering*, Vol. 6, No. 2, pp. 216-233, 2012.

Watson, L., *The Platinum Ten*, Halstead: ResQMed, Halstead, 2001.

Womack, J.P. and Jones, D.T., *Lean Thinking*, Free Press, 1996.

Zidel, T., *A Lean Guide to Transforming Healthcare: How to Implement Lean Principles in Hospitals, Medical Offices, Clinics, and other Healthcare Organizations*, Quality Press, Milwaukee, WI.,2006.

Biographies

Bernardo Villarreal is a full professor of the Department of Engineering of the Universidad de Monterrey. He holds a PhD and an MSc of Industrial Engineering from SUNY at Buffalo. He has 20 years of professional experience in strategic planning in several Mexican companies. He has taught for 20 years courses on industrial engineering and logistics in the Universidad de Monterrey, ITESM and Universidad Autonoma de Nuevo León. He has made several publications in journals such as Mathematical Programming, JOTA, JMMA, European Journal of Industrial Engineering, International Journal of Industrial Engineering, Production Planning and Control, International Journal of Logistics Research and Applications, Industrial Management and Data Systems and the Transportation Journal. He is currently a member of the IIE, INFORMS, POMS and the Council of Logistics Management.

Jose Arturo Garza-Reyes is a reader in Operations Management and Business Excellence at the Centre for Supply Chain Improvement, Derby Business School, the University of Derby, UK. He has published a number of articles in leading international journals and conferences, including International Journal of Production Research, International Journal of Production Economics, Production Planning & Control, Journal of Cleaner Production, Robotics and Computer Integrated Manufacturing, Journal of Manufacturing Technology Management, International Journal of Quality and Reliability Management, TQM & Business Excellence, International Journal of Productivity and Performance Management, among others. Dr Garza-Reyes has also written two books in the areas of quality management systems and manufacturing performance measurement systems. He has participated as guest editor for special issues in the Supply Chain Management: An International Journal, International Journal of Lean Six Sigma, International Journal of Lean Enterprise Research, International Journal of Engineering Management and Economics and International Journal of Engineering and Technology Innovation. Dr Garza Reyes is co-founder and current Editor of the International Journal of Supply Chain and Operations

Resilience (Inderscience). He is currently serving in the editorial board of several international journals as well as has contributed as member of the scientific and organizing committees of several international conferences. His research interests include general aspects of operations and manufacturing management, operations and quality improvement and performance measurement.

Jenny Díaz-Ramírez is professor at the University of Monterrey. She has worked previously as a professor at Tecnologico de Monterrey, Mexico and Pontificia Universidad Javeriana Cali, Colombia. She got a MSc in operations research from Georgia Tech and the PhD in Industrial Engineering from Tecnologico de Monterrey, Campus Toluca in 2007. Her research topics are applied optimization and statistics in topics such as health systems, air quality and logistics.

Arturo Quezada is an Industrial Engineer just graduated from Universidad de Monterrey (UDEM). He has participated on several projects such as the Improvement of the routing operations of a brewing company. He also applied Lean Thinking principles for Improving the Productivity of several assembly lines for Metalsa. Arturo is a member of the IISE and APICS Societies.

Gabriela Morales is a SUMA CUM LAUDE Industrial Engineer just graduated from Universidad de Monterrey (UDEM). She has participated on several projects such as the Improvement of the routing operations of a meat and poultry food producer firm. She also applied Lean Thinking principles for Improving the Productivity of several assembly lines for a Mexican subsidiary of GE company. She has started graduate work for a master degree in Industrial Engineering at UDEM. Gabriela is a member of the IIE and APICS Societies.

Aracely Carranza is a CUM LAUDE Industrial Engineer just graduated from Universidad de Monterrey (UDEM). She has participated on several projects such as the Improvement of the routing operations of a leading convenience store firm. She also applied Lean Thinking principles for Improving the Productivity of several metal assembly lines for a Mexican metal mechanic company. Currently, she has started to work at a Mexican firm leader in the manufacturing of frozen and refrigerated food as a transportation and traffic analyst. Aracely is a member of the IIE and ASQ Societies.