Abstract

Traditionally healthcare systems have paid little attention to the management of inventories. However, with the implementation of diagnostic related groups by the United States government (which resulted in a pre-fixed level of compensation for specific medical services), these systems have turned their attention to cost containment as a means of increased profitability. This research addresses the issue of managing inventory costs in a healthcare setting.

The specific problem addressed in this paper is a comparison of inventory costs and service levels of an in-house three-echelon distribution network vs. an outsourced two-echelon distribution network. In comparing inventory policies in both networks, we focus on non-critical inventory items. Based on our analysis, we find that the recent trend of outsourcing to distribute non-critical medical supplies directly to the hospital departments using them (i.e., the two-echelon network) results not only in inventory cost savings but also does not compromise the quality of care as reflected in service levels.

Keywords: Health services; Inventory; Supply chain management

1. Introduction

Consider the following scenario:

A major healthcare provider operated 7 hospitals within the state of Florida in the United States with approximately 20 patient departments within each hospital. Non-critical inventory items
were received by the central warehouse owned and operated by the provider and were then distributed to each hospital warehouse. In turn, these items were supplied by the hospital warehouses to departments in each hospital. Each hospital in the network distributed stock only to departments within the particular hospital and thus, there was no stock transshipment between hospitals.

A partnership with a major distributor of medical supplies led to a restructuring of the distribution network which consolidated the central and hospital warehousing functions into a single Service Center. The center would be operated by the distributor and would directly supply items to individual departments within each hospital. Although this restructuring effort could in the long-run reduce the supplier base, a key issue of immediate concern in this conversion was whether the healthcare provider realized cost savings and if so, did these savings come at the expense of lower customer service?

This scenario is reflective of a general trend of outsourcing specific organizational activities to “expert” third-party providers. In general, the decision to outsource any organizational activity is justified by ensuring that either: (a) the outsourcing agency can provide the product or service more efficiently than an internal agency/department while maintaining the same base level of the quality of the product or service; (b) the quality of the product or service provided by the outsourcer will be higher than the quality of the same product or service delivered by the internal agency/department while maintaining the same level of efficiency in providing the product or service (Lunn, 2000); and/or (c) the supplier base would in the long-run be reduced leading to a reduction in procurement costs which could be passed on downstream to the end users. Obviously, if both efficiency and quality increase (in the short-run) and procurement costs are lower (in the long-run) due to outsourcing, then the organization tends to gain substantially by choosing to outsource immediately.

In healthcare, outsourcing of inventory decisions is indicative of current practices (Veral and Rosen, 2001) and is primarily driven by three factors. First, inventory investments in this industry are substantial and are estimated to be between 10% and 18% of net revenues (Holmgren and Wentz, 1982; Jarett, 1998). For example, for the quarter ending March 31, 2001, IASIS Healthcare generated net revenues of $232,619,000 and its inventory investment was $23,354,000 (10.04% of net revenues) while University Health Services generated net revenues of $561,790,000 and its inventory investment was $89,709,000 (15.97% of net revenues). Thus, any cost savings which can be generated through a more efficient management of inventories can lead to direct increases in profitability. Second, as reported by Li and Benton (1996), healthcare providers focus on quality of service both from an internal and external perspective. It has been argued that the move towards outsourcing of inventory can lead to improved internal performance as assessed through service levels (Jarett, 1998). Further, improving internal service levels also positively impacts patient care and this, in turn, should be lead to increases in the external measures of customer satisfaction and customer perceptions of service quality. Third and finally, there has been an increase in the expertise and in the number of third-party providers which offer inventory management services in healthcare. These providers (which include large management consulting firms with healthcare specific services) have over a short-period of time accumulated substantive anecdotal evidence of their successes in stockless inventory management through outsourcing (see, for example, Rivard-Royer et al., 2002). These “success” stories are also another key driver of the current trend of outsourcing inventory management in healthcare settings.

One missing aspect from the accumulated anecdotal evidence in healthcare is whether the conversion of an in-house three-echelon distribution network to an outsourced two-echelon distribution network actually results in cost savings and the possible impact on quality (assessed in terms of service levels for departments within hospitals). This forms the basis of our investigation in comparing inventory costs and service levels
across two distribution networks managing non-critical supplies. The specific focus of this research is two-fold.

- To develop normative tools/methodologies through which we can examine the non-critical inventory item decisions within each type of network. This would facilitate a comparison of total non-critical item related inventory costs across both networks under a range of scenarios.
- Using these models and data for one healthcare provider, we were interested in estimating the potential savings (if any) and related impact on service levels that would result by switching from the three-echelon network to a two-echelon network.

The remainder of this paper is organized as follows. In the next section, we review the relevant literature on inventory management in healthcare and on multi-echelon inventory systems. The general models developed to analyze inventory policies under each scenario are described next followed by a discussion of how each of these models was operationalized in this research. This is followed by a comparison of the networks in terms of service levels and total inventory costs, and finally, we conclude the paper with implications and conclusions of our research.

2. Healthcare inventory management: A literature review

The management and distribution of inventory of all kinds among and within hospitals is usually discussed under the broad heading of material management. One of the distinct features of material management in a hospital is the use of a periodic review par level (or order-up to level) servicing approach. For the three-echelon distribution network, this approach would be implemented as follows. The inventory position at each department is reviewed at the beginning of a review interval and the difference between the department’s par level and inventory in hand is ordered from the hospital. Orders from each department are transmitted to the hospital instantaneously and they are executed as received. After fulfilling these orders, each hospital, in turn, reviews the difference between the par level and inventory in hand and places an order with the central warehouse. Once again, these orders are transmitted instantaneously and executed as received. Finally, the central warehouse also orders from external suppliers, the difference between the par level and inventory in hand, and suppliers replenish central warehouse stocks by the end of the review period. The only difference in this approach for the two-echelon network (Scenario B) is that the hospital is bypassed in the process (i.e., departments communicate directly with the service center). One key aspect of such a policy is the setting of the review interval and practical considerations lead to this interval being identical across echelons.

