Vlerick Leuven Gent Management School
the Autonomous Management School
of Ghent University and Katholieke Universiteit Leuven

Vlerick Leuven Gent Working Paper Series 2011/05

PANDEMIC INFLUENZA VACCINE ALLOCATION PROTOCOL

ANN VEREECKE

Ann.Vereecke@vlerick.com

BEHZAD SAMII

Behzad.Samii@vlerick.com

D/2010/6482/20

PANDEMIC INFLUENZA VACCINE ALLOCATION PROTOCOL

ANN VEREECKE

Vlerick Leuven Gent Management School

BEHZAD SAMII

Vlerick Leuven Gent Management School

Contact:

Behzad Samii

Reep 1

9000 Gent

Tel: 09/210.98.50

ABSTRACT

This research investigates the impact of alternative allocation mechanisms that can be employed in the context of vaccine inventory rationing. We develop service level expressions for high priority (target groups such as healthcare professionals) and low priority (non-target groups) demand classes using Partitioned Allocation (PA), Standard Nesting (SN), and Theft Nesting (TN) allocation mechanisms. These service level expressions are then utilized to address the two interdependent problems of selecting an appropriate allocation mechanism first and then determining the optimal reserved vaccine quantity accordingly. We conduct numerical experiments on instances with different levels of inventory slackness. From the results of numerical experiments, we observe that there exist certain inventory slackness conditions under which one of the allocation mechanisms outperforms the others. We have evidence that health planners in different regions may exact-order, under-order, or over-order the vaccine inventory with respect to the actual demand for vaccines. Our analysis shows that, depending on the ordering policy devised by the health planners, it is important for the decision maker to choose PA when vaccines are under-ordered and TN when over-ordered. In the case of exact-ordering, if the target groups are in minority TN and when in majority SN is the preferred allocation mechanism.

INTRODUCTION

The recent pandemic of H1N1 and the much feared pandemic of H5N1 avian influenza have demonstrated that the threat of influenza pandemics looms large over us today. The society expects that governments, manufacturers, and international agencies will have strategies in place to quickly produce and distribute sufficient amount of effective vaccines and antiviral medicines. However, in the absence of a breakthrough in the supply of these drugs and vaccines, it is likely that we will face acute shortages in the initial period after a pandemic starts. New strains of influenza appear every year and it takes time to identify the new strain and design a targeted vaccine against it. Also, the efficacy of a more generic vaccine is likely to be very low against a novel strain that has pandemic potential. Thus, there is broad consensus that we will not have sufficient volumes of vaccines to vaccinate the entire population when there is an outbreak and we will have to rely on good targeting and rationing strategies for the quantity of vaccine available. Even in a regular influenza season like the 2004-2005 in the United States, we experienced vaccine shortages because of demand and supply uncertainties (Williams and Yadav 2005). For both pandemic and regular influenza public health officials have to develop protocols for rationing the available vaccine based on multiple competing objectives. Priority for vaccination may be given to certain population segments. It is evident that personnel in healthcare institutions who are critical in assisting those who are sick should be given priority in obtaining the vaccine. This will decrease the risk of transmission from health care personnel to patients and avoid absenteeism of essential medical personnel during outbreaks. Some population segments may be important to prioritize for vaccination due to the usefulness of their work in controlling the pandemic e.g. those who work in companies that manufacture vaccines. Those who are medically the neediest and are likely to face the highest mortality from the disease should also be given high priority. Similarly, from the standpoint of slowing the progression of the outbreak priority can be given to those who present the highest risk of transmitting the disease onwards e.g. school children. On the other hand, those who argue in favor of equity and egalitarian health care will contend that the general population should also have sufficient access to the vaccine and availability only to high priority segments is unfair. At least some amount of the vaccine should be made available to the general population even if on a strictly First Come First Served (FCFS) basis. Public health planners thus face the challenging problem of allocating scarce quantities of influenza vaccine to a population consisting of priority segments and non-priority segments. The key question they face is how to set the rationing levels or service levels for each of the population segments. Some would argue that the objective is to minimize the overall social costs due to mortality and morbidity from the pandemic. Thus a rationing model should be built to achieve

this objective while incorporating disease transmission, inventory to treat those infected, and other such dynamics. Such a model however suffers from numerous limitations as highlighted below: (1) Although there are measures to capture the social cost of mortality and morbidity such as loss of Disability Adjusted Life Year (DALY), Case Fatality Rate (CFR), and (as in van Hoek et al. 2011) loss of Quality Adjusted Life Year (QALY) there is little consensus on their applicability and use in large scale pandemic situations; (2) Modeling disease transmission and healthcare inventory availability in the event of a pandemic require data on contagiousness of influenza strain in different groups. Given that the pandemic strains are novel, such data does not exist and we can only extrapolate from parameters observed in past virus strains. Due to the complexities of the decision and the non-applicability of historical data on different parameters, public health planners typically create a list of high priority segments (also called target groups) that need to obtain the vaccine (or anti-viral drug like Tamiflu®) and set a service level they want to achieve for these segments. These service levels are set based on subjective assessment that the decision makers want to see as their impact on the overall wellbeing of the society. Setting the high priority service level (and the corresponding quantity of vaccine reserved) too high leads to the general population not being able to get vaccinated; whereas setting it too low implies that a larger fraction of the high priority segments will remain unvaccinated. This decision dilemma of the public health vaccine allocator is captured by the model presented in this research. This model outputs the service levels of both priority classes for all possible reserved vaccine quantities, given the three inputs by the health planner of the size of the target and non-target groups and the available vaccine inventory at the beginning of the vaccination season. In inventory rationing problems, a FCFS approach will provide suboptimal results if the long term penalty costs of stock-outs are different among customer classes. For this reason, alternative mechanisms for allocating arriving demand to reserved and unreserved inventory, under the assumption that decisions upon acceptance/rejection are taken in real time, have been employed (Nahmias and Demmy 1981). Partitioned Allocation (PA), Standard Nesting (SN), and Theft Nesting (TN) are the most commonly used allocation mechanisms discussed in the literature (Talluri and van Ryzin, 2005). It is important to note that the model presented in this research can show the impact of these three allocation mechanisms on the outputs expected by the decision maker. Consequently, the health planner should decide about the choice of the allocation mechanism as well as the desired reserved vaccine quantity associated with the selected mechanism. In the two-class system which we are addressing demand from the low priority class (non-target groups) can only be fulfilled from the unreserved portion of the vaccine inventory. However, the allocation of the reserved and unreserved vaccine inventory to the incoming high priority demand is triggered differently under PA, SN, or TN. In Partitioned Allocation, as the name suggests, demand from the high (low) priority class consumes

only the reserved (unreserved) quantity. Thus, no competition exists between the two demand classes. In Standard Nesting, demand from the high priority class first consumes the reserved quantity; once and if the reserved quantity is exhausted, the high priority class competes equally on a FCFS basis with classes of lower priority for any remaining unreserved vaccine inventory. In a different fashion at Theft Nesting, demand from high and low priority classes initially compete for the unreserved vaccine inventory on a FCFS basis. Once and if the unreserved vaccine inventory has been exhausted, additional demand from the low priority class is rejected and demand from the high priority class consumes the remaining inventory. For each allocation mechanism, we develop service level expressions for the low and high priority demand classes. Such formulations have, so far, not been available in the literature. We utilize the service level expressions to address the two interdependent problems of selecting an appropriate allocation mechanism and determining the optimal reserved quantity. We conduct numerical analysis on numerous instances of low and high priority demand rates to observe that there exists certain inventory slackness condition under which one of the allocation mechanisms is preferred. This research contributes to the existing literature in revenue management and inventory rationing in two distinct ways. First, while imposing no order arrival pattern, we model a dual-criteria inventory rationing problem based on system fill rates rather than widely used backlog costs which are obviously hard to estimate in the human life context. Second, we provide closed form expressions for the service levels of the two customer classes under each allocation mechanism. Our analysis shows that, depending on the availability of vaccine inventory and the demand rates of the different customer classes, it is important for the decision maker to choose the right allocation mechanism. Although we limit our analysis to a single vaccination period two demand classes vaccine inventory rationing problem, we expect that our approach and our results can be utilized in a wide range of vaccine supply chains with multiple service differentiated demand classes. The remainder of this paper is organized as follows: in section 2 we review the related literature in both inventory rationing and revenue management and highlight our relative contribution. In section 3 we provide a formal characterization of the vaccine allocation time structure under portioned, standard nesting, and theft nesting allocation mechanisms; and consequently develop exact expressions for the high and low priority classes service levels. The results of the numerical experiments are presented in section 4. In section 5 we summarize our findings and point toward the future research that can be conducted based on the results of this research.

