

LOGISTICAL MODELS FOR PLANNING AND OPERATING MEDICAL COUNTERMEASURE DISTRIBUTION NETWORKS DURING PUBLIC HEALTH EMERGENCIES

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Kathleen Allison King

August 2012

© 2012 Kathleen Allison King
ALL RIGHTS RESERVED

LOGISTICAL MODELS FOR PLANNING AND OPERATING MEDICAL
COUNTERMEASURE DISTRIBUTION NETWORKS DURING PUBLIC
HEALTH EMERGENCIES

Kathleen Allison King, Ph.D.

Cornell University 2012

Public health emergencies require rapid responses from federal, state, and local authorities to prevent widespread mortality and morbidity. However, existing response plans seldom account for the variety of risks and uncertainties inherent in emergency scenarios. Our goal is to construct models that will help policy makers respond effectively to two different potential emergencies: an inhalational anthrax bioterrorist attack and an influenza pandemic. We present a three-echelon capacitated distribution network model of the United States' antibiotic mass-dispensing system for responding to a large-scale anthrax attack. We construct two inventory allocation policies and present a numerical study that compares their performance to that of planned allocation methods. We also present detailed simulation models of an antibiotic-dispensing clinic and the multi-echelon supply chain that operate to support such clinics. Along with the results of our earlier numerical study, these simulations can be used to demonstrate the importance of flexible clinic staffing plans, show the value of centralized command and control during emergency response operations, and provide other public health policy insights. Finally, we investigate the value of using the commercial pharmaceutical supply chain to dispense antiviral medication during an influenza pandemic. We construct historically-based regional antiviral demand scenarios, simulate the performance of the supply chain, and

describe inventory allocation and staffing models that could be used to improve system operations.

BIOGRAPHICAL SKETCH

Kathleen King grew up in St. Charles, Illinois with her parents, Karen and Ralph King, and sister, Margaret King. She graduated from the Illinois Mathematics and Science Academy in Aurora, IL in 2001. Kathleen then spent a year as an Olin Partner at the Franklin W. Olin College of Engineering, where she worked with faculty and 29 other students to develop the new college's curriculum. She earned a Bachelor of Science degree in Engineering as part of the inaugural class of Olin College in 2006. She began her graduate career at Cornell University in 2006, earning a Master of Science degree in Operations Research in 2009 and a Doctor of Philosophy degree in Operations Research in 2012. Her research at Cornell was supported by a Department of Energy Computational Science Graduate Fellowship. While at Cornell, Kathleen was also fortunate to meet and marry Steve Raciti, a fellow Cornell PhD.

This dissertation is dedicated to my parents, Karen and Ralph King, for their unwavering love and support. I am so grateful for all of your encouragement. Thank you both for helping me to keep my sense of humor and perspective.

ACKNOWLEDGEMENTS

First, I want to thank my advisor, Jack Muckstadt, for his excellent teaching, thoughtful guidance, and generous advice over the years. His ability to constantly generate great ideas and look at problems in creative ways is an inspiration to me. I am grateful for all of the time he has spent helping me learn how to think and understand how the world works. Professor Muckstadt has also been generous in sharing his knowledge and enjoyment of German Riesling and French Bordeaux wines.

I feel fortunate to have had the opportunity to work with Professor Nathaniel Hupert, who has been my guide in understanding the world of public health. I strive to emulate his thoughtful perspective and ability to draw connections between engineering and public health. I am also grateful to Professor Kathryn Caggiano for her guidance on all matters from technical topics to resume writing and job hunting. Kathryn is very insightful and has been generous with her time and advice over the years. I want to extend special thanks to Professors Shane Henderson, Peter Jackson, Huseyin Topaloglu, and David Williamson for their helpful ideas, questions, and guidance of my research.

I am grateful to my colleagues Lisa Koonin, Martin Meltzer, and Anita Patel of the Centers for Disease Control and Prevention for expanding my knowledge of a vital set of public health concerns and problems. They are impressive individuals who are able to look at public health problems from many different perspectives. I am very happy to know that the health of the United States is in such good hands. I want to thank them for allowing me to contribute to their work on potential alternative antiviral distribution strategies for responding to future influenza pandemics.