A major issue in setting par levels for various items in a healthcare setting is that these levels tend to reflect the desired inventory levels of the patient caregivers rather than the actual inventory levels needed in a department over a certain period (see Prashant, 1991). In most cases these par levels are experience-based and politically driven, rather than data-driven. This poses a problem for warehouse managers since the...
inventory they hold is typically based on aggregate hospital demands while requirements of departments when aggregated are not in line with such estimates. The literature analyzing the setting of optimal par levels and review periods for multiple echelons draws upon prior work in multi-echelon inventory systems which is discussed next.

Allen (1958) was one of the first researchers to analyze multi-echelon distribution systems. He considered the problem of determining an optimal redistribution of the stock among $K$ locations. There have been a number of extensions to Allen’s model (see Simpson, 1959; Krishnan and Rao, 1965; Das, 1975; Hoadley and Heyman, 1977). Clark and Scarf’s (1963) study represents the first attempt to formulate and characterize an optimal policy in a multi-period, multi-echelon, inventory/distribution model that involves uncertain demand. Additional work by Schwarz (1981) and others (e.g., Deuermeyer and Schwarz, 1981; Eppen and Schrage, 1981; Nahmias and Smith, 1994) have produced a large set of models that generally seek to identify optimal lot sizes and safety stocks in a multi-echelon framework.

A number of researchers have analyzed an arborescent distribution system with no stock at the warehouse. Thus, the warehouse acts as a “break-bulk” facility by ordering goods in bulk and upon receipt, breaking these quantities into smaller units to ship to retailers (see Silver et al., 1998). The question of whether or not a warehouse should keep stock extends to one of centralization vs. decentralization of stocks. More complexity is involved in the analysis when the warehouse can hold stock, since the number of decision variables increases. These variables include the amount the warehouse should order from its supplier, the amount to ship from the warehouse to the retailers each period, and the amount to allocate to each retailer in case of shortage.

The most relevant modeling studies in the context of this research are those of Sinha and Matta (1991) and Rogers and Tsubakitani (1991). Both studies analyzed a two-echelon inventory system under stochastic demand with fixed lead times and a periodic review par level service system as described above. In Rogers and Tsubakitani (1991), the focus is on finding the optimal par levels for the lower echelons to minimize penalty costs subject to the maximum inventory investment across all lower echelons being constrained by a budgeted value. They show that the optimal par levels are determined by a critical ratio (for the newsboy model) adjusted by the Lagrange multiplier related to the budget constraint. Sinha and Matta (1991) analyze a multi-product system where they focus on minimizing holding costs at both echelon levels plus penalty costs at the lower echelon level. Their results are that the par levels at the lower echelon level is determined by the critical ratio while the par level for the upper echelon is determined by a search of the holding cost function at that level.

In order to adapt these modeling approaches to our setting, we needed to make several significant changes as follows:

- To model the three-echelon network (Scenario A), we extend the two-echelon models of Sinha and Matta (1991) and Rogers and Tsubakitani (1991). Essentially, when we model Scenario A, our model differs from those proposed in these studies since it incorporates costs and par levels for an extra echelon.
- In modeling Scenario B (the two echelon network), we also needed to incorporate the fact that over and above the backorder cost, the service center was also going to charge departments an additional penalty cost if the backorder exceeded a certain fixed amount. This led us to incorporate a set of additional 0–1 decision variables in our model to capture such effects. Consequently, our modeling effort for this scenario while relying on the basic models of Rogers and Tsubakitani (1991) and Sinha and Matta (1991) also differs significantly in the context of these binary variables to capture the realities of the healthcare system being modeled.
- Finally, given the par level servicing policy described earlier, note that orders for each department within a hospital are fulfilled through hospital stocks prior to supplies being received by the hospital from the central warehouse. Similarly, the central warehouse fulfills all hospital orders prior to supplies being received from external distributors. This led us to incorporate the fact that each of the higher echelons
needed to set par levels such that the inventories were at least as large as the expected aggregate orders which will be placed by all the echelons directly below. This is another critical difference in our modeling effort as compared to prior work.

In the next section, we present the details of the two models developed, one for each network scenario.

3. Model development

3.1. Assumptions

Our investigation of the two healthcare networks assumes that each is represented as an arborescent multi-echelon distribution system. Essentially, each system is represented as one where a lower echelon receives items from only one source at a higher echelon. In line with current practices, under both scenarios no transshipments among departments or among hospitals are allowed. Based on the par level ordering policy, we also assume a periodic review par level inventory management system for each echelon. Each model attempts to capture the key realities underlying the scenarios being investigated. However, there are also certain assumptions we make in order to facilitate the modeling effort. These assumptions and the underlying rationale behind each is discussed next.

- We focus on developing single non-critical item inventory models for each network (see Footnote 2 for estimated investment levels and examples of non-critical items). Although each department within a hospital does stock a large number of these types of items, we observed that in general, demand for individual items was independent of the demand for other items (discussed below). If this assumption does not hold in other situations, the models could still apply provided items within a department could be grouped such that within each group item demands are highly correlated while between groups, item demand correlation is low. In such a case, the inventory decisions for each group could be separately analyzed.
- Demand for the single item is assumed to be independent across departments within a hospital. The assumption of independent demands at the department levels is reasonable since a large number of identical non-critical items are stocked within each department in a hospital. The required stocks of these items are primarily used to satisfy department specific patient demands which are usually independent of one another. Although in exceptional situations, our assumption may not hold (e.g., in times of certain epidemics, the same item demand may be higher at all locations), we feel it is reasonable in general.
- The relevant costs incorporated in our models focus on the traditional inventory holding and backorder costs at each echelon. Thus, we explicitly assume no item price differentials across echelon scenarios. In practice, this may not be the case, since the outsourcer may also be in a position to obtain smaller prices given its procurement volume. Further, the outsourcer could, in the long-run, reduce the total supplier base and fixed cost savings due to such a reduction could also be passed on to the end users. However, to carry out a fair and unbiased comparison of the two alternative networks, we felt that it was important to remove this external effect.
- Given that we focus primarily on inventory related costs, there is another issue that needs to be addressed. In the three-echelon network (Scenario A), the healthcare system incurred inventory costs at

---

Footnote 2: The primary difference between management of inventories for critical vs. non-critical items was that, as would be expected, the healthcare providers set substantially higher service levels for these items (e.g., a 99% service level would be required for a critical item as compared to a 90% service levels for a non-critical item). A secondary difference was that transshipments of critical supplies between departments within a hospital occurred quite frequently while for non-critical supplies, this was not the case.
all levels while in the two-echelon network, the outsourcer was responsible for costs incurred at the service center and the inventory costs for the departments would be the responsibility of the healthcare system. Thus, when comparing the costs across networks, the healthcare system would probably incur lower total costs. However, this ignores the fact that the healthcare system would need to also reimburse the service provider (under Scenario B) and such a reimbursement would be based on the service center inventory costs. Thus, when developing models for each network scenario, we include all inventory related costs at all echelons.