2. LITERATURE REVIEW

In the literature on inventory theory, the early work of Veinott (1965) considers multiple demand classes in a multi-period, single-product setting using critical level policy. Topkis (1968) considers a periodic review inventory model using the critical level policy and also solves the problem of how inventory should be allocated between demand classes. Each demand class is characterized by a different shortage cost and the allocation is based on the trade-off between the benefit of filling demand for low class items in the current period versus reserving the available inventory to fill higher class items in subsequent periods. Similar to his work, we also consider that there is only a single replenishment opportunity which is at the beginning of the period. Evans (1968) and Kaplan (1969) present similar periodic review models to Topkis (1968) but with a different set of assumptions about the operating characteristics or customers repurchase behavior. In the first continuous review model, Nahmias and Demmy (1981) evaluate fill rates for given rationing and reorder levels in a *(s , Q)* inventory system with Poisson demand and two demand classes and inventory is rationed according to critical level policy suggested by Veinott (1965). Other more recent examples of research characterizing the optimal inventory rationing policy are Ha (1997a), Ha (1997b), and de Vericourt et.al (2002). Benjaafar et al. (2006) extend these approaches to account for multiple products and facilities. Deshpande et al. (2003) present a model for inventory rationing under multiple demand classes which includes setup costs, lead times, and customer backorders in a continuous time framework. They derive closed-form expressions for performance measures, such as average backorders and fill rate, for a given *(Q, r, K)* type threshold rationing policy and also present an algorithm for computing the optimal policy parameters. They formulate a cost minimization problem to choose policy parameters to minimize the sum of setup costs, holding costs, and customer shortage costs. Melchiors et al. (2000) also assume a *(s , Q)* inventory model with two demand classes. They consider a lost sales environment where low priority class demand is rejected when inventory level drops below the critical level. This assumption allows them to derive an exact expression for the expected cost, unlike the approximate results of Nahmias and Demmy (1981). Melchiors (2003) extends this work and allows a non-stationary critical level policy where critical levels depend on the elapsed time since the outstanding order is triggered. Arslan et al. (2007) develop a model for cost evaluation and optimization under the assumptions of Poisson demand, deterministic replenishment lead time and a continuous review (*Q,R*) policy with rationing. They analyze the service level problem, the cost minimization problem, and the service time problem. The service level problem aims to achieve a pre-specified service level for each demand class by incurring the minimum long run average inventory holding costs. One of the major differences between the inventory rationing papers described above and our work is that optimal control policies and

rationing levels are determined on the basis of holding and backorder cost or lost sales cost in these papers. We consider that the optimal reservation decision consists of two interdependent problems of selecting an appropriate allocation mechanism and determining the optimal reservation quantity both determined based on the service levels of high and low priority demand classes. Contrary to our work, since they do not deal with human lives specifically, the models described above explicitly assume cost functions for high priority orders that cannot be fulfilled in time. The task of setting protection levels on perishable resources such as hotel rooms or flight seats inventory for different demand classes also exists in many revenue management problems. Again, in contrast to our work they generally utilize the standard revenue management approach of maximizing profit (or revenue) by reserving inventory for future orders with a higher revenue or profit potential. They also typically make two important assumptions: (1) demand for low fare classes arrives before demand for high fare classes (LBH) and (2) no long-term implications if demand cannot be fulfilled (i.e. no loss of goodwill, reputation, etc.). Both assumptions are realistic in the revenue management's commonly applied fields such as the airline and hotel industry. There are also examples of research in the revenue management literature that do not assume LBH. Lee and Hersh (1993) divide the booking period into several time intervals with diminishing length as they get closer to the end of the booking period. Request for booking in those intervals follow a Poisson process. They introduce a discrete-time dynamic programming model for finding an optimal booking policy which can be reduced to a set of critical values for the multiple fare class problems. Robinson (1995) does not assume LBH, instead considers that requests for different fare classes arrive at non-overlapping intervals. Also, Talluri and van Ryzin (2004) analyze the complex multiple fare class problem based on consumer choice that requires information not only on the actual arrival process but also on the choice behavior. There are several other papers in revenue management (especially in the single-leg problem) with various assumptions on demand distribution or arrivals sequence. For further discussion see Brumelle and Walczak (2003), McGill and van Ryzin (1999), and Talluri and van Ryzin (2004). In all models described above, either LBH or other rigid assumptions about the arrival processes have been made and the resulting optimal booking control policies depend on these specific assumptions. In our model, in line with the arrival of the vaccine customers, we assume that demand from the two customers' classes arrive according to independent Poisson processes and we do not impose a specific sequence of arrival, e.g. LBH. Modeling demand arrival through stochastic processes (rather than probability densities) is a prerequisite for developing exact service level expressions for different types of nesting (i.e. Standard and Theft Nesting). Therefore, our modeling approach is distinct from previous models developed in the field of revenue management. According to Talluri and van Ryzin (2005), the Standard and Theft Nesting are in fact equivalent if demand

arrives in low-to-high class order. However, they acknowledge that in practice demand rarely arrives in low-to-high order and the choice of Standard versus Theft Nesting matters. In our model, we not only provide closed form expressions for the service levels of two customer classes under each allocation mechanism but also provide examples of strict preference for Partitioned Allocation, Standard Nesting, or Theft Nesting inventory allocation mechanisms based on relevant system parameters.

3. MODEL FORMULATION

In our model we consider a single perishable influenza vaccine with an available inventory of $x > 0$ units at the beginning of the vaccination period $[0, \tau]$ with no replenishment opportunity during this period. Under ideal circumstances, the available vaccine inventory in a region should be the total population times the number of doses of vaccine required per person. However, the long production lead time and the production capacity of vaccine providers may dictate a lower available quantity. This vaccine inventory is used to immunize two customer classes $k \in \{L, H\}$ with high (H) and low (L) priority. Class H includes the susceptible target groups such as pregnant women, healthcare professionals, and children under the age of six. Class L corresponds to the population not considered as class H. The healthcare planners decide on the inclusion of different population groups into the class H before the vaccination period starts. We assume that demand of class *k* follows a Poisson process with rate λ_k and that each unit of demand requires one unit of vaccine inventory. The demand rate should be determined based on the severity of the new influenza strain and the public perception of vaccine efficacy. For example, in the less sever influenza season of 2008-2009 only 46% of healthcare workers opt to get vaccinated against influenza (HIDA 2009). We denote by $D_k(t)$ the random demand from class *k* in the time interval $[0, t]$ and by $d_k(t)$ its corresponding realizations. Upon arrival, the decision maker accepts or rejects an arriving individual based on vaccine inventory availability and a predefined allocation mechanism. When employing conventional FCFS allocation, demand is fulfilled regardless of the customer class until the vaccine inventory *x* is entirely consumed. We consider a setting in which a decision maker wants to dedicate a portion $r(0 \le r \le x)$ of the available vaccine inventory exclusively to class H to ensure a higher probability to fulfill all demand (i.e. service level) for this customer class than under FCFS allocation. In the revenue management literature (Talluri and van Ryzin 2005), r is usually referred to as the "protection level". Correspondingly, the unreserved amount of inventory $x - r$ can be considered as

a “booking limit” for class L demand. Any class L demand exceeding $x-r$ will not be fulfilled. Three common mechanisms for allocating demand of classes H and L to the reserved and unreserved portion of the vaccine inventory can be distinguished: (1) Partitioned Allocation (PA) when class H and class L demand are strictly allocated to their corresponding vaccine inventory buckets defined by r and $x-r$, respectively. Class H demand can only utilize r and class L demand only $x-r$ units of the overall available vaccine inventory. After r units of demand from class H have arrived, any additional demand from class H will be rejected. Similarly, after $(x-r)$ units of demand from class L have arrived, any additional demand from class L will be rejected. (2) Standard Nesting (SN) when demand from class H first consumes the reserved quantity r*;* once and if r is entirely consumed, class H demand competes equally on a FCFS basis with class L demand for any remaining unreserved vaccine inventory. Demand from class L can only be fulfilled from the unreserved portion $x-r$. (3) Theft Nesting (TN) when demand from classes H and L first compete on a FCFS basis for the first $x-r$ units of the unreserved vaccine inventory. Once and if $x-r$ units of inventory have been entirely consumed, additional class L demand is rejected and class H demand consumes the remaining r units of vaccine inventory. This implies that during the first $x-r$ arrivals, class H demand “steals” from the unreserved portion of inventory and the reserved units are consumed by class H demand only after $x-r$ units of demand have arrived to the system. Although PA is of less practical interest, its analysis will help us develop service level measures for SN and TN. In the following we first provide expressions for class H and class L service levels under PA. Based on these expressions we will then develop the corresponding service level measures for SN and TN.

3.1 Service levels under Partitioned Allocation (PA)

When employing PA, computation of the service level for both customer classes is straightforward. Let $\alpha_k^{PA}(.)$ denote the probability that the entire demand of class *k* in the vaccination period $[0,\tau]$ is fulfilled using PA mechanism. Let $F_k(.)$ denote the cdf of class *k* demand in the planning period. Because r and $x-r$ are exclusively dedicated to class H and L respectively, we know that $\alpha_H^{PA}(r)=F_H(r)$ and $\alpha_L^{PA}(r)=F_L(x-r)$. Before turning to the service level expressions for SN and TN, we use the alternative approach of Samii et al. (2011) in calculating the service levels for class H and L under PA. Although this approach slightly complicates matters for PA, it will later be helpful in deriving the SN and TN service level expressions. Let $V_H(r)\equiv\inf\{t\geq 0: D_H(t)=r\}$ denote the random time required to observe r units of demand

from class H, and v_H denotes its corresponding realization. The entire demand of class H can only be fulfilled if: (1) r is not exhausted before the end of the planning period (i.e. $\{V_H(r) \geq \tau\}$) ;or (2) r is exhausted before the end of the period, but no further demand of class *H* arrives thereafter, i.e., $\{(V_H(r) < \tau) \wedge (D_H(\tau - V_H(r)) = 0)\}$. For notational ease, we denote the two event sets $\{V_H(r) \geq \tau\}$ and $\{(V_H(r) < \tau) \wedge (D_H(\tau - V_H(r)) = 0)\}$ by $\xi_{H,1}^{PA}$ and $\xi_{H,2}^{PA}$. Because the two event sets are mutually exclusive, we can express the class H service level by:

$$\alpha_H^{PA}(r) = \Pr\{\xi_{H,1}^{PA}\} + \Pr\{\xi_{H,2}^{PA}\} \quad (1)$$

Analogously, we can write the class L service level under PA. Let $V_L(x-r) \equiv \inf\{t \geq 0 : D_L(t) = x-r\}$ denote the random time required to observe $x-r$ units of demand from class L, by v_L its corresponding realization, and by $\xi_{L,1}^{PA}$ and $\xi_{L,2}^{PA}$ the two event sets $\{V_L(x-r) \geq \tau\}$ and $\{(V_L(x-r) < \tau) \wedge (D_L(\tau - V_L(x-r)) = 0)\}$. Then,

$$\alpha_L^{PA}(r) = \Pr\{\xi_{L,1}^{PA}\} + \Pr\{\xi_{L,2}^{PA}\} \quad (2)$$

Knowing that $V_H(r)$ and $V_L(x-r)$ are Erlang distributed random variables, we can explicitly write (1) and (2) as $\alpha_H^{PA}(r) = \int_{\tau}^{\infty} f_e(v_H; r, \lambda_H) dv_H + \int_0^{\tau} f_e(v_H; r, \lambda_H) f_p(0, \lambda_H(\tau - v_H)) dv_H$ and $\alpha_L^{PA}(r) = \int_{\tau}^{\infty} f_e(v_L; x-r, \lambda_L) dv_L + \int_0^{\tau} f_e(v_L; x-r, \lambda_L) f_p(0, \lambda_L(\tau - v_L)) dv_L$; where $f_e(v_H; r, \lambda_H)$ denotes the pdf of the Erlang distributed random arrival time $V_H(r)$ with rate parameter λ_H and shape parameter r; i.e. $V_H(r) \sim Erlang(\lambda_H, r)$. Also, $f_p(0, \lambda_H(\tau - v_H))$ denotes the Poisson pmf, i.e. probability of observing zero units of demand from a Poisson process with rate λ_H in the time interval $[v_H, \tau]$ (details provided in Online Supplement OS.1).