I have also been fortunate to collaborate with several incredibly bright and hard-working Cornell undergraduate and Master of Engineering students: Michelle Castorena, Caitlin Hawkins, and Cindie Wu, who worked on D-PODS; Kenneth Chu, who worked on ESCOE; and especially Christine Barnett, with whom I have worked for over two years on ESCOE and the antiviral modeling project. Christine is thoughtful, dedicated, and creative; I am excited to see what she will do in her future research at the University of Michigan.

I want to thank the Department of Energy Computational Science Graduate Fellowship (DOE CSGF) and the Krell Institute for supporting my research over the last four years. I am grateful both for the financial support and for introducing me to a community of interesting and intelligent people in the world of computational science. I hope to find opportunities to collaborate with other CSGF Fellows throughout my career.

Finally, I want to acknowledge my husband, Steve Raciti, for his love and encouragement. I am grateful to have a partner who is brilliant and thoughtful and who has the wonderful ability to balance work and the rest of life. Steve has been amazingly supportive of my work and eternally willing to discuss my research trials, tribulations, and triumphs with me. He is always there to remind me that research is supposed to be hard; that's what makes it interesting and satisfying.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	vii
List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Emergency Preparedness Literature Review	9
2 Anthrax Response Network Model	19
2.1 Model Description and Notation	20
2.1.1 Constraints on Inventory Allocation Decisions	33
2.1.2 Costs	36
2.1.3 Dynamic Programming Formulation	41
2.2 Inventory Allocation Strategies	42
2.2.1 Wait-and-See Solution	43
2.2.2 Expected Value Solution	44
2.2.3 Truncated Cumulative Approximation	46
2.2.4 Lagrangian Relaxation Method	53
2.3 Decentralized Allocation Methods	92
2.3.1 Fair Share Method	92
2.3.2 Independent Ordering Method	94
2.4 Computational Results and Discussion	97
2.4.1 Cost Comparison of Allocation Methods	101
2.4.2 Policy Implications	110
2.5 Future Work	116
3 Mass Prophylaxis Simulation Models	121
3.1 Dynamic Point of Dispensing Simulator	122
3.1.1 Model Description	124
3.1.2 Model Interface	125
3.1.3 Staffing Policy Implications	127
3.1.4 Discussion	141
3.2 Emergency Supply Chain Operations Evaluator	142
3.2.1 Model Formulation	144
3.2.2 Example Scenario: Modeling an Inhalational Anthrax At- tack	149
3.2.3 Discussion	156

4	Antiviral Dispensing Models For an Influenza Pandemic	159
4.1	Constructing Antiviral Demand Scenarios	165
4.2	Simulating System Performance	170
4.2.1	Example Results	175
4.3	Antiviral Allocation Models	180
4.3.1	SNS-Distributor Inventory Allocation	182
4.3.2	Distributor-Pharmacy Inventory Allocation	188
4.4	Ongoing and Future Work	191
A	D-PODS User Manual	192
A.1	Introduction	192
A.2	Glossary of Important Terms	193
A.3	Working with D-PODS	194
A.3.1	Getting Started	194
A.3.2	Creating or Selecting a Case	195
A.3.3	D-PODS Menu	197
A.3.4	Step 1: Constructing the Model	197
A.3.5	Step 2: Input Arrival Rate Information	202
A.3.6	Step 3: Service Time Parameters	204
A.3.7	Step 4: Staffing Requirements	206
A.3.8	Step 5: Simulation Parameters	209
A.3.9	Step 6: Input Case Name	210
A.3.10	Step 7: Run the Simulation	211
A.3.11	Step 8: View the Results	211
A.4	Navigating the Access Database	215
B	ESCOE User Manual	218
B.1	Introduction	218
B.2	Model Assumptions	219
B.2.1	Glossary of Important Terms	221
B.3	Working with ESCOE	222
B.3.1	Getting Started	222
B.3.2	Selecting an EXCOE Case	224
B.3.3	ESCOE Menu and Step 1	226
B.3.4	Step 2: Constructing the Network	229
B.3.5	Step 3: Define Lead Times	232
B.3.6	Step 4: Describe the Inventory	233
B.3.7	Step 5: Describe the SNS	235
B.3.8	Step 6: Describe the FDSs	238
B.3.9	Step 7: Describe the RSSs	240
B.3.10	Step 8: Describe the POD Types	242
B.3.11	Step 9: Describe the Simulation Experiment	245
B.3.12	Step 10: Run the Simulation	247
B.3.13	Step 11: View the Results	248