- In line with current practice in the healthcare industry, the models we develop are single review period models, and we assume that all echelons have the same review period (Veral and Rosen, 2001).
- We also assume that replenishment lead-times (including transportation times) are zero (or small enough in relation to the review period). The routine nature of delivery, from supplier to warehouse, from warehouse to hospitals, and from hospitals to departments makes this a valid assumption.
- Consistent with current practices in healthcare, we assume that all backorders are satisfied by emergency deliveries at a cost higher than normal delivery cost. We also assume that outside suppliers hold sufficient stock to satisfy all demand from the warehouse (service center).
- In setting par levels for each department, we follow a policy typically established in the healthcare setting. Essentially, this policy requires each department to meet a minimum service level (in terms of the minimum fraction of demand which should be satisfied). Thus, department service levels (and in turn par levels) are established such that a pre-specified minimum service level is maintained for each department.

Based on these assumptions, we now proceed to describe each model developed to compare the two network structures. A comprehensive list of notation used in presenting both models is given in Table 1.

3.2. Models

Based on the assumptions stated earlier, each healthcare network is analyzed using a single-item, single review period inventory model. These models are developed based on the assumption that the underlying review period demand distribution for each department is known. In general, we assume that if \( f(\cdot) \) represents the pdf of review period demand then, \( f(\cdot) = 0 \) if \( \cdot < 0 \). Let \( \mu_{ij} \) represent the average review period demand with associated standard deviation \( \sigma_{ij} \) for department \( j \) in hospital \( i \). The models for each scenario are presented next.

3.2.1. Model for Scenario A

In this model, the decision variables are the par levels for each department \( (S_{ij}^{DA}) \), for each hospital \( (S^h) \) and the central warehouse \( (S^w) \). In determining the department par levels, we also ensure that a minimum pre-specified service level \( \delta_{ij} \) (which represents the proportion of demand which is satisfied through inventory) for each department \( j \) in hospital \( i \) is maintained. Thus, the par level \( S_{ij}^{DA} \) for each department \( j \) in hospital \( i \) is determined such that it’s associated service level \( (\beta_{ij}^{DA}) \) is at least as large as the minimum pre-specified service level \( \delta_{ij} \). Based on this, the model is presented below.

\[
\text{Minimize } EC^A(S^w, S^h, S_{ij}^{DA}) = H^w \left[ \int_0^{S^w} (S^w - x^w) f(x^w) \, dx^w \right] + \left\{ \sum_{i=1}^{N} H^h_i \left[ \int_0^{S^h_i} (S^h_i - x_i) f(x_i) \, dx_i \right] \right\}
\]

\[
+ \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} H_{ij}^{DA} \left[ \int_0^{S_{ij}^{DA}} (S_{ij}^{DA} - x_{ij}) f(x_{ij}) \, dx_{ij} \right] \right\}
\]

\[
+ \sum_{i=1}^{N} P_i^A \left[ \sum_{j=1}^{n_i} \int_{S_{ij}^{DA}}^{\infty} (x_{ij} - S_{ij}^{DA}) f(x_{ij}) \, dx_{ij} \right] \quad (1)
\]
subject to:

$$\int_{0}^{S_{ij}^{DA}} (x_{ij} - S_{ij}^{DA}) f(x_{ij}) \, dx_{ij} - (1 - \beta_{ij}^{A}) \mu_{ij} \leq 0 \quad \forall i \text{ and } \forall j,$$

$$\beta_{ij}^{A} \geq \delta_{ij} \quad \forall i \text{ and } \forall j,$$

$$S_{i}^{h} = \sum_{j=1}^{n_{i}} \left[ S_{ij}^{DA} - \int_{0}^{S_{ij}^{DA}} (S_{ij}^{DA} - x_{ij}) f(x_{ij}) \, dx_{ij} \right] \geq 0 \quad \forall i,$$

$$S_{i}^{w} - \sum_{i=1}^{N} \left[ S_{i}^{h} - \int_{0}^{S_{i}^{h}} (S_{i}^{h} - x_{i}) f(x_{i}) \, dx_{i} \right] \geq 0,$$

$$S_{w}, S_{i}^{h}, S_{ij}^{DA} \geq 0 \quad \forall i \text{ and } \forall j.$$
The objective function sums up the holding costs based on expected inventory levels at all three echelons (central warehouse, hospital warehouses, and departments within each hospital) and also assesses a penalty cost for expected backorders for departments (see Eq. (1)). Constraint set (2) is the relationship between the expected demand satisfied from inventory for department \( j \) in hospital \( i \) based on service level \( \beta_{ij}^A \). Obviously \( \beta_{ij}^A \) is constrained to be at least as large as \( \delta_{ij} \) which is the pre-specified minimum service level that should be maintained for the department (see constraint set (3)). Constraints (4) and (5) capture the realities of the three-echelon network for the healthcare system studied in this paper. Recall that orders are communicated by the departments at the beginning of the review interval and are executed as received by the hospital warehouse, and subsequently orders are communicated by hospital warehouses at the beginning of the review interval and executed as received by the central warehouse. Based on this, each hospital warehouse needs to maintain adequate inventories (as determined by the par levels \( S_{ij}^h \)) so as to satisfy the expected orders for all departments it supplies and the central warehouse needs to maintain inventories (reflected in the par level \( S^c \)) in order to satisfy the expected orders for all hospital warehouses. This feature is incorporated in this set of constraints. Finally, the non-negativity constraints on the decision variables are incorporated in constraint sets (6).