3.2 Service levels under Standard Nesting (SN)

As described previously, under SN class H demand competes equally with class L demand for the remaining part of $x-r$ that has not been consumed until time $V_H(r)$, i.e.

$\max\{0, x-r-D_L(V_H(r))\}$. It is intuitive to see that for $r < x$ the class H (L) service level under SN will be higher (lower) than under PA due to additional equal competition opportunities. Let $\delta_H^{SN}(r)$ denote the class H service level increment and $\delta_L^{SN}(r)$ the class L service level decrement induced by SN as compared to PA. In most general terms, we can express the service levels of class H and L under SN by $\alpha_H^{SN}(r) = \alpha_H^{PA}(r) + \delta_H^{SN}(r)$

and $\alpha_L^{SN}(r) = \alpha_L^{PA}(r) - \delta_L^{SN}(r)$. In the following, we develop expressions for $\delta_H^{SN}(r)$ and $\delta_L^{SN}(r)$ and use these expressions to derive exact formulations for the service level under SN for class H and class L.

3.2.1 Class H service level under SN

To determine the service level increment $\delta_H^{SN}(r)$ induced by SN, we expand our previous approach for calculating the class H service level under PA. Based on the random time $V_H(r)$ we identified two mutually exclusive sets of events $\xi_{H,1}^{PA}$ and $\xi_{H,2}^{PA}$ that lead to fulfillment of the entire class H demand. Now, in SN, we are interested in class H demand that can be fulfilled from the unreserved portion $x-r$ between $V_H(r)$ and the time the overall inventory *x* is exhausted. We denote by $S(r)$ the random time at which the overall inventory x is exhausted; $S(r) \equiv \inf\{t > V_H(r) : D_H(t) + \min(r, x - D_H(t)) = x\}$. For time interval $]V_H(r), S(r)]$, we can identify two additional event sets that lead to fulfillment of the entire class H demand. Figure 1 provides an illustration that will help introducing the two events sets. The graph on the left-hand side of Figure 1 illustrates the case in which $V_H(r) < \tau \le S(r)$, reserved quantity r is exhausted within the planning horizon ($V_H(r) < \tau$) but the overall inventory *x* is sufficient to fulfill the entire (classes H and L) demand arriving until the end of the period ($\tau \le S(r)$). Note that we are only interested in those instances in which some class H demand materializes in time interval $]V_H(r), \tau]$, i.e. $D_H(\tau - V_H(r)) > 0$; instances with $D_H(\tau - V_H(r)) = 0$ are already included in the event set $\xi_{H,2}^{PA}$. We define the associated events set $\xi_{H,3}^{SN} = \{(V_H(r) < \tau \le S(r)) \wedge (D_H(\tau - V_H(r)) > 0)\}$.

Insert Figure 1 About Here

The graph on the right-hand side of Figure 1 illustrates the case in which $V_H(r) < S(r) < \tau$, implying that the overall inventory x is exhausted before the end of the planning horizon. In such instances, the entire class H demand will only be fulfilled if no further class H demand materializes in the time interval $]S(r), \tau]$, i.e. $D_H(\tau - S(r)) = 0$. We define the associated events set $\xi_{H,4}^{SN} = \{(V_H(r) < S(r) < \tau) \wedge (D_H(S(r) - V_H(r)) > 0) \wedge (D_H(\tau - S(r)) = 0)\}$. Because instances with $D_H(\tau - V_H(r)) = 0$ are already captured by $\xi_{H,2}^{PA}$, we require $D_H(S(r) - V_H(r)) > 0$ so that $\xi_{H,4}^{SN} \cap \xi_{H,2}^{PA} = \varnothing$. Since $\xi_{H,3}^{SN}$ and $\xi_{H,4}^{SN}$ are also mutually exclusive, the class H increment compared to PA can be expressed as $\delta_H^{SN}(r) = \Pr\{\xi_{H,3}^{SN}\} + \Pr\{\xi_{H,4}^{SN}\}$. Using the previous conventions for denoting the pdf of Erlang and pmf of Poisson distributed random variables, we can express

$$\Pr\{\xi_{H,3}^{SN}\} = \int_0^{\tau} \sum_{i=r+1}^{x} \sum_{j=0}^{x-i} f_e(v_H; r, \lambda_H) f_p(j, \lambda_L \tau) f_p(i-r, \lambda_H(\tau - v_H)) dv_H$$

(3)

Because (3) is a convolution of continuous Erlang and discrete Poisson distributions, it is rather complex to compute. Expanding and rearranging terms yields:

$$\Pr\{\xi_{H,3}^{SN}\} = \sum_{i=r+1}^{x} \sum_{j=0}^{x-i} f_p(j, \lambda_L \tau) \int_0^{\tau} \frac{\lambda_H^r e^{-\lambda_H v_H} v_H^{r-1}}{(r-1)!} \frac{(\lambda_H(\tau - v_H))^{i-r} e^{-\lambda_H(\tau - v_H)}}{(i-r)!} dv_H$$

$$\Rightarrow \sum_{i=r+1}^{x} \sum_{j=0}^{x-i} f_p(j, \lambda_L \tau) \frac{\lambda_H^i e^{-\lambda_H \tau}}{(r-1)!(i-r)!} \int_0^{\tau} (\tau - v_H)^{i-r} v_H^{r-1} dv_H$$

Since the integral part of the product is an Euler Hypergeometric Integral (EHI), we can substitute $\int_0^{\tau} (\tau - v_H)^{i-r} v_H^{r-1} dv_H$ by $\frac{(r-1)!(i-r)!\tau^i}{i!}$ to get $\Pr\{\xi_{H,3}^{SN}\} = \sum_{i=r+1}^{x} \sum_{j=0}^{x-i} f_p(j, \lambda_L \tau) \frac{\lambda_H^i e^{-\lambda_H \tau}}{(r-1)!(i-r)!} \frac{(r-1)!(i-r)!\tau^i}{i!}$. Simplification of terms yields:

$$\Pr\{\xi_{H,3}^{SN}\} = \sum_{i=r+1}^{x} \sum_{j=0}^{x-i} f_p(j, \lambda_L \tau) f_p(i, \lambda_H \tau)$$

(4)

The interpretation of (4) is rather interesting. The expression represents the probability of meeting the overall class L and H demand when the available vaccine inventory is sufficient, despite receiving more class H demand than what is exclusively reserved. Since (4) is a convolution of two (truncated) Poisson distributions, it allows us to compute the probability of the events in set $\xi_{H,3}^{SN}$ without explicitly having to account for the time structure in the arrival process. To obtain an explicit formulation for $\Pr\left\{\xi_{H,4}^{SN}\right\}$, we need to consider whether the last unit of inventory is consumed by class H or class L demand. We denote by $z\ (z>r)$ the last unit of demand from class H that arrives in the time interval $[0,S(r)]$. If $x-z$ units of demand from class L have already arrived, the last unit of class H demand consumes the last available inventory unit and $S(r)\sim Erlang(\lambda_H,z)$. We denote by $\xi_{H,4,1}^{SN}\subseteq\xi_{H,4}^{SN}$ the subset of events in which the last unit of inventory is consumed by class H demand. If, however, the last unit of inventory is consumed by class L, then $S(r)\sim Erlang(\lambda_L,x-z)$. We denote by $\xi_{H,4,2}^{SN}\subseteq\xi_{H,4}^{SN}$ the subset of events in which the last unit of inventory is consumed by class L demand. Because $\xi_{H,4,1}^{SN}\cap\xi_{H,4,2}^{SN}=\varnothing$;

$$\Pr\left\{\xi_{H,4}^{SN}\right\}=\Pr\left\{\xi_{H,4,1}^{SN}\right\}+\Pr\left\{\xi_{H,4,2}^{SN}\right\} \tag{5}$$

The individual probabilities can explicitly be written as:

$$\Pr\left\{\xi_{H,4,1}^{SN}\right\}=\int_0^{\tau}\sum_{z=r+1}^{x}\sum_{w=1}^{\infty}f_e(s;z,\lambda_H)\,f_p(x-z,\lambda_L s)\,f_p(w,\lambda_L(\tau-s))\,f_p(0,\lambda_H(\tau-s))ds \tag{6}$$

and

$$\Pr\left\{\xi_{H,4,2}^{SN}\right\}=\int_0^{\tau}\sum_{z=r+1}^{x}\sum_{w=1}^{\infty}f_e(s;x-z,\lambda_L)\,f_p(z,\lambda_H s)\,f_p(w,\lambda_L(\tau-s))\,f_p(0,\lambda_H(\tau-s))ds \tag{7}$$

.