B.4 Navigating the Access Database	250
Bibliography	255

LIST OF TABLES

2.1	Table of Model Notation. Variables names written in boldface, with one or more subscripts suppressed, indicate a vector (e.g., $\mathbf{x}_t = (x_{0t}, x_{1t}, \dots, x_{M+N,t})$). Adding the superscript <i>past</i> to a variable indicates the vector of past values of the variable, starting one lead time ago, and, if it is known, the current value (e.g., $\mathbf{x}_{mt}^{\text{past}} = (x_{m,t-\tau_m}, x_{m,t-\tau_m+1}, \dots, x_{mt})$ and $\mathbf{d}_{nt}^{\text{past}} = (d_{n,t-\tau_n}, d_{n,t-\tau_n+1}, \dots, d_{n,t-1})$).	25
2.2	Simulation Parameters. *All simulations began with initial RSS inventories set at two periods of expected demand and initial SNS inventories set at five periods of expected demand, except for simulations 3.23 and 3.24, which were initialized with five periods of expected demand at the RSSs and nine periods of expected system demand at the SNS.	99
2.3	Average total costs scaled by the Wait-and-See (WS) and Lagrangian Relaxation (LR) lower bounds. The minimum value in each row is indicated by bold font.	107
2.4	Number of iterations for which each solution method gives the smallest / second smallest cost.	108
2.5	Number of Minima and Second Place (Imperfect Ordering). . . .	109
3.1	Simulation parameter values.	130
4.1	HHS Regions.	165
4.2	First day on which a stockout of adult antivirals occurs at pharmacies in New Jersey (NJ) and Georgia (GA) in moderate pandemic scenarios under the Patients Leave policy.	178
4.3	First day on which demand exceeds service capacity at pharmacies in New Jersey (NJ) and Georgia (GA) in moderate pandemic scenarios under the Patients Return policy.	178
4.4	Influenza model notation.	184

LIST OF FIGURES

1.1	The United States Strategic National Stockpile distribution network.	6
2.1	Location numbers in our model of the SNS network.	21
2.2	Timeline of events within a single time period.	27
2.3	Total cost incurred for all simulations.	103
2.4	Average per patient waiting time for experiments 1 and 2.	116
2.5	Average units of inventory required per person served for experiments 1 and 2.	117
2.6	Average per patient waiting time and inventory required per person for simulations 1.11, 1.22, 2.21-2.23, and 3.11-3.12.	118
2.7	Average per patient waiting time and inventory required per person for simulations 1.11, 1.12, 1.14, 1.21, 1.22, and 1.24.	119
2.8	Average per patient waiting time and inventory required for experiment 3.	120
3.1	The main menu of the D-PODS interface.	126
3.2	POD patient flow diagram.	129
3.3	Expected patient arrival scenarios.	131
3.4	Average patient time spent in the POD and staff-hours required each day, given constant or dynamic staffing plans.	132
3.5	The length of the queue at the Greeting Station for Scenario A for Constant Staffing (top) and Dynamic Staffing (bottom). The small blue dots show the queue lengths from 10 simulation replications for every five minute interval; the yellow-green circles show the 95th percentile queue length for each five minute interval.	133
3.6	The length of the queue at the Triage Station for Scenario A for Constant Staffing (top) and Dynamic Staffing (bottom). The small blue dots show the queue lengths from 10 simulation replications for every five minute interval; the yellow-green circles show the 95th percentile queue length for each five minute interval.	134
3.7	Average patient time spent in the POD for patient arrival Scenario A for each of the four staffing plans.	136
3.8	Queue lengths at the Greeting station for Plans 1 and 4 with patient arrival Scenario A.	137
3.9	Queue lengths at the Triage station for Plans 1 and 4 with patient arrival Scenario A.	138
3.10	ESCOE interface main menu.	144
3.11	Emergency supply chain diagram, including Forward Deployed Stockpiles (FDSs).	145