### 3.2.2. Model for Scenario B

In this model, the decision variables are the par levels for each department (\( S_{ij}^{dB} \)), and the service center (\( S^c \)). An additional set of decision variables for handling a context specific issue are also included in modeling the problem under this scenario. The outsourcer operating the service center negotiated with the healthcare provider that if total emergency deliveries to a hospital (reflected in total backorders) exceeded a constant \( a \) units it would charge a higher penalty cost per unit of \( B^B \) while backorders within this limit would incur a penalty cost of \( P^B \). In the model described below, the decision variables \( y_i \) enable us to incorporate this issue. As with the previous models, the department par levels are determined to ensure that a minimum pre-specified service level \( \delta_{ij} \) (which represents the proportion of demand which is satisfied through inventory) for each department \( j \) in hospital \( i \) is maintained. Thus, the par level \( S_{ij}^{dB} \) for each department \( j \) in hospital \( i \) is determined such that its associated service level \( \beta_{ij}^{dB} \) is at least as large as the minimum pre-specified service level \( \delta_{ij} \). Based on this, the model is presented below.

Minimize \( EC^{dB}(S^c, S_{ij}^{dB}, y_i) = H^c \left[ \int_0^{S^c} (S^c - x^c)f(x^c) \, dx^c \right] + \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} H_{ij}^{dB} \left[ \int_0^{S_{ij}^{dB}} (S_{ij}^{dB} - x_{ij})f(x_{ij}) \, dx_{ij} \right] \right\} + \sum_{i=1}^{N} (1 - y_i) \left\{ (P_i^B z_i) + P^B \left[ \sum_{j=1}^{n_i} \int_{S_{ij}^{dB}}^{\infty} (x_{ij} - S_{ij}^{dB})f(x_{ij}) \, dx_{ij} - z_i \right] \right\} + \sum_{i=1}^{N} y_i P^B \sum_{j=1}^{n_i} \int_{S_{ij}^{dB}}^{\infty} (x_{ij} - S_{ij}^{dB})f(x_{ij}) \, dx_{ij} \right\} \) subject to:

\[
\int_{S_{ij}^{dB}}^{\infty} (x_{ij} - S_{ij}^{dB})f(x_{ij}) \, dx_{ij} - (1 - \beta_{ij}^{dB}) S_{ij}^{dB} \leq 0 \quad \forall i \text{ and } \forall j,
\]

\[
\beta_{ij}^{dB} \geq \delta_{ij} \quad \forall i \text{ and } \forall j,
\]

---

4 If we do not include these constraints in our model, then we can obtain an optimal solution to our problem quite easily as follows. Use a newsboy type analysis at each department to determine the service levels \( \beta_{ij}^{dB} \) and associated par level \( S_{ij}^{dB} \). Based on these par levels, determine the appropriate hospital warehouse par levels \( S_{ij}^h \) and finally use these to determine the central warehouse par level \( S^c \).
The objective function sums up the holding costs based on expected inventory levels at two echelons (service center and departments within each hospital) and also assesses a penalty cost for expected backorders for all departments within each hospital (see Eq. (7)). If \( y_i = 0 \), then total backorders for all departments within a hospital exceed \( a_i \), and hence the penalty costs \( P_B_i \) (to the first \( a_i \) units) and \( P \) (for the excess units) are charged while if \( y_i = 1 \), then only the penalty costs \( P_B_i \) are charged per unit. As with the earlier model, constraint set (8) is the relationship between the expected demand satisfied from inventory for department \( j \) in hospital \( i \) based on service level \( B_{ij} \). Obviously \( B_{ij} \) is constrained to be at least as large as \( d_{ij} \) which is the pre-specified minimum service level that should be maintained for the department (see constraint set (9)). Constraint (10) captures the fact that service center inventories must be adequate to satisfy the expected orders for all departments in all hospitals. Constraint sets (11) and (12) force \( y_i \) to take on the appropriate value of 0 or 1 and \( M \) is a sufficiently “large” constant. Finally, the binary and non-negativity constraints on the decision variables are incorporated in constraint sets (13) and (14).

The models presented in this section apply to any general distribution of review period demand. In the next section, we describe: (a) how each of these models was operationalized based on empirical data for the healthcare provider; and (b) procedures developed in order to obtain “good” solutions for each model.

4. Model operationalization and solution methods

The models developed in the previous section are general in the sense that they do not assume any specific underlying demand distribution. After collecting demand data for several non-critical routinely used items in the healthcare network, we first determined that the normal distribution could be used as a close approximation to represent the underlying demand process (done through a Q–Q plot analysis). Although such a distribution does have a positive probability associated with negative review period demand, such a possibility can be minimized if the observed coefficient of demand is relatively small. In our case, the largest observed value for this coefficient was 0.40 and this effectively eliminated the practical possibility of a negative review period demand (essentially this translates to a probability of negative review period demand of 0.0062 or 0.62%). Further, the normal distribution has been used extensively in prior research on inventory models to model the underlying demand process (Silver et al., 1998). The exact specification of both models assuming a normal distribution of review period demand is described next.

Under the assumption of independent review period demand for each department in a hospital, we first observe that the following relationships hold:
1. Average review period demand at each hospital \(i (\mu^h_i)\) is the sum of total average review period demands within each department in the hospital. Thus:

\[
\mu^h_i = \sum_{j=1}^{n_i} \mu^d_{ij}.
\]

2. The standard deviation of average review period demand for a hospital \(i (\sigma^h_i)\) is

\[
\sigma^h_i = \sqrt{\sum_{j=1}^{n_i} \sigma^d_{ij}}.
\]

3. The average review period demand at the central warehouse \((\mu^w)\) is the sum of total average review period demands for all hospitals. Thus:

\[
\mu^w = \sum_{i=1}^{N} \mu^h_i.
\]

4. The standard deviation of average review period demand for the central warehouse \((\sigma^w)\) is

\[
\sigma^w = \sqrt{\sum_{i=1}^{N} \sigma^h_i}.
\]

5. For the two echelon scenario, the service center average review period demand \((\mu^c)\) and the associated standard deviation \((\sigma^c)\) is as follows:

\[
\mu^c = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mu^d_{ij},
\]

\[
\sigma^c = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sigma^d_{ij}}.
\]

Based on this, if we assume that the review period demand for each department is normally distributed with parameters \(\mu^d_{ij}\) and \(\sigma^d_{ij}\), then we also know that: (a) aggregate review period demand for each hospital \(i\) is normally distributed with parameters \(\mu^h_i\) and \(\sigma^h_i\); (b) aggregate review period demand for the central warehouse is normally distributed with parameters \(\mu^w\) and \(\sigma^w\); and (c) aggregate review period demand for the service center is normally distributed with parameters \(\mu^c\) and \(\sigma^c\).