To simplify expression (6), we write

$$\Pr\left\{\xi_{H,4,1}^{SN}\right\}=\int_0^{\tau}\sum_{z=r+1}^{x}f_e(s;z,\lambda_H)\,f_p(x-z,\lambda_L s)\,f_p(0,\lambda_H(\tau-s))ds$$

$$-\int_0^{\tau}\sum_{z=r+1}^{x}f_e(s;z,\lambda_H)\,f_p(x-z,\lambda_L s)\,f_p(0,\lambda_L(\tau-s))\,f_p(0,\lambda_H(\tau-s))ds$$

Expanding the expressions for the Erlang pdf and Poisson pmf gives us:

$$\Pr\left\{\xi_{H,4,1}^{SN}\right\}=\int_0^{\tau}\sum_{z=r+1}^{x}\frac{\lambda_H^z s^{z-1}e^{-\lambda_H s}}{(z-1)!}\frac{(\lambda_L s)^{x-z}e^{-\lambda_L s}}{(x-z)!}e^{-\lambda_H(\tau-s)}ds-\int_0^{\tau}\sum_{z=r+1}^{x}\frac{\lambda_H^z s^{z-1}e^{-\lambda_H s}}{(z-1)!}\frac{(\lambda_L s)^{x-z}e^{-\lambda_L s}}{(x-z)!}e^{-\lambda_L(\tau-s)}e^{-\lambda_H(\tau-s)}ds \tag{8}$$

Defining $u \equiv \lambda_L s$; the first term in (8) can be written as:

$$\sum_{z=r+1}^{x} \frac{\lambda_H^z \lambda_L^{x-z} e^{-\lambda_H \tau}}{(z-1)!(x-z)!} \int_0^{\lambda_L \tau} \left(\frac{u}{\lambda_L}\right)^{x-1} e^{-u} \left(\frac{du}{\lambda_L}\right) \tag{9}$$

The integral part of the product in (9) is a Lower Incomplete Gamma Function (LIGF). Therefore, the first term in (8) can be expressed as:

$$\sum_{z=r+1}^{x} \frac{(x-1)!}{(z-1)!(x-z)!} \left(\frac{\lambda_H}{\lambda_L}\right)^z e^{-\lambda_H \tau} \left\{1 - \sum_{y=0}^{x-1} f_p(y, \lambda_L \tau)\right\} \tag{10}$$

Simplifying the second term in (8) yields: $\sum_{z=r+1}^{x} \frac{\lambda_H^z \lambda_L^{x-z} e^{-(\lambda_L+\lambda_H)\tau} \tau^x}{(z-1)!(x-z)!x}$ (11)

Subtracting (11) from (10) gives an equivalent formulation for (8):

$$\Pr\left\{\xi_{H,4,1}^{SN}\right\} = \sum_{z=r+1}^{x} \sum_{y=x}^{\infty} f_p(z, \lambda_H \tau) f_p(y, \lambda_L \tau) \frac{(x-1)!z}{(x-z)!(\lambda_L \tau)^z} - \sum_{z=r+1}^{x} \frac{\lambda_H^z \lambda_L^{x-z} e^{-(\lambda_L+\lambda_H)\tau} \tau^x}{(z-1)!(x-z)!x} \tag{12}$$

Analogously, after some algebra, (7) can be written as:

$$\Pr\left\{\xi_{H,4,2}^{SN}\right\} = \sum_{z=r+1}^{x} \sum_{y=x}^{\infty} f_p(z, \lambda_H \tau) f_p(y, \lambda_L \tau) \frac{(x-1)!}{(x-z-1)!(\lambda_L \tau)^z} - \sum_{z=r+1}^{x} \frac{\lambda_H^z \lambda_L^{x-z} e^{-(\lambda_L+\lambda_H)\tau} \tau^x}{z!(x-z-1)!x}$$

Now, (5) can be expressed as follows:

$$\Pr\left\{\xi_{H,4}^{SN}\right\} = \sum_{z=r+1}^{x} \sum_{y=x}^{\infty} f_p(z, \lambda_H \tau) f_p(y, \lambda_L \tau) \frac{x!}{(x-z)!(\lambda_L \tau)^z} - \sum_{z=r+1}^{x} f_p(z, \lambda_H \tau) f_p(x-z, \lambda_L \tau)$$

Defining $y \equiv z + j$ and simplifying yields:

$$\Pr\left\{\xi_{H,4}^{SN}\right\} = \sum_{z=r+1}^{x} \sum_{j=x-z+1}^{\infty} f_p(j, \lambda_L \tau) f_p(z, \lambda_H \tau) \frac{x!\, j!}{(x-z)!(z+j)!}$$

$$f_h(z; z+j, z, x) = \frac{\binom{z}{z}\binom{j}{x-z}}{\binom{z+j}{x}}$$

Because where $f_h(z; z+j, z, x)$ denotes the pmf of a random variable z that follows a Hypergeometric distribution with parameters $z+j$, z, x; we can write:

$$\Pr\left\{\xi_{H,4}^{SN}\right\} = \sum_{z=r+1}^{x} \sum_{j=x-z+1}^{\infty} f_p(j, \lambda_L \tau) f_p(z, \lambda_H \tau) f_h(z; z+j, z, x) \tag{13}$$

Expression (13) again has a rather interesting interpretation. It captures the probability that the total demand of class H and L exceeds the available vaccine inventory *x*, and that z units of class H demand are observed within the first x arrivals. Expression (13) again allows us to compute the

probability of the event $\xi_{H,4}^{SN}$ without explicitly having to account for the time structure in the arrival process. Replacing z with i for notational brevity, we can now express the class H service level increment induced by SN as:

$$\delta_H^{SN}(r) = \sum_{i=r+1}^{x} \sum_{j=0}^{\infty} f_p(j, \lambda_L \tau) f_p(i, \lambda_H \tau) f_h\left(i; max(i+j, x), i, x\right)$$

. Finally, the resulting class H service level in SN can be expressed as:

$$\alpha_H^{SN}(r) = \alpha_H^{PA}(r) + \sum_{i=r+1}^{x} \sum_{j=0}^{\infty} p(j; \lambda_L, \tau)\, p(i; \lambda_H, \tau) h\left(i; max(i+j, x), i, x\right)$$

(14)

Expectedly, the class H service level in SN is increasing in *r*, with the marginal increase of:

$$\alpha_H^{SN}(r+1) - \alpha_H^{SN}(r) = \sum_{j=x-r}^{\infty} \sum_{k=0}^{r} f_p\left(r+1, \lambda_H \tau\right) f_p\left(j, \lambda_L \tau\right) f_h\left(k; j+r+1, r+1, x\right) > 0$$

3.2.2 Class L service level under SN

We now utilize a similar approach as in the previous section to determine an expression for $\delta_L^{SN}(r)$, the class L service level decrement induced by SN. In calculating $\delta_L^{SN}(r)$ we have to consider those instances in which the entire class L demand cannot be fulfilled because class H consumes some portion of $x-r$ that (under PA) would have been sufficient to fulfill the entire class L. Such instances occur when the entire vaccine inventory is exhausted before the end of the planning horizon (i.e. $S(r) < \tau$) and $D_L(\tau) \le x-r$. The corresponding set $\xi_{L,3}^{SN} = \left\{\left(S(r) < \tau\right) \wedge \left(D_L(\tau) \le x-r\right) \wedge \left(D_L(\tau - S(r)) > 0\right)\right\}$ includes all events in which class L demand cannot be completely fulfilled because of SN. The service level decrement induced by SN is then simply the probability of such an event: $\delta_L^{SN}(r) = \Pr\left\{\xi_{L,3}^{SN}\right\}$.

Finally (details in Online Supplement OS.2), the class L service level under SN can thus be written as:

$$\alpha_L^{SN}(r) = \alpha_L^{PA}(r) - \sum_{i=r+1}^{x} \sum_{j=1}^{i-r} \sum_{k=0}^{\infty} f_p(x-i+j, \lambda_L \tau) f_p(i+k, \lambda_H \tau) f_h\left(i; x+j+k, i+k, x\right) \quad (15)$$

The class L service level in SN is decreasing in *r*, with the marginal decrease of:

$$\alpha_L^{SN}(r+1) - \alpha_L^{SN}(r) = -\sum_{i=0}^{\infty} f_p\left(i, \lambda_H \tau\right) f_p\left(x-r, \lambda_L \tau\right) f_h\left(x-r; i+x-r, x-r, x-r+\min(i, r)\right) < 0$$

3.3 Service levels under Theft Nesting (TN)

Under TN, the order in which the reserved and unreserved portions of inventory are consumed is reversed in comparison to SN. Class H and L demand first compete equally for the unreserved portion $x-r$ of the inventory. After the first $x-r$ units of demand have arrived from both classes, additional class L demand is rejected and the remaining r units are exclusively dedicated to demand from class H. Figure 2 provides an illustration of TN in our specific setting.

Insert Figure 2 About Here

Again it is intuitive that when $r < x$, the class H (L) service level under TN will be higher (lower) than PA due to existing equal competition opportunity. Let $\delta_H^{TN}(r)$ denote the class H service level increment and $\delta_L^{TN}(r)$ the class L service level decrement over PA, induced by TN. Similar to our approach in Section 3.2, we can express the service levels of class H and L in TN by $\alpha_H^{TN}(r) = \alpha_H^{PA}(r) + \delta_H^{TN}(r)$ and $\alpha_L^{TN}(r) = \alpha_L^{PA}(r) - \delta_L^{TN}(r)$. In the following sections, we first develop expressions for the class H and L service levels, individually. Then, to follow our previous structure, we equivalently express the class H and L service levels based on the increment $\delta_H^{TN}(r)$ and the decrement $\delta_L^{TN}(r)$ induced by TN.