3.12	The ESCOE interface for describing the distribution network. . .	146
3.13	Average queue lengths at a small POD for the 0 and 2 FDS cases, with Early (4th hour) and Late (12th hour) POD opening times. .	154
3.14	Average queue lengths at a large POD for the 0 and 2 FDS cases, with Early (4th hour) and Late (12th hour) POD opening times. .	154
3.15	Inventory on-hand over time in the seventh and eighth simulation replications for one small POD that opens in the twelfth hour without FDS support.	156
3.16	Queue lengths over time in the seventh and eighth simulation replications for one small POD that opens in the twelfth hour without FDS support.	157
3.17	Inventory on-hand over time in the seventh and eighth simulation replications for one small POD that opens in the twelfth hour without FDS support.	157
4.1	Expected patient demands for each of the ten HHS regions during a moderate 2009-like pandemic scenario.	161
4.2	Burn rate for adult antivirals in a severe 1918-like scenario. . . .	163
4.3	Burn rate for adult antivirals in a severe 1957-like scenario. . . .	164
4.4	Number of patients served in the distribution network under two possible patient behavior strategies: patients leave the system stop seeking antivirals if inventory is unavailable or patients return until they are served.	177
4.5	The top graph shows the average demand for antivirals each day and the bottom shows the maximum demand on a single day, under the Patients Leave policy.	180
4.6	The top graph shows the average newly arriving demand for antivirals each day and the bottom shows the maximum actual demand (new arrivals + returning patients) on a single day, under the Patients Return policy.	181
A.1	Case selection screen shot.	196
A.2	Load case data screen shot.	196
A.3	D-PODS menu screen shot.	198
A.4	Manage cases screen shot.	199
A.5	Step 1 screen shot.	199
A.6	Step 2 screen shot.	203
A.7	Step 3 screen shot.	205
A.8	Step 4 screen shot.	207
A.9	Output tables initial screen shot.	212
A.10	Output tables station performance measures screen shot.	213
A.11	Screen shot of output graphs after location selection and replication selection.	214

B.1	Emergency supply chain diagram.	220
B.2	Case Selection screen shot.	224
B.3	Load Case Data screen shot.	225
B.4	ESCOE Menu screen shot.	227
B.5	Manage cases screen shot.	228
B.6	Step 2 screen shot.	230
B.7	RSS/POD Type Relationship Table Shot.	231
B.8	Step 3 screen shot.	233
B.9	Step 4 screen shot.	234
B.10	Step 5 screen shot.	237
B.11	Step 6 screen shot.	239
B.12	Step 7 screen shot.	241
B.13	Step 8 screen shot.	243
B.14	Step 9 screen shot.	246
B.15	Output tables screen shot.	249
B.16	Screen shot of output graphs page, before data to graph has been selected.	250
B.17	Screen shot of output graphs of an patient demand at PODs of a particular type that are served by one of the RSSs.	251

CHAPTER 1

INTRODUCTION

Public health emergencies require rapid responses from federal, state, and local authorities to prevent widespread mortality and morbidity. However, existing response plans seldom account for the variety of risks and uncertainties inherent in emergency scenarios. Our goal is to construct models that will help policy makers respond effectively to potential emergencies.

Throughout this thesis, we use the phrase “public health emergency” to refer to events that are caused by biological agents, including viruses, bacteria, and other toxins, and we focus on events that could occur in the United States. Examples of potential emergencies include naturally occurring events, like influenza pandemics, as well as deliberate bioterror incidents, such as a smallpox or anthrax attack. Since it is impossible to anticipate and prevent all public health emergencies, preparedness is of utmost importance. Public health emergency preparedness involves creating response plans that minimize the impact of an emergency event and ensure that the individuals and organizations that will be called upon to respond have the requisite resources, training, and tools. In the United States the Centers for Disease Control and Prevention (CDC) coordinates preparedness efforts at the national level; but states, counties, and some private organizations also dedicate significant time and energy to preparedness planning. The CDC provides funding and guidance to the states and some cities; the states, however, are largely responsible for managing the response to emergencies that affect their populations.

This thesis will present models and simulation tools that will contribute to emergency preparedness at the federal, state, and local levels. We have de-

signed quantitative methods to help policy-makers better understand and plan for emergencies. We focus on two particular types of emergencies: a large-scale inhalational anthrax attack and an influenza pandemic. These two emergencies were chosen because they have been designated as significant risks to the population of the United States and because they require very different response systems.