Then following the analysis for the newsboy model under normally distributed demand (Silver et al., 1998), we note that

\[
\int_0^{S^w} (S^w - x^w)f(x^w) \, dx^w = \sigma^w[Z^w + G(Z^w)],
\]

\[
\int_0^{S^h_i} (S^h_i - x_i)f(x_i) \, dx_i = \sigma^h_i[Z^h_i + G(Z^h_i)],
\]

\[
\int_0^{S^d_{ij}} (S^d_{ij} - x_{ij})f(x_{ij}) \, dx_{ij} = \sigma^d_{ij}[Z^d_{ij} + G(Z^d_{ij})],
\]
\[ \int_{S^d_{ij}}^{\infty} (x_{ij} - S^d_{ij}) f(x_{ij}) \, dx_{ij} = \sigma^d_{ij} G(Z^d_{ij}), \]
\[ \int_{0}^{S^h} (S^h - x^h) f(x^h) \, dx^h = \sigma^h [Z^h + G(Z^h)], \]
\[ \int_{0}^{S^w_{ij}} (S^w_{ij} - x_{ij}) f(x_{ij}) \, dx_{ij} = \sigma^w_{ij} [Z^w_{ij} + G(Z^w_{ij})], \]
\[ \int_{S^d_{ij}}^{\infty} (x_{ij} - S^d_{ij}) f(x_{ij}) \, dx_{ij} = \sigma^d_{ij} G(Z^d_{ij}), \]
\[ \mu^A_{ij} = F(Z^A_{ij}), \]
\[ \mu^B_{ij} = F(Z^B_{ij}), \]

where \( F(\cdot) \) represents the cdf of review period demand and \( G(\cdot) \) is the unit normal loss function. The two models developed in the prior section are operationalized by substituting these expressions. Obviously, the decision variables for these revised models are the standard normal deviates: (a) \( Z^d_{ij}, Z^h_{ij}, \) and \( Z^w_{ij} \) for Scenario A; and (b) \( Z^dB_{ij}, \) and \( Z^c \) for Scenario B. Once estimates of these are obtained, then the par levels can be set as follows:

\[ S^d_{ij} = \mu^d_{ij} + Z^d_{ij} \sigma^d_{ij}, \]
\[ S^dB_{ij} = \mu^dB_{ij} + Z^dB_{ij} \sigma^dB_{ij}, \]
\[ S^h_{ij} = \mu^h_{ij} + Z^h_{ij} \sigma^h_{ij}, \]
\[ S^w = \mu^w + Z^w \sigma^w, \]
\[ S^c = \mu^c + Z^c \sigma^c. \]

In both of these operationalizations, we are faced with the challenge of minimizing an objective function (convex in the decision variables) over a non-convex constraint set. Thus, we need to solve a reverse convex programming problem which has been shown to be NP-hard (Horst and Padalos, 1995). Further, there are no effective procedures that have been developed for specifying lower bounds to such programming problems but upper bounding heuristic methods can of course, be developed. A first approach to obtaining such an upper bound was through the use of LINGO where we “solved” each model separately with the same parameter settings. The term “solve” is being used loosely here since we do not obtain optimal solutions to our models. Rather, we identified a local optimum by running LINGO with a limit on run time. A second approach implemented exploits the newsboy type structure of both models. In this case, we developed greedy heuristics to obtain feasible solutions to each model and these are described in the next section.

### 4.1. Heuristic 1 (for Scenario A)

The motivation for this heuristic stems from the fact that based on a newsboy type analysis, the service level for a department could be estimated to be the maximum of the ratio of penalty to holding cost or the prespecified minimum service level \( \delta_{ij} \). Based on these estimated departmental service levels for each department, we can determine feasible solutions to the problem.
hospital, we could determine the appropriate service levels for each hospital which would satisfy constraint (4). Finally, moving up to the highest echelon level (central warehouse), we could impute the service levels for the central warehouse based on these estimated hospital service levels to satisfy constraint (5). A formal statement of this greedy procedure is outlined below and it also contains details on how par levels are set based on estimated service levels for each echelon.

1. For each department $j$ in hospital $i$, let
   
   $$ \beta^A_{ij} = \max \left\{ \frac{\rho^A_{ij}}{\sigma^A_{ij}}, \delta_{ij} \right\}. $$

   Based on this, determine $Z^A_{ij}$ such that $F(Z^A_{ij}) = \beta^A_{ij}$.

   Set the par level $S^A_{ij} = \mu^A_{ij} + Z^A_{ij}\sigma^A_{ij}$.

2. Determine for each hospital $i$:
   
   $$ Z^h_i = \frac{1}{\sigma^h_i} \left\{ \left[ \sum_{j=1}^{n_i} (\mu^h_{ij} - \sigma^h_i G(Z^A_{ij})) \right] - \mu^h_i \right\}. $$

   Set the par level for hospital $i$ as $S^A_{hi} = \mu^h_i + Z^h_i\sigma^h_i$.

3. Determine for the central warehouse:
   
   $$ Z^w = \frac{1}{\sigma^w} \left\{ \left[ \sum_{i=1}^{N} (\mu^w_i - \sigma^w_i G(Z^h_i)) \right] - \mu^w \right\}. $$

   Set the par level for the central warehouse $S^w = \mu^w + Z^w\sigma^w$.

4. Determine the expected total cost based on these par levels.

4.2. Heuristic 2 (for Scenario B)

The motivation for developing this heuristic is similar to that described above. However, note that in this case, if we set department service levels to be the maximum of the ratio of penalty to holding cost or the prespecified minimum service level ($\delta_{ij}$), we also need to be concerned with the resulting expected total backorders which may be greater than the allowable limit $\alpha$. The reason for this is that any expected backorders greater than $\alpha$ incur a higher penalty cost $P_B$. Thus, we modify the departmental service levels to incorporate this higher additional penalty cost. Once we determine these departmental service levels, then in line with the prior heuristic, we use these to impute the service center service levels to satisfy constraint (10). A formal statement of this greedy procedure is outlined below and, as with the earlier heuristic, it also contains details on how par levels are set based on these estimated service levels for each echelon.