3.3.1 Class H service level under TN

In TN, we can identify two events sets that lead to fulfillment of the entire class H demand. Let $V(x-r) \equiv \inf\{t \geq 0 : D_L(t) + D_H(t) = x-r\}$ denote the random time at which the unreserved vaccine inventory $x-r$ is consumed by class H and L demand. We denote by $S^{TN}(r)$ the random time at which the entire inventory x is exhausted. Because demand from class L can only utilize the part of the unreserved vaccine inventory $x-r$ that is not consumed by the class H demand arriving in the first $x-r$ arrivals, we know that $S^{TN}(r) \equiv \inf\{t > V(x-r) : D_L(V(x-r)) + D_H(t) = x\}$. We can identify two mutually exclusive sets of events $\xi_{H,1}^{TN}$ and $\xi_{H,2}^{TN}$ that lead to fulfillment of the entire class H demand occurring during the vaccination period: $\xi_{H,1}^{TN} = \{\tau \leq V(x-r)\}$ and $\xi_{H,2}^{TN} = \{(V(x-r) < \tau) \wedge (D_H(\tau - V(x-r)) \leq r)\}$. Events set $\xi_{H,1}^{TN}$ captures those instances in which the entire class H demand can be fulfilled because the unreserved portion $x-r$ is sufficient to fulfill

the overall (class H and L) demand in the planning horizon. The entire class H demand will also be filled, if $x-r$ is not sufficient to fulfill all demand (i.e. $V(x-r)<\tau$) but the reserved quantity r is sufficient to meet any class H demand $D_H(\tau-V(x-r))\le r$ that arrives in the time interval $]V(x-r),\tau]$. The events are captured in set $\xi_{H,2}^{TN}$. Since $\xi_{H,1}^{TN}\cap\xi_{H,2}^{TN}=\varnothing$, sets are mutually exclusive and $\delta_H^{TN}(r)=\Pr\{\xi_{H,1}^{TN}\}+\Pr\{\xi_{H,2}^{TN}\}$ (details in Online Supplement OS.3). Consequently, we can express the class H service level in TN as:

$$\alpha_H^{TN}(r)=\sum_{n=0}^{x-r-1}f_p(n,(\lambda_L+\lambda_H)\tau)+\sum_{i=0}^{x-r}\sum_{j=0}^{r}\sum_{k=0}^{\infty}f_p(i+j,\lambda_H\tau)\,f_p(x-r-i+k,\lambda_L\tau)\,f_h\left(i;x-r+k+j,i+j,x-r\right)$$

Equivalently, the class H service level in TN can be expressed as an increment to the class H service level in PA (details in Online Supplement OS.4). Similar to SN, the class H service level in TN is also increasing in *r*.

$$\alpha_H^{TN}(r)=\alpha_H^{PA}(r)+\sum_{i=r+1}^{x}\sum_{j=0}^{x-i}p(j;\lambda_L,\tau)\,p(i;\lambda_H,\tau)+\sum_{i=r+1}^{x}\sum_{j=x-i+1}^{\infty}\sum_{k=i-r}^{x-r}p(j;\lambda_L,\tau)\,p(i;\lambda_H,\tau)h\left(k;i+j,i,x-r\right)$$

(16)

It can also be shown (details in Online Supplement OS.5) that TN and SN provide exactly same class H service levels if all class L demand arrive before the first class H demand arrival, the LBH assumption.

3.3.2 Class L service level under TN

To determine the class L service level under TN, we can identify two sets of events leading to fulfillment of the entire class L demand. It is straightforward that if $\tau\le V(x-r)$, not only class H demand but also the entire class L demand will be fulfilled. We denote the corresponding event set by $\xi_{L,1}^{TN}\left(=\xi_{H,1}^{TN}\right)$. If $V(x-r)<\tau$ the class L demand will only be completely fulfilled if $D_L(\tau-V(x-r))=0$. We denote the corresponding events set by $\xi_{L,2}^{TN}=\left\{\left(V(x-r)<\tau\right)\wedge\left(D_L(\tau-V(x-r))=0\right)\right\}$. For $\xi_{L,1}^{TN}$, we know from section 3.3.1 that: $\Pr\{\xi_{L,1}^{TN}\}=\sum_{n=0}^{x-r-1}f_p(n,(\lambda_L+\lambda_H)\tau)$. The expression for the class L service level in TN will be (details in Online Supplement OS.6):

$$\alpha_L^{TN}(r)=\sum_{n=0}^{x-r-1}f_p(n,(\lambda_L+\lambda_H)\tau)+\sum_{i=0}^{x-r}\sum_{j=0}^{\infty}f_p(x-r-i,\lambda_L\tau)\,f_p(i+j,\lambda_H\tau)\,f_h\left(i;x-r+j,i+j,x-r\right)$$

Equivalently, the class L service level in TN can be expressed as a decrement to the service level in PA (details in Online Supplement OS.7):

$$\alpha_L^{TN}(r) = \alpha_L^{PA}(r) - \sum_{i=x-r-j}^{\infty} \sum_{j=1}^{x-r} \sum_{k=0}^{j-1} f_p(j, \lambda_L \tau) f_p(i, \lambda_H \tau) f_h(k; i+j, j, x-r)$$

(17)

Similar to SN, the class L service level in TN is also decreasing in *r*.

4. NUMERICAL EXPERIMENTS

Having derived exact expressions for class H and L service levels for Partitioned Allocation, Standard Nesting, and Theft Nesting allocation mechanisms in the previous section we now provide numerical insights into the vaccine inventory reservation problem. In selecting the optimal reservation policy, the decision maker faces two interdependent problems: (1) selecting the allocation mechanism (i.e. PA, SN, or TN) and (2) determining the optimal *r* based on his individual preferences (i.e. aspired class H service level). With respect to the latter problem, our service level expressions derived in chapter 3 provide valuable support: given expressions (14) and (15) for SN as well as (16) and (17) for TN, the decision maker can calculate the tradeoff between the service levels of class H and class L.

Lemma 1 *Let* $\Delta_k^{PA}(r)$, $\Delta_k^{SN}(r)$, *and* $\Delta_k^{TN}(r)$ ($0 \le r < x$) *denote the class k* ($k \in \{L, H\}$) *increase in service level under PA, SN, and TN for increasing the reserved quantity from r to r+1.*

a) *i)* $\Delta_H^{PA}(r) > 0$ *ii)* $\Delta_L^{PA}(r) < 0$

b) *i)* $\Delta_H^{SN}(r) > 0$ *ii)* $\Delta_L^{SN}(r) < 0$

c) *i)* $\Delta_H^{TN}(r) > 0$ *ii)* $\Delta_L^{TN}(r) < 0$

Corollary Under PA, SN, and TN every reserved quantity *r* ($0 \le r < x$) is Pareto-efficient.

We observe from Lemma 1 and its corollary that under PA, SN, and TN an increase in the class H service level induced by reserving an additional unit of vaccine inventory for the exclusive use of class H will always lead to a decrease in the class L service level. Therefore, after selecting the preferred allocation mechanism, the decision maker's optimal *r* can only be determined based on his individual preferences regarding the tradeoff between the gain in class H and loss in class L service levels. Interestingly, in following the H1N1 flu vaccine procurement news during the 2009 outbreak, we observed three distinct ordering policies devised by the health planners among the studied countries: (1) Exact ordering when health officials ordered enough to vaccinate each resident of the country with one dose of vaccine. Belgium, for example, followed the exact ordering policy (GSK Press Release 2009). (2) Under-ordering when by under-estimating the impact of the outbreak or supply shortage due to late ordering, countries like Russia ordered an amount of vaccine doses inadequate to vaccinate the whole population (White 2009). (3) Over-ordering when in countries

such as the Netherlands with big influx of tourists and in-transit passengers, and also speculation about the severity of the outbreak and the need for multiple doses of vaccine, the vaccine order size was more than double the size of the population (Gray-Block 2010). In this section, we run numerical experiments to observe the impact of PA, SN, and TN allocation mechanisms when those ordering policies have been implemented. Ideally, we strive to identify the allocation mechanism that outperforms the other two by providing higher vaccination coverage to non-target groups (class L) for the aspired service level of target groups (class H) given the same available vaccine inventory. We assume the availability of 100 doses of vaccine (*x=100*) at the start of the vaccination season with no chance of replenishment during the season; and use the formulations derived in chapter 3 to calculate the service levels for multiple instances of class L and class H demand (λ_L, λ_H) rates. As mentioned before, the demand rate should be determined based on the severity of the new influenza strain and the public perception of vaccine efficacy. To facilitate the categorization of instances, we define the class H slackness $\theta_H = \frac{x}{\lambda_H}$ as the ratio of the available vaccine inventory to the class H demand rate; and the overall slackness $\theta = \frac{x}{\lambda_L + \lambda_H}$ as the ratio of the available vaccine inventory to the overall demand rate.

4.1 Exact-ordering

The population is assumed to be 100 and one dose of vaccine has been ordered and received before the vaccination season for every resident (*x=100*). At first, we envisage the realistic scenario that the target groups (class H) are in minority. We focus on four (λ_L, λ_H) instances of (90,10), (80,20), (70,30), and (60,40). For example, figure 3 shows the class L and class H service levels in PA, SN, and TN allocation mechanisms for every possible reserved vaccine quantity (from no reservation *r=0* to full reservation *r=100*) when $(x = 100, \lambda_L = 70, \lambda_H = 30)$. Although graphs represent discrete reserved vaccine quantities, for illustration purposes the points have been connected and shown as curves.

Insert Figure 3 About Here

At this stage, it would be difficult to draw any conclusion about which allocation mechanism should be used. However, if we show the same information but having the class H service levels on the x-axis and the class L service levels on the y-axis, the preferred allocation mechanism will reveal

itself. According to figure 4, TN allocation mechanism outperforms PA and SN by providing higher class L and class H service levels in all non-extreme reserved quantities. Any point on the graph represent a given reserved vaccine quantity for the exclusive use of class H. By definition at *r=0* (i.e. FCFS), TN and SN should provide identical results and at *r=100* (i.e. full dedication of vaccine inventory to target groups) all three allocation mechanisms must perform equally.