Anthrax has existed as a disease in humans and animals for hundreds of years [Inglesby, 2002, Cieslak & Eitzen, 1999]. It is caused by the bacteria *Bacillus anthracis*, which forms spores that can survive for decades in soils [Cieslak & Eitzen, 1999]. Humans can contract anthrax in three ways: by inhaling, consuming, or coming into prolonged contact with anthrax spores; the disease is not contagious between humans [Cieslak & Eitzen, 1999]. In nature, these spores tend to clump and bind to soils, mitigating the likelihood that humans will contract the disease, but it is possible to produce weapons-grade anthrax powder with a high concentration of spores and low electrostatic charge to minimizing clumping [Cieslak & Eitzen, 1999, Inglesby, 2002]. The existence of weapons-grade anthrax has made the possibility of a large scale inhalational anthrax attack one the most serious bioterrorist threats facing the United States [Henderson, 1999, Cieslak & Eitzen, 1999, Inglesby, 2002].

Inhalational anthrax begins with an incubation period of 1 to 6 days, followed a prodromal phase during which cold and flu-like symptoms appear [Cieslak & Eitzen, 1999, Wilkening, 2008]. Major organ failure follows not long after. If treatment is not begun within 48 hours after the onset of symptoms, death is a likely outcome for as many as 95% of patients [Cieslak & Eitzen, 1999]. An anthrax vaccine has been developed, but it is not widely used due to lim-

ited supplies and low production capacity, as well as concerns about adverse reactions [Inglesby, 2002]. The recommended treatment for inhalational anthrax is a 60 day regimen of antibiotic prophylaxis [Brookmeyer *et al.*, 2003, Inglesby, 2002]. To minimize the number of mortalities following an inhalational anthrax attack, the CDC's goal is to ensure that all individuals exposed to anthrax spores begin a course of antibiotic prophylaxis within 48 hours of the time the attack is detected [CDC, 2004].

Influenza is a very different type of disease. It is caused by the rapidly mutating influenza virus. There are two main types of influenza, A and B, but many different strains of the virus exist and new ones appear each year [Barr *et al.*, 2010]. Most strains are seasonal, which means that they are contagious and can be fatal for high-risk groups, such as the elderly, the very young, and the immunocompromised, but for most of the population, the severity is relatively low [CDC, 2010]. A pandemic strain of influenza is different from seasonal ones because of its much higher rates of severe complications and fatalities, combined with an unusually high reproductive rate so that the disease spreads much more quickly than the seasonal strains.

The World Health Organization (WHO) coordinates ongoing international influenza surveillance to identify the current strains of influenza circulating in the population [Gerdil, 2003, Barr *et al.*, 2010]. Every six months, the WHO compiles the results of this surveillance to identify the strains that will likely be circulating during the upcoming influenza season [Gerdil, 2003, Barr *et al.*, 2010]. In the Northern Hemisphere, the influenza season runs from approximately November through April; in the Southern Hemisphere, it runs from May through October. To produce, validate, and distribute a tailored influenza vac-

cine takes six to eight months [Gerdil, 2003, Barr *et al.*, 2010]. Hence, the WHO makes its recommendations in February and September for the Northern and Southern Hemisphere influenza seasons, respectively [Gerdil, 2003, Barr *et al.*, 2010]. The efficacy of each vaccine depends on the degree to which the actual influenza strains circulating in a population match those predicted by the WHO.

Influenza is spread through particles, droplets, and direct contact, which makes it easy to pass between humans [CDC, 2010]. Infected individuals experience a 1 to 4 day asymptomatic incubation period; for most people this is followed by a 3 to 7 day period of illness after which they recover and retain immunity to the virus [CDC, 2010]. The influenza virus may also cause viral pneumonia and other secondary respiratory illnesses, and it may exacerbate existing cardiac, pulmonary, and other medical conditions [CDC, 2010]. Infected individuals are infectious to others during the last day of the incubation period and remain so for 5 to 10 days following the onset of illness [CDC, 2010]. However, young children and immunocompromised individuals may remain infectious for significantly longer periods of time [CDC, 2010].

When a vaccine is unavailable, antiviral drugs may be used as prophylaxis to reduce the likelihood of infection. For maximum effectiveness, prophylaxis must be continued for ten days to six weeks [Fiore *et al.*, 2011]. Antiviral drugs may also be used to treat patients who contract influenza. Patients who begin taking antivirals before becoming symptomatic experience reduced symptoms and have a lower likelihood of transmitting the virus to others [Fiore *et al.*, 2011]. Antivirals, when used as treatment, are usually prescribed for five days [Fiore *et al.*, 2011]. Antiviral drugs have the advantage of being fairly flexible; they are specific only to influenza type (A or B), rather than particular strains. Given a

limited supply of antivirals, far fewer individuals can be provided with prophylaxis, compared to treatment. For this reason, during an influenza pandemic the CDC pandemic influenza plan calls for antivirals to be used exclusively for treatment of those to whom a physician has issued a prescription [HHS, 2005]. As noted earlier, a targeted vaccine is unlikely to be available during the early stages of a pandemic.