1. For each department $j$ in hospital $i$, let $CR^B_{ij} = \frac{\rho^A_{ij}}{\delta_{ij}}$.

   Based on this, determine $Q^B_{ij}$ such that $F(Q^B_{ij}) = CR^B_{ij}$.

   Then, determine the expected backorder $E^B_{ij} = \sigma^B_{ij} G(Q^B_{ij})$ where $G(Q^B_{ij})$ is the unit loss normal function.

2. For each hospital $i$: if $\sum_{j=1}^{n_i} E^B_{ij} \leq \alpha_i$ set $\beta^B_{ij} = \max \{ CR^B_{ij}, \delta_{ij} \}$;

   else determine $\overline{P}$, as

   $$ \overline{P} = p^B \left( \frac{\sum_{j=1}^{n_i} E^B_{ij} - \alpha_i}{\sum_{j=1}^{n_i} E^B_{ij}} \right) + p^B \left[ 1 - \left( \frac{\sum_{j=1}^{n_i} E^B_{ij} - \alpha_i}{\sum_{j=1}^{n_i} E^B_{ij}} \right) \right] $$

   and set

   $$ \beta^B_{ij} = \max \left\{ \frac{\overline{P}}{(\overline{P} + H^B_{ij})}, \delta_{ij} \right\}. $$
Determine the par levels for each department \( j \) in hospital \( i \) as follows. First, determine \( Z_{ij}^B \) such that 
\[
F(Z_{ij}^B) = \beta_{ij}^B.
\]
Set the par level \( S_{ij}^{dB} = \mu_{ij} + Z_{ij}^B \sigma_{ij}^d \).
4. Determine for the service center:

\[
Z^c = \frac{1}{\sigma^c} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} (\mu^d_{ij} - \sigma^d_{ij} G(Z^d_{ij})) \right\}.
\]

Set the par level for the central warehouse \( S^c = \mu^c + Z^c \sigma^c \).

5. Determine the expected total cost based on these par levels.

Both heuristic procedures were implemented in Microsoft Excel using the built-in function for generating the CDF for the normal distribution. A comparison of the solutions obtained using LINGO with those obtained using the heuristics are shown in Tables 2 and 3. The percentage differences reflect the averages from various combinations of inventory holding and penalty costs. For example, a coefficient of variation of 40% was examined under different holding cost ratios (e.g. \( H^h_{ij}/H^d_{ij} = 0.57, 0.63, 0.70 \) for the model for Scenario A and \( H^c_{ij}/H^d_{ij} = 0.57, 0.63, 0.70 \) for the model for Scenario B). Although the percentage differences reflect averages, it is clear that the solutions obtained by LINGO dominate those obtained using the heuristics. Further, we also observed that in all individual cases the LINGO solutions dominated those obtained using the heuristic methods. Thus, all the remaining discussion in this paper is based on the use of LINGO for obtaining solutions to both models.

5. Results

5.1. Experimental details

In order to carry out the comparison of the in-house and outsourced networks, we collected/obtained data on review period (a review period being defined as a week) for multiple items experienced by a single department of one hospital in the health-care system being analyzed. Based on this information and in consultation with the relevant system personnel, we set the review period demand parameters as follows. First, for a department in a hospital, we set the average demand to be an average of the actual demand for all items. Then, we adjusted this average demand by their estimate of the average demand at each of the other departments. In this manner, we generated the average demand for all departments in a single hospital. For departments in another hospital in the network, we adjusted the average demand for each unit by a constant fraction (which again was estimated based on input from the network personnel). An example is given below to clarify this process of setting average demand for each department in each hospital in a given distribution network.

Let us arbitrarily assume that we want to specify average single item demand for two hospitals each with 10 departments and assume that the average actual demand for the data obtained was 5 units. Based on this, the average demand for department 1 in hospital 1 is set to 5. Average demand for each of the other departments in hospital 1, is adjusted upwards/downwards by a constant. Thus, for example, departments 2–4 will face an average demand of 4 (i.e., 80% of 5); departments 5–7 will face an average demand of 6 (i.e., 120% of 5); while departments 8–10 will face an average demand of 7 (i.e., 140% of 5). Note that the percentage constants (i.e., 80%, 120% and 140%) are based on inputs from the hospital personnel. Next, in order to specify the average demands for the 10 departments in hospital 2, we simply multiply the demands for each corresponding department in hospital 1 by a constant. If we assume a constant of 90% (individual constants for each hospital are estimated through discussions with the health-care network personnel), average demand for each department in hospital 2, is simply 90% of the demand of the corresponding department in hospital 1 (i.e., department 1 in hospital 2 will face a demand of 4.5 (90% of 5); departments 2–4 in hospital 2 will face a demand of 3.6 (90% of 4); and so on).

The second parameter of interest related to average demand is the standard deviation/variance of demands for individual items in each department in each hospital. Although, we could have estimated this
parameter for our experiments in the same manner as described for the average demand, we found that this parameter was not really different across departments and hospitals but it varied more across items. Thus, we choose to vary this parameter in an experimental manner and the range of values explored is based on the variances in multiple item demand data that was collected/obtained from the healthcare network. Given the average demand for an item described above, we varied the standard deviation of demand such that the coefficient of variation levels were set at 0.40, 0.35, and 0.30. Once demand parameters for departments in each hospital was set, we now describe how the cost and network size parameters were set in order to facilitate an analysis and evaluation of the network structures being compared. One of the key issues to note in our experimentation is that unit holding costs were set to be equal across all departments which led to equal unit holding costs for each hospital and equal unit backorder costs for each hospital. Table 4 below summarizes the settings used for our experiments and a discussion and rationale for each setting follow.