Insert Figure 4 About Here

If no vaccine is reserved and we allocate them on a FCFS basis, TN and SN at *r=0* result a class H service level of 61% and class L service level of 54% (class H service level in PA at *r=0* is close to zero and off-scale). One can observe that the positive impact of reservation on class H service level in TN is more pronounced than the PA and SN. By reserving 2 units of vaccine in TN, the class H service level is almost as high as reserving 31 units in SN and 33 units in PA. Using TN and at low reservations, the class H early arrivals benefit from a relatively large *x*-*r* unreserved vaccine pool in a fair competition; whereas the late class H arrivals can still be vaccinated since the last *r* vaccines are exclusively for their use without the fear of spoiling vaccines due to over-reservation and no-shows. If the aspired class H service level is 95%, it can be achieved using TN and reserving 8 vaccines. Repeating the same exercise with the other three instances and generating the class H and L service level graphs will provide the results reported in table 1 sorted on θ and then θ_H. In all $(\theta = 1)$ instances, TN is the preferred mechanism.

Insert Table 1 About Here

We also examined the unlikely instances that the target groups exceed or match half the population size $(\lambda_H \geq 50)$. For example, figure 5 illustrates the service levels in the $(\lambda_L = 10, \lambda_H = 90)$ instance.

Insert Figure 5 About Here

Similar to the previous example, we show the information by having the class H service levels on the x-axis and the class L service levels on the y-axis as in figure 6.

Insert Figure 6 About Here

According to figure 6, SN allocation mechanism outperforms PA and TN and consequently will be the preferred choice. TN is clearly over-protective of class H arrivals and causes vaccine spoilage that otherwise could have been used to vaccinate the non-target groups. PA performs as well as SN at high reserved quantities since the resulting small unreserved portion is guaranteed to be consumed by the non-target groups; and the target groups are large enough to consume the reserved portion completely. If, for example, the aspired class H service level is 95%, it cannot be achieved as even full reservation *r=100* provides only a class H service level of 86%. Table 2 summarizes the instances where the target groups are in majority and exact ordering policy is implemented.

Insert Table 2 About Here

If the target groups are as large as the non-target groups ($\lambda_L = \lambda_H = 50$), figure 7 shows that no allocation mechanism is preferred at all times and the decision maker selects the allocation mechanism according to his class H aspiration level (i.e. minimum pre-set required service level for target groups).

Insert Figure 7 About Here

If the desired service level for class H is less than 92%, TN will be the preferred allocation mechanism. Otherwise, SN performs best if higher class H service level has been planned, as shown in figure 8 with zoom in at class H service level above 90%. If, for example, the aspired class H service level is 95%, it can be achieved using SN and reserving 62 vaccines.

Insert Figure 8 About here

Although we observe the crossover from TN preference to SN preference at $(x=100, \lambda_L=50, \lambda_H=50)$, detailed analysis of (λ_L, λ_H) instances with $(40<\lambda_L<60)$ and $(\lambda_H=100-\lambda_L)$ will reveal the mechanism of this preference shift.

4.2 Under-ordering

In these instances, the health planners failed to secure enough doses of vaccine for the whole population. It can also apply to situations where the vaccine efficacy is low and repeat vaccinations are required. Figure 9 depicts one such instance where the demand for vaccines exceeds the supply by 20% $(x=100, \lambda_L=90, \lambda_H=30)$.

Insert Figure 9 About here

According to figure 10, PA outperforms the other two mechanisms in this instance. It can be explained by the fact that in such instances the demand is high enough to justify the partitioning of the available vaccine inventory between demand classes without expecting any spoilage of vaccines, thus eliminating the need for equal competition between class L and class H for the unreserved part of the vaccine inventory.

Insert Figure 10 About Here

If no vaccine is reserved and we allocate them on a FCFS basis, TN and SN at *r=0* result a class H service level of 7% and class L service level of 3.7% (class L service level in PA at *r=0* is 86% and off-scale). If the aspired class H service level is 95%, it can be achieved using PA and reserving 39 vaccines. The preferred allocation mechanism in such under-ordering $(\theta<1)$ instances is PA as shown in Table 3.

Insert Table 3 About Here

We have deliberately eliminated the extreme cases where the available vaccines are not even sufficient for vaccinating the target groups. Obviously in those cases the whole available vaccine inventory will be allocated to the class H according to PA (or equivalently to SN or TN with full reservation at $r = x$).

4.3 Over-ordering

In these instances, the health planners order more vaccines than the population. Reasons can be the expected need for multiple vaccinations per person or the influx of travelers in the flu season. Perceived low efficacy of the vaccine or low severity of the flu can also lead to more vaccine than the combined class L and H demand. Figure 11 represents the service levels at one such instance $(x = 100, \lambda_L = 10, \lambda_H = 80)$.

Insert Figure 11 About Here

As shown in figure 12, in this instance no allocation mechanism is preferred at all times and the decision maker should select the mechanism according to his class H aspiration level. For example, if the minimum required class H service level for target groups is greater than 92%, SN allocation mechanism is preferred. If, for example, the aspired class H service level is 95%, it can be achieved using SN and reserving 95 vaccines.

Insert Figure 12 About Here

When over-ordered $(\theta > 1)$ and the available vaccine inventory is much greater than the class H demand $(\theta_H > 2.5)$ TN is the preferred allocation mechanism regardless of the class H aspiration level, as shown in Table 4. In extreme over-ordering, the available vaccine inventory can be so high that even FCFS allocation (equivalent to SN or TN at *r=0*) would satisfy the decision maker's aspiration.

Insert Table 4 About here

5. CONCLUSIONS

In this research, we investigate the impact of alternative allocation mechanisms that can be employed in the context of vaccine inventory rationing. For partitioned, standard nesting, and theft nesting allocation mechanisms we develop service level expressions for low priority (non-target group) and high priority (target group) demand classes. Such formulations have, so far, not been available in the literature. We utilize the service level expressions to address the two interdependent problems of selecting the preferred allocation mechanism first and then determining the optimal reserved vaccine quantity that fulfills the aspired class H service level. We conduct numerical experiments on numerous instances with different levels of inventory slackness. From the results of our numerical experiments, we observe that there exist certain inventory slackness conditions under which one of the allocation mechanisms outperforms the others. This paper contributes to the existing literature in revenue management as well as inventory rationing. The dual-criteria inventory reservation model based on system service levels, and the random arrival order pattern had not been explored previously in the literature. Our closed-form expressions for the service levels of the two customer classes under each allocation mechanism will help determining the preferred inventory allocation mechanisms based on relevant system parameters. We have found evidence that health planners in different regions may exact-order, under-order, or over-order the vaccine inventory with respect to the actual demand for vaccines in their region. Our analysis shows that, depending on the ordering policy devised by the health planners, the optimal allocation mechanism differs; i.e. it is important for the decision maker to choose PA when under-ordered and TN when over-ordered. In the case of exact-ordering, if the target groups are in minority TN and when in majority SN is the preferred allocation mechanism. Although we limit our analysis to a single vaccination period single vaccine inventory reservation problem, we expect that our approach and our results can be utilized in a wide range of vaccine supply chains. We believe that this is an initial promising finding that should be considered in future research and potentially enables the development of efficient techniques for optimally solving this problem rather than using heuristics. Future research can take numerous directions. Despite we limit our analysis to two demand classes; the basic structure of the model can be extended to handle any arbitrary number of service-differentiated classes. The analysis for multiple demand classes under the nested allocation mechanism can become excessively complex. Therefore, the study of alternative nested and non-nested inventory allocation mechanisms will be an interesting future work. A more realistic compound Poisson demand distribution and demand fill rates as the performance measure could be

other viable alternatives. Throughout this paper, we assumed the reservation policies as static, i.e. the decision about the allocation mechanism and the reserved vaccine quantity can be taken only at the start of the vaccination period. Study of dynamic reservation policies that update the choice of the allocation mechanism and the corresponding reserved vaccine quantity will be valuable for both research and practice of healthcare supply chain management.

A.1 Proof of Lemma 1

a) Holds due to the properties of the Poisson cdf.

b) For any given $r_2 > r_1$ and after simplifications:

$$\alpha_H^{SN}(r_2) - \alpha_H^{SN}(r_1) = \sum_{i=r_1+1}^{r_2} \sum_{j=x-i+1}^{\infty} \sum_{k=0}^{i-1} f_p(i, \lambda_H \tau)\, f_p(j, \lambda_L \tau)\, f_h(k; j+i, i, x) > 0$$

Replacing r_2 with $r+1$ and r_1 with r leads to the result for $\Delta_H^{SN}(r)$.

$$\alpha_H^{SN}(r+1) - \alpha_H^{SN}(r) = \sum_{j=x-r}^{\infty} \sum_{k=0}^{r} f_p(r+1, \lambda_H \tau)\, f_p(j, \lambda_L \tau)\, f_h(k; j+r+1, r+1, x) > 0$$

For any given $r_2 > r_1$ and after simplifications:

$$\alpha_L^{SN}(r_2) - \alpha_L^{SN}(r_1) = -\sum_{i=0}^{x-j} \sum_{j=x-r_2+1}^{x-r_1} f_p(i, \lambda_H \tau)\, f_p(j, \lambda_L \tau)$$

$$- \sum_{i=x-j+1}^{\infty} \sum_{j=x-r_2+1}^{x-r_1} f_p(i, \lambda_H \tau)\, f_p(j, \lambda_L \tau)\, f_h(x-j; j+i, i, x) < 0$$

Replacing r_2 with $r+1$ and r_1 with r leads to the result for $\Delta_L^{SN}(r)$.