For both anthrax and pandemic influenza, the CDC has taken responsibility for stockpiling the antibiotics and antivirals that would be used to mitigate these emergencies [CDC, 2004, HHS, 2005]. The United States Department of Health and Human Services (HHS) and the Centers for Disease Control and Prevention created the Strategic National Stockpile (SNS), formerly the National Pharmaceutical Stockpile, in 1999 to ensure that essential pharmaceuticals and medical supplies will be available to the American population during emergencies [CDC, 2011]. The SNS stores its inventory in multiple warehouses around the country to ensure that it can deliver an initial shipment of antibiotics and medical supplies to any affected state rapidly [CDC, 2011]. Supplies for selected fast-acting emergencies like anthrax have been collected in “Push Packages,” which are stored in ready-to-ship pallets that can be shipped to states within twelve hours after the SNS is activated [CDC, 2011]. Additional shipments tailored to the particular emergency may be sent over time.

Each state has at least one Receiving Staging and Storing (RSS) warehouse, which will be opened to accept materials from the SNS and send them to the appropriate dispensing locations [Nelson *et al.*, 2008]. There are several proposed methods for dispensing medical supplies at the local level, but the most common is the use of Points of Dispensing (PODs), which are ad hoc clinics set up

throughout the affected region [SNS, 2008]. These PODs may be operated by state or local health departments to distribute antibiotics to the general public or they may be “closed” locations that serve well-defined populations such as hospital patients, nursing home residents, or prison inmates [SNS, 2008]. In the case of an anthrax attack, this distribution network, shown in Figure 1.1, may only remain in place for days. Public health planners must ensure that the logistics of the system work seamlessly to ensure that the affected population is served within the 48 hour window to minimize mortality and morbidity.

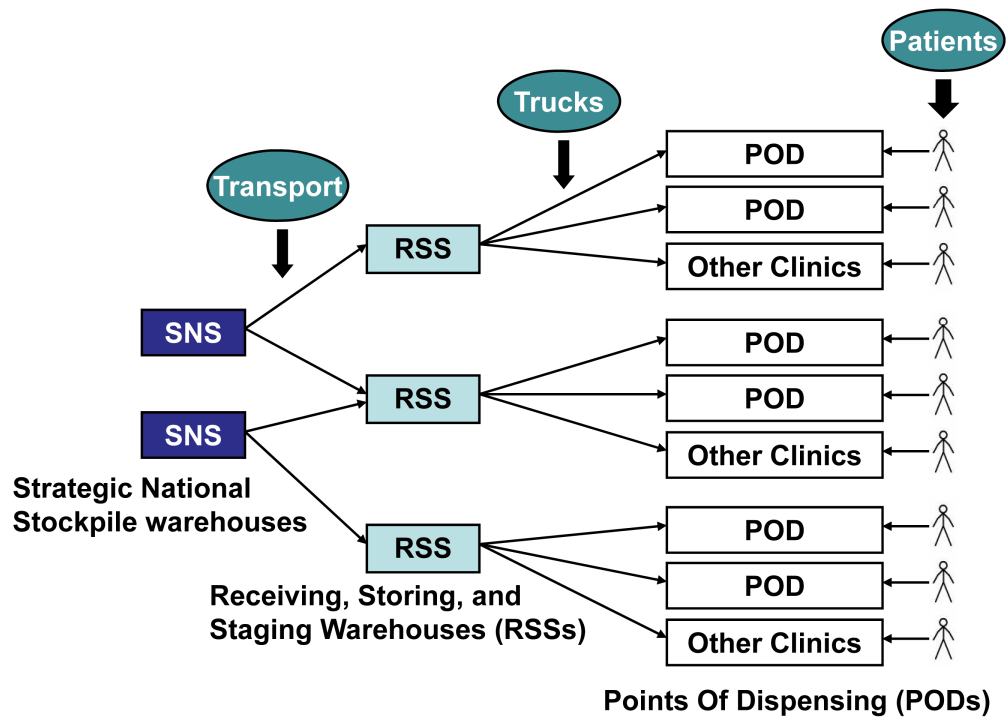


Figure 1.1: The United States Strategic National Stockpile distribution network.