As is obvious from the Table 4, all cost parameters are either an explicit or implicit function of the holding cost per unit for a department in a hospital ($H_{ij}^{A}$ and $H_{ij}^{B}$). Hence, to obtain the parameter settings shown in the table above, we first arbitrarily set the value of $H_{ij}^{A}$ (and equal to $H_{ij}^{B}$) to be a constant for all departments in all hospitals. Then:

1. Based on the coefficient of variation setting to be investigated, fix the standard deviation of demand for each department in each hospital (see Table 4 for different settings investigated for this parameter).
2. Fix the unit holding cost for each department in each hospital. For:
   - the three-echelon network, fix the unit holding cost for each hospital ($H_i^h$) and the warehouse ($H^w$) depending upon the parameter setting to be investigated (see Table 4); and
   - the two-echelon network, fix the unit holding cost for the service center ($H^e$) depending upon the parameter setting to be investigated (see Table 4).
3. Set the unit backorder cost for each hospital ($P_A^i = P_B^i$) depending upon the parameter setting to be investigated (see Table 4).
4. Based on discussions with the outsourcer, the emergency unit delivery cost for Scenario B ($P_B^i$) is set to be 120% of $P_B^i$.
5. Run the experiment using LINGO and collect information on the relevant performance measures.

For each experiment we carried out, we collected information on the service levels for each department in each hospital, service levels for each hospital (for the three-echelon distribution network—Scenario A),

<table>
<thead>
<tr>
<th>Parameter sets</th>
<th>Ratios</th>
<th>Parameter levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holding costs</td>
<td>$H_i^h/H_{ij}^{da}$</td>
<td>0.70, 0.63, 0.50</td>
</tr>
<tr>
<td></td>
<td>$H^w/H_i^h$</td>
<td>0.70, 0.63, 0.57</td>
</tr>
<tr>
<td></td>
<td>$H^e/H_{ij}^{db}$</td>
<td>0.70, 0.63, 0.57, 0.50, 0.44</td>
</tr>
<tr>
<td>Penalty costs</td>
<td>$P_A^i/H_{ij}^{da}$</td>
<td>15.30, 23.00, 46.00</td>
</tr>
<tr>
<td></td>
<td>$P_B^i/H_{ij}^{db}$</td>
<td>15.30, 23.00, 46.00</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>$\sigma_{ij}/\mu_{ij}$</td>
<td>0.30, 0.35, 0.40</td>
</tr>
<tr>
<td>Number of departments</td>
<td>$n_i \forall i$</td>
<td>10, 15, 20</td>
</tr>
<tr>
<td>Number of hospitals</td>
<td>$N$</td>
<td>3, 5, 10</td>
</tr>
</tbody>
</table>
service level for the warehouse/service center, and total costs. Before discussing the results in detail, three input parameters were set as follows:

- The minimum required service level for a department set at 90% (this was indicative of current practice in the healthcare setting for non-critical inventory items).
- The unit holding cost for each department in each hospital set at $10 (i.e., $H^d_{ij} = H^d_{ij} = $10 ∀i and ∀j). Although we did investigate different starting values for the unit holding cost for each department in each hospital, we found that this did not change the resulting service levels within departments, within hospitals, and at the warehouse/service center. However, total costs were obviously greater when we increased the unit holding cost for each department in each hospital.
- When running the experiments for Scenario B, we set the value of $a_i = 15$ units ∀i. Thus, if expected backorders for all departments in a hospital were greater than 15 units, the emergency delivery penalty cost $P^B$ was charged.

We now proceed to separately discuss the results in terms of service levels, total costs and a comparison of Scenarios A and B.

5.2. Results for service levels

For both scenarios, the average service levels for departments are quite comparable and changes in these levels are primarily due to changes in the ratio of penalty costs to department holding costs. As would be expected, when the ratio of penalty cost to departmental holding costs increases, average department service levels also increase. More specifically, when $P^A_{ij}/H^d_{ij} = P^B_{ij}/H^d_{ij} = 15.3$, average service levels for departments are approximately 93.75%, when $P^A_{ij}/H^d_{ij} = P^B_{ij}/H^d_{ij} = 23$, average service levels for departments are approximately 95.77%, and when $P^A_{ij}/H^d_{ij} = P^B_{ij}/H^d_{ij} = 46$, average service levels for departments are approximately 97.86%. Average service levels for hospitals are, of course, only relevant for Scenario A. In this case, these service levels increase as the ratio of penalty costs to holding costs increases and these are in the range 45.19–49.00%.

The interesting differences in service levels that are noticeable across scenarios are for the warehouse as compared to the service center. In general, the service center service levels are always higher than those of the warehouse (warehouse service levels are in the range 7.20–23.81% while service center service levels are in the range 35.22–48.28%). Thus, if we compare scenarios, this indicates that greater inventories are held at the service center (Scenario B) as compared to the warehouse (Scenario A). Further, these service levels are impacted by the ratio of Penalty to departmental holding costs and the size of the network (in terms of hospitals and departments for Scenario A and the number of departments for Scenario B). As earlier, warehouse and service center service levels increase when the ratio of penalty costs to average department holding costs increases. Further, in line with prior research on multi-echelon inventory models, higher service levels for the warehouse and service center are achieved when the size of the network decreases since the possibility of backorders decreases upstream in a smaller network as compared to a larger network.

5.3. Results for total costs

After the discussion related to service levels, two results for total costs are obvious:

- For both scenarios, as the ratio of penalty costs to holding costs increases, there is an increase in aggregate total costs. These results are a direct function of the fact that this ratio tends to increases service levels (and hence, inventories) which leads to an increase in total costs.
In line with the traditional single item newsboy analysis, increases in the coefficient of variation of demand leads to increases in inventory levels at each echelon and this, in turn, increases aggregate costs.

In terms of network size, the results are graphically represented in Figs. 1 and 2. The primary reason for this increase in costs due to an increase in network size stems from the fact that when increasing the number of departments in a hospital or the number of hospitals, there is an increased demand that each system must accommodate. Thus, if we compare a network of three hospitals with 10 departments vs. a network of three hospitals with 15 departments, the latter system experiences higher demand due to a larger number of departments. Although we could have compared costs by assuming the same demand spread over different size networks, this was not carried out for two reasons. First, note that our focus is primarily on comparing the impact of the two generic network structures (the three-echelon vs. the two-echelon) rather than the impact of demand being spread over fewer or greater number of departments. Second, and more importantly, we found that in practice, demand was typically greater in larger hospital networks and hence, our experiments reflect this scenario.