$$\alpha_L^{SN}(r+1) - \alpha_L^{SN}(r) = -\sum_{i=0}^{\infty} f_p(i, \lambda_H \tau)\, f_p(x-r, \lambda_L \tau)\, f_h(x-r; i+x-r, x-r, x-r+\min(i, r)) < 0$$

c) In the most general case, and for any given $r_2 > r_1$ after simplifications:

$$\alpha_H^{TN}(r_2) - \alpha_H^{TN}(r_1) = \sum_{i=r_1+1}^{x} \sum_{y=x-i+1}^{\infty} \sum_{w=0}^{i-r_1-1} f_p(i, \lambda_H \tau)\, f_p(y, \lambda_L \tau)\, f_h(w; y+i, i, x-r_1)$$

$$- \sum_{i=r_2+1}^{x} \sum_{y=x-i+1}^{\infty} \sum_{w=0}^{i-r_2-1} f_p(i, \lambda_H \tau)\, f_p(y, \lambda_L \tau)\, f_h(w; y+i, i, x-r_2)$$

Replacing r_2 with $r+1$ and r_1 with r leads to the result for $\Delta_H^{TN}(r)$.

$$\Delta_H^{TN}(r) = \sum_{i=r+1}^{x} \sum_{y=x-i+1}^{\infty} \sum_{w=0}^{i-r-1} p\left(i;\lambda_H,\tau\right) p\left(y;\lambda_L,\tau\right) h\left(w; y+i, i, x-r\right)$$

$$-\sum_{i=r+2}^{x} \sum_{y=x-i+1}^{\infty} \sum_{w=0}^{i-r-2} p\left(i;\lambda_H,\tau\right) p\left(y;\lambda_L,\tau\right) h\left(w; y+i, i, x-r-1\right) > 0$$

In the most general case, and for any given $r_2 > r_1$ after simplifications (see proof of section b):

$$\Delta_L^{TN}(r) = \alpha_L^{PA}(r_2) - \sum_{i=x-r_2-j}^{\infty} \sum_{j=1}^{x-r_2} \sum_{k=0}^{j-1} f_p(j,\lambda_L\tau) f_p(i,\lambda_H\tau) f_h\left(k; i+j, j, x-r_2\right)$$

$$-\alpha_L^{PA}(r_1) + \sum_{i=x-r_1-j}^{\infty} \sum_{j=1}^{x-r_1} \sum_{k=0}^{j-1} f_p(j,\lambda_L\tau) f_p(i,\lambda_H\tau) f_h\left(k; i+j, j, x-r_1\right) < 0$$

REFERENCES

Arslan. H., S. C. Graves, T. Roemer. 2007. A single-product inventory model for multiple demand classes. *Management Science* **53**(9) 1486-1500.

Benjaafar, S. and M. El Hafsi. 2006. Production and Inventory Control of a Single Product Assemble-to Order System with Multiple Customer Classes. *Management Science* **52** 1896-1912.

Brumelle, S. L., D. Walczak. 2003. Dynamic airline revenue management with multiple semi-Markov demand. *Operations Research* **51**(1) 137-148.

de Vericourt, F., F. Karaesmen, Y. Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Science* **48**(11) 1486-1501.

Deshpande, V., M. A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated classes, *Management Science* **49**(6) 683-703.

Evans, R. V. 1968. Sales and restocking policies in a single item inventory system, *Management Science* **14**(7) 463-472.

Gray-Block, A. 2010. Ministry says GSK reduces order by 3 million vaccines. *Reuters*. 12 May 2010. Website at http://www.reuters.com/article/2010/05/12/dutch-h1n-idUSLDE64B25M20100512

GSK Press Release. 2009. GlaxoSmithKline Update: A H1N1 influenza vaccine update. *Press release of 15 May 2009*. Website at http://www.gsk.com/media/pressreleases/2009/2009_pressrelease_10054.htm

Ha, A. Y. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science* **43**(8) 1093–1103.

Ha, A. Y. 1997b. Stock rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics* **44** 457–472.

HIDA. 2009. Season 2008-2009 Influenza vaccine production and distribution. *Health Industry Distributors Association (HIDA) annual report*.

Kaplan, A.1969. Stock Rationing. *Management Science* **15**(5) 260-267.

Lee, T.C. and M. Hersh. 1993. A Model for Dynamic Airline Seat Inventory Control with Multiple Seat Bookings. *Transportation Science* **27** 252-265.

McGill, J., G. van Ryzin. 1999. Revenue Management: Research Overview and Prospects. *Transportation Science* **33** 233-256.

Melchiors, P., R. Dekker, M. J. Kleijn. 2000. Inventory rationing in an (s, Q) inventory model with lost sales and two demand classes. *Journal of the Operational Research Society* **51**(1) 111-122.

Nahmias, S., W. S. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Science* **27**(11) 1236-1245.

Robinson, L. W. 1995. Optimal and approximate control policies for airline booking with sequential non-monotonic fare classes. Operations Research **43** 252-263.

Samii, A. B., R. Pibernik, P. Yadav. 2011. An inventory reservation problem with nesting and fill rate-based performance measures. *International Journal of Production Economics*. Forthcoming.

Talluri, K. and G. van Ryzin. 2004. Revenue Management under a general discrete choice model of customer behavior. *Management Science* **50** 15-33.

Talluri, K., G. van Ryzin. 2005. The theory and practice of revenue management. Springer, ISBN 0-387-24376-3

Topkis, D. M. 1968. Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes, *Management Science* **15** 160-176.

van Hoek A. J., A. Underwood, M. Jit, E. Miller, W. J. Edmunds. 2011. The Impact of Pandemic Influenza H1N1 on Health-Related Quality of Life: A Prospective Population-Based Study. *PLoS ONE* **6**(3): e17030. doi:10.1371/journal.pone.0017030

Veinott, A. F. 1965. Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. *Operations Research* **13**(5) 761-778.

White, G. L. 2009. Doctor says Russia understating swine-flu cases. *The Wall Street Journal*, 23 September 2009 Page A10. Website at http://online.wsj.com/article/SB125366412674432421.html

Williams, S., P. Yadav. 2005. A supply chain fix for flu. *MIT Center for Transportation and Logistics Supply Chain Frontiers* **12.**

FIGURE 1: CLASS H DEMAND FULFILLMENT IN STANDARD NESTING

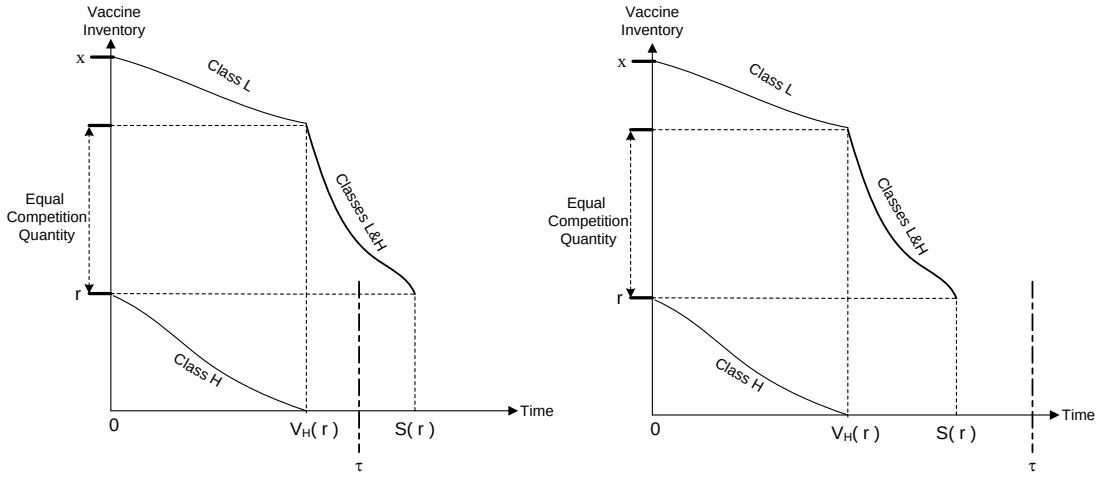

FIGURE 2: DEMAND FULFILLMENT TIME STRUCTURE IN TN

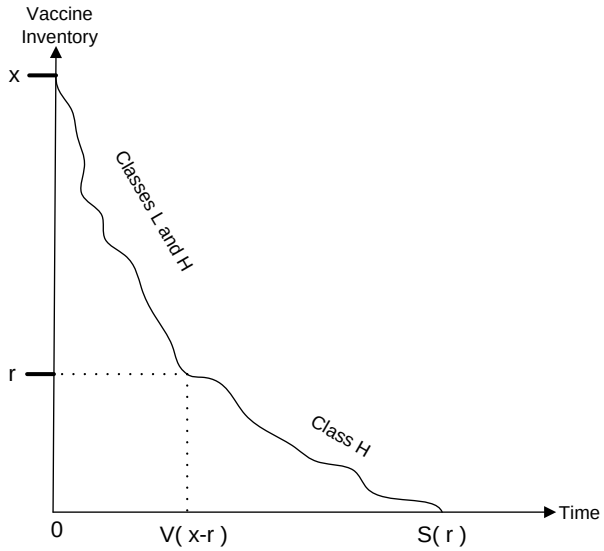

FIGURE 3: CLASS H AND L SERVICE LEVELS FOR $(x=100, \lambda_L=70, \lambda_H=30)$

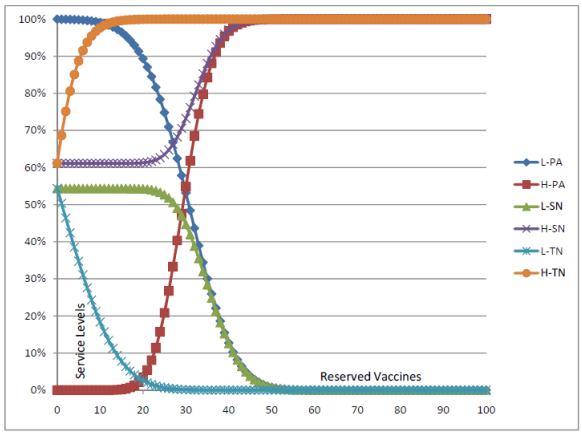

FIGURE 4: PERFORMANCE OF ALLOCATION MECHANISMS FOR

$(x = 100, \lambda_L = 70, \lambda_H = 30)$

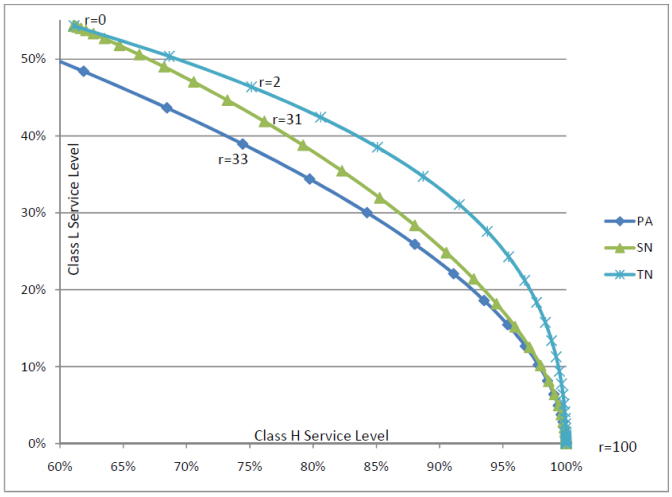

TABLE 1 TN-PREFERENCE IN EXACT-ORDERING AND MINORITY TARGET GROUPS

λ_L	λ_H	x	θ_H	θ	Preference
60	40	100	2,50	1,00	TN
70	30	100	3,33	1,00	TN
80	20	100	5,00	1,00	TN
90	10	100	10,00	1,00	TN