The distribution of medical supplies to all people living in an affected area following a large-scale emergency like an anthrax attack is called a “mass-dispensing campaign.” A large mass-dispensing campaign would also be carried out during an influenza pandemic once a targeted vaccine is available, al-

beit over a longer period of time than is allowed after an anthrax attack. However, both before and after a vaccine is available antivirals will be distributed to those who seek treatment for influenza, but no one will receive them without a written prescription. Such a scenario calls for a “controlled-dispensing campaign,” in which only a subset of the population in any given region is served. Current CDC plans call for state health departments to run the antiviral dispensing in accordance with their emergency response plans. The SNS distributed 11 million courses of antivirals to the states during early stages of the 2009 H1N1 pandemic [HHS, 2009]. The states were responsible for the distribution of these antivirals to their populations. The states were also responsible for the initial distribution of the H1N1 vaccine, but when the vaccine became widely available to the public in December 2009, the CDC worked with states to provide H1N1 vaccine directly to commercial pharmacies in an effort to make the vaccine more accessible [Koonin *et al.*, 2011]. The use of pharmacies to dispense vaccines was considered highly successful [Koonin *et al.*, 2011]. Lisa Koonan, the Senior Advisor for the Influenza Coordination Unit in the Office of Infectious Diseases at CDC, and others have pointed out that the commercial pharmaceutical supply chain is well-suited to a long-term controlled antiviral dispensing campaign [Koonin *et al.*, 2011, Lien *et al.*, 2006]. This would avoid straining the resources of state health departments, leaving them free to focus on providing antivirals to special groups that cannot reach commercial pharmacies, such as nursing home patients or residents of Native American reservations.

The proposed system would take advantage of a highly effective supply chain that already serves the vast majority of Americans. There are more than 60,000 commercial pharmacies in the United States; over 95% live within five miles of one of these locations [SK&A, 2011, CDC, 2012]. These pharmacies cur-

rently fill over 90% of the prescriptions in the United States; the remainder use mail-order or other sources [CDC, 2012]. A small number of national pharmaceutical distributors make daily deliveries to over 93% of the pharmacies, and computerized systems are already used to transfer information about inventory needs from many of the pharmacies to their distributors [CDC, 2012]. Leveraging this system to include antiviral distribution would allow the states to avoid recreating a system that already exists. However, while the idea of using the commercial system may be appealing, it raises many questions about which pharmacies should be included in the dispensing effort, how inventory allocation decisions will be made, and how the CDC can ensure that inventory is allocated “fairly,” without favoring or marginalizing particular populations.

In the following three chapters, we present models of the SNS distribution network responding to both an anthrax attack and an influenza pandemic. In Chapter 2 we describe a resource allocation model of the full SNS-RSS-POD distribution network under an anthrax attack scenario. We present several inventory allocation strategies, including a novel Lagrangian relaxation approach. We show computational results from testing these strategies with simulations and discuss public health policy implications. In Chapter 3 we present two simulation models, one of a single POD and one of the full three echelon supply chain, under an anthrax scenario. We describe how these models can assist planners who are responsible for designing these response systems. In Chapter 4 we focus on the influenza pandemic scenario. We explore the proposed use of the commercial supply chain to dispense antivirals. We present models that could be used to make inventory and staff allocation decisions during the pandemic, describe a method for estimating historical pandemic curves, and use a simulation of the dispensing network to identify potential strengths and pitfalls of the

proposed system. To give context to this work, the next section will review the relevant public health modeling literature.