5.4. Comparing Scenario A vs. Scenario B

Given that the prior results are in line with expectations and thus, provide face validity for our modeling effort, we now address the issue of comparing the two alternative network structures. Our analysis in this context attempts to: (1) evaluate the cost savings associated with switching from an in-house three echelon network to an outsourced two-echelon network; and (2) compare the service levels for each department in these two scenarios. In carrying out such a comparison, we first had to equalize the echelon holding costs across scenarios. Two specific holding cost cases where the holding costs across echelons were approximately equal are:

![Graph 1: Scenario A—effects of network size.](image)

![Graph 2: Scenario B—effects of network size.](image)
Case 1. \(H^h_{ij}/H^d_{ij} = 0.70\) and \(H^w/H^h = 0.70\) for Scenario A vs. \(H^c/H^d_{ij} = 0.50\) for Scenario B; and

Case 2. \(H^h_{ij}/H^d_{ij} = 0.70\) and \(H^w/H^h = 0.63\) for Scenario A vs. \(H^c/H^d_{ij} = 0.44\) for Scenario B.

The results described below are based on these two cases.

For both these scenarios, Scenario B always dominates Scenario A in terms of total costs for any fixed network size, a fixed coefficient of variation of demand, and a fixed penalty cost to department holding cost ratio. In terms of magnitude of savings, these are obviously dependent on the specific holding costs, penalty costs and emergency delivery costs for the system and a summary of cost savings across all experimental parameters are shown in Table 5. As can be observed from Table 5, average savings across both cases are approximately $200 with a range of $70–500 (with percentage savings ranging between 2% and 3%). Given that our analysis is for a single inventory item, obviously these savings would be significantly larger for several non-critical items which are typically stocked. Further, note that here we are comparing cost savings in terms of total costs for both healthcare networks. If the service center is actually operated by an outsourcer, then this entity would be responsible for the costs of the service center inventories at that stage. Thus, the advantages to the current healthcare system would be even larger and could be used to estimate how much the system would be willing to “pay” to contract out inventory operations to the third-party and/or in negotiations with third-party providers.

The surprising result that also stems from our analysis is that related to the department service levels since Scenario B provides equivalent average department service levels for departments as compared to Scenario A. Further, the service center service levels for Scenario B are always higher than the warehouse service levels and lower than the hospital service levels for Scenario A. This points to the fact that the service center holds smaller inventories than the total inventories held by the central warehouse and hospitals. Analytically from a multi-echelon inventory management perspective, the result that the inventory costs of the two echelon system are lower than that of the three echelon system is in line with expectations since the former system eliminates an echelon level (and related inventories) in the analysis. However, what is unusual is that the highest echelon service levels are greater for the two echelon network as compared to the three echelon network. Thus, this points to the fact that although end-user (i.e., department) service levels are equivalent between the two scenarios, we are also able to provide higher customer service at the highest echelon level.

6 Lower total costs due to an outsourced hybrid stockless system are also reported in a case study by Rivard-Royer et al. (2002). In addition to these cost reductions, they also found that outsourcing resulted in a significant reduction in full-time equivalent labor hours required by the healthcare provider.
This concludes our analysis of the two alternative networks. As we have seen, the results indicate that our modeling approaches have strong face validity. Further, these approaches can also be used successfully to compare alternative networks for managing inventories.

6. Conclusions and recommendations

Outsourcing the distribution of non-critical items in the health industry is always a viable alternative. The network structure of Scenario B is a typical mode of outsourcing, where an outside agent manages the receiving, inventory holding, and distribution components of the distribution system. Based on observation and reports from administrative healthcare network personnel, this arrangement resulted in freeing up the hospitals’ facilities and staff. The implication for this on a broader scale should allow those in the health sector to concentrate more on their core responsibilities of being health providers.

Scenario B is more or less an implementation of a centralized distribution system (a pull system, since the lowest echelon dictates distribution of materials). However, in order for such a centralized system to be fully effective, there needs to be a well-coordinated information system to support the distribution of items. The problems encountered in getting data on the distribution of basic items for this research suggests that there is little coordination among the various facilities. Our models are also appropriate for other settings where par level servicing policies are used to manage inventory investments (e.g., retailing). However, care should be taken in translating our cost savings in terms of other settings since these are a function of the parameter settings used in our paper.

Both models have some shortcomings, which should help to form a platform for future research agendas in the area of distribution systems within the healthcare industry. One of these shortcomings is the assumption of zero lead-times between echelons. While there are many situations, such as the one that is the focus of this research that can fit within this assumption; there are many other situations that deviate somewhat from this assumption. Thus, one extension of this research could be the inclusion of positive lead-time between echelons, especially between the outside supplier and the warehouse/service center. This is not as straightforward as it may seem on the surface, as the structure of the model might have to be modified to accommodate this. A second issue is to extend these models to incorporate the fact that single item demand may be correlated across departments within a hospital or departments across hospitals. This would require the formulation of more complicated cost functions reflecting joint probability density functions of item demands as well as constraints where the service levels are also jointly determined. A third extension of our models would reflect multi-item demands and hence, looking at joint replenishment policies for these items.

Two other more direct extensions of the model could be (i) to investigate the trade-off of having a predetermined minimum system-wide service level compared to a predetermined minimum service level at the department level; and (ii) to investigate the trade-offs of basing the desired target inventory for the warehouse (or service center) on incremental echelon cost, as against installation cost. The first case would be a complex undertaking since one would have to be concerned with the joint probabilities among the echelons. This complexity would be increased as the number of hospitals ($N$) and the number of departments within each hospital ($n_i$) increase. The second case would require fundamental changes in the formulation to ensure that the order-up-to-level at each upper echelon becomes a function of the echelon stock rather than the installation stock. In addition, the expressions for ending inventory are designed for installation stock (Diks et al., 1996). In examining a 2-echelon inventory system in a capacitated production-inventory system, Rappold and Muckstadt (1998) showed that the desired target inventory for each distribution center based on installation costs will be less than or equal the target inventory level based on echelon costs.
Acknowledgements

This research was partially supported by a Summer Research Grant awarded by the Warrington College of Business Administration to Asoo J. Vakharia.

References