FIGURE 5: CLASS H AND L SERVICE LEVELS FOR $(x=100, \lambda_L=10, \lambda_H=90)$

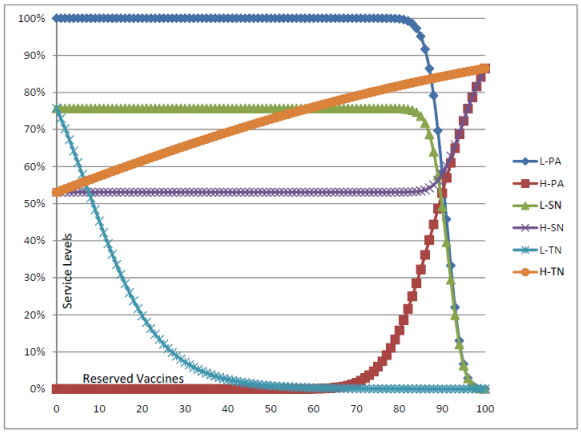

FIGURE 6: PERFORMANCE OF ALLOCATION MECHANISMS FOR

$(x = 100, \lambda_L = 10, \lambda_H = 90)$

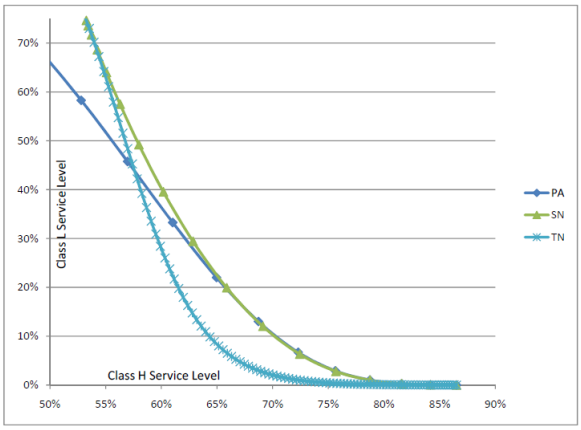

TABLE 2 SN-PREFERENCE IN EXACT-ORDERING AND MAJORITY TARGET GROUPS

λ_L	λ_H	x	θ_H	θ	Preference
10	90	100	1,11	1,00	SN
20	80	100	1,25	1,00	SN
30	70	100	1,43	1,00	SN
40	60	100	1,67	1,00	SN

FIGURE 7: PERFORMANCE OF ALLOCATION MECHANISMS FOR

$(x = 100, \lambda_L = 50, \lambda_H = 50)$

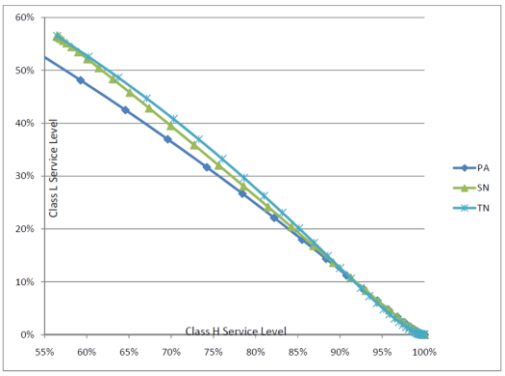

FIGURE 8: PERFORMANCE AT HIGH VALUES OF CLASS H SERVICE LEVEL FOR $(x = 100, \lambda_L = 50, \lambda_H = 50)$

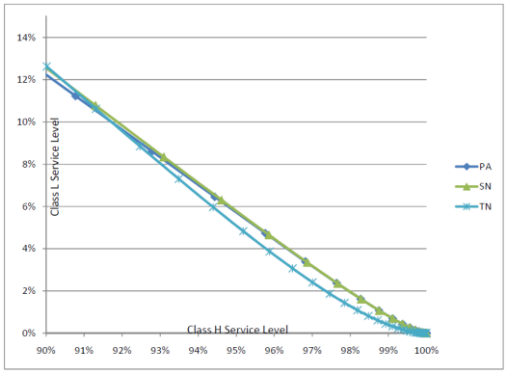

FIGURE 9: CLASS H AND L SERVICE LEVELS FOR $(x=100, \lambda_L=90, \lambda_H=30)$

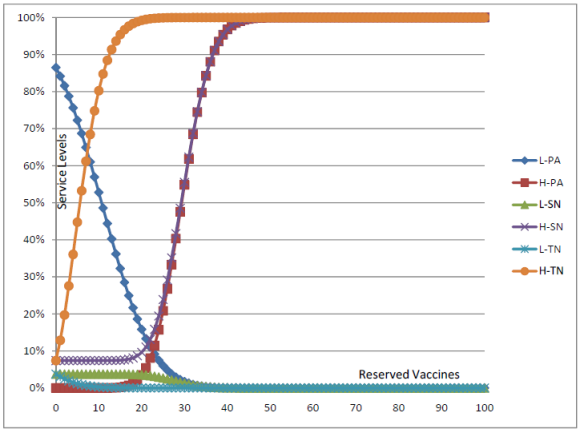

FIGURE 10: PERFORMANCE OF ALLOCATION MECHANISMS FOR $(x = 100, \lambda_L = 90, \lambda_H = 30)$

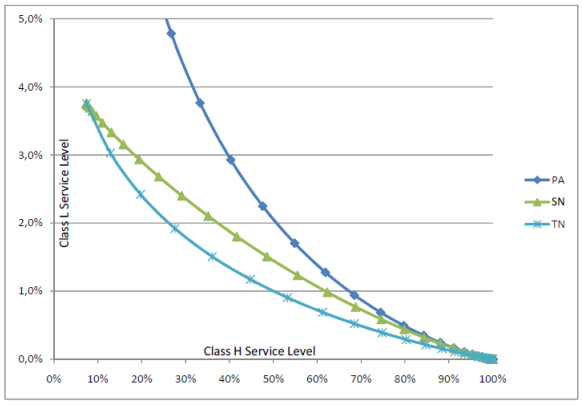

TABLE 3 PA-PREFERENCE IN VACCINE UNDER-ORDERING

λ_L	λ_H	x	θ_H	θ	Preference
30	90	100	1,11	0,83	PA
90	30	100	3,33	0,83	PA
20	90	100	1,11	0,91	PA

FIGURE 11: CLASS H AND L SERVICE LEVELS FOR $(x=100, \lambda_L=10, \lambda_H=80)$

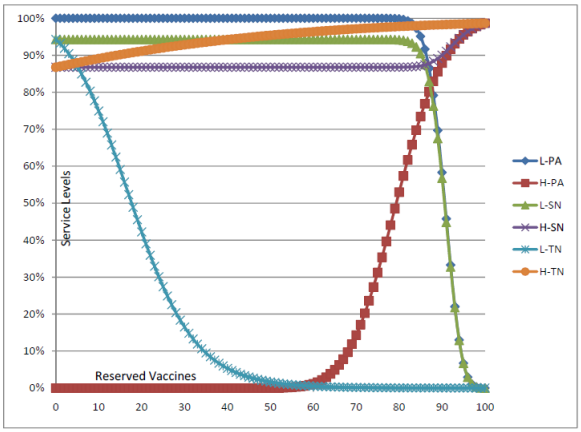

FIGURE 12: PERFORMANCE OF ALLOCATION MECHANISMS FOR

$(x = 100, \lambda_L = 10, \lambda_H = 80)$

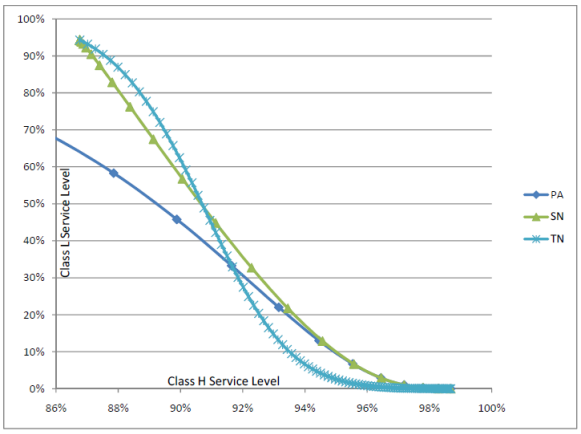

TABLE 4 ALLOCATION MECHANISM PREFERENCE BASED ON CLASS H ASPIRATION LEVEL IN OVER-ORDERING

λ_L	λ_H	x	θ_H	θ	**Preference**
10	80	100	1,25	1,11	SN, for class H Service Level>92%
70	20	100	5,00	1,11	TN
10	70	100	1,43	1,25	SN, for class H Service Level>99.7%