1.1 Emergency Preparedness Literature Review

In the last decade, a great deal of attention has been given to planning an effective response to an anthrax attack, and operations researchers have made some significant contributions to these efforts. A number of papers have presented high-level models and simulations of the anthrax response network to help policy-makers invest in the most cost-effective mitigation strategies. [Craft *et al.*, 2005], building off earlier work in [Wein *et al.*, 2003], constructed a simulation of an anthrax attack and the response, from spore dispersal to detection to prophylaxis and treatment. They identify the need for an education program to impress on people the importance of fully adhering to the antibiotic prophylaxis regime; the importance of minimizing the delay in beginning prophylaxis and maximizing POD throughput rates; and the need for increased hospital surge capacity. [Bravata *et al.*, 2006] used a compartmental simulation model to study the cost-effectiveness of various emergency response strategies. They found that POD dispensing capacity is the most important determinant in the success of a response plan, and show that the cost-effectiveness of many strategies is sensitive to the probability that an anthrax attack occurs. [Braithwaite *et al.*, 2006] considered several different anthrax attack scenarios and concluded that effective surveillance and rapid prophylaxis is more cost-effective than large-scale anthrax vaccination, in agreement with the previously mentioned papers.

However, there have been many papers that have questioned whether the

CDC's antibiotic prophylaxis plan is the best possible medical response. Other possible responses include pre- or post-exposure vaccination as well as antibiotic treatment for those who have become symptomatic. [Fowler *et al.*, 2005] and [Fowler & Shafazand, 2011] found that post-exposure vaccination and antibiotic prophylaxis is the most effective and least costly strategy. However, they noted that if antibiotic prophylaxis cannot be distributed quickly after exposure, then pre-exposure vaccination would become cost-effective. [Brookmeyer *et al.*, 2004] and [Baccam & Boechler, 2007] modeled the impact of various antibiotic prophylaxis and vaccination strategies. They both agreed that rapidly beginning post-exposure antibiotic prophylaxis and high levels of adherence are essential, but Baccam and Boehler concluded that vaccination may also be valuable, while Brookmeyer *et al.* disagreed. [Hupert *et al.*, 2009] found that the total time required to complete the antibiotic dispensing campaign and the initial delay before beginning antibiotic prophylaxis must be carefully controlled to minimize mortality and morbidity. Others have confirmed these conclusions for various scenarios [Mitchell-Blackwood *et al.*, 2011, Schmitt *et al.*, 2007]. Despite a wide variety of model parameter values and assumptions, there is general agreement among all of these papers that antibiotic prophylaxis is a highly effective strategy, provided that the dispensing campaign is begun soon after the initial anthrax attack and that it is concluded promptly, within the CDC's 48 hour goal.

Another set of papers have focused on modeling the PODs in great detail, trying to find optimal designs and staffing strategies that will help the clinics operate as efficiently as possible. [Hupert *et al.*, 2002] used simulation to determine POD staffing levels based on patient arrivals; this work was an early example of a model that was widely used to assist in POD planning. [Washington,

2009] simulated POD operations to show that assigning staff efficiently within a POD allows the clinic to serve more patients over time. [Lee *et al.*, 2006b] and Lee06b present RealOpt, a decision support tool that planners can use in real time to determine staffing and the layout of the POD stations. They describe how RealOpt has been successfully used in POD drills around the country. [Lee *et al.*, 2010b] extended this work to model the spread of disease within PODs, when PODs are operated during the outbreak of an infectious disease such as influenza. [Aaby *et al.*, 2006] developed the Clinic Planning Model Generator, a queuing-based capacity-planning model that estimates POD performance based on user-specified inputs.

Other papers have focused on addressing practical planning questions related to other parts of the anthrax response plan. [Zaric *et al.*, 2008] present a compartmental simulation model in conjunction with a simple spreadsheet-based interface to let policy-makers use the model to address many planning questions. They emphasize the need for rapid dissemination of the news that an attack has occurred to the affected population, and, like [Wein *et al.*, 2003], they stress the importance of education to ensure a high level of adherence to the prophylaxis regimes. They also question the value of push packages and show that direct tailored shipments may be more useful, if these can be delivered rapidly. [Lee *et al.*, n.d.] present a tool for determining optimal POD locations using a facility location optimization model. [Berman *et al.*, 2011] address the same problem using game theoretic methods to avoid the likelihood of a terrorist attack on the PODs. [Lu *et al.*, 2010] present an algorithm that uses Markov switching models to improve the performance of surveillance systems. [Montjoy & Herrmann, 2010] provide algorithms for routing delivery vehicles in emergency scenarios that will provide as much flexibility as possible

to account for potential disruptions in traffic. [Lee, 2008] considers the response plan to more general emergencies and shows that the inclusion of a basic supply chain model can help planners better understand the resource needs and requirements of a response plan .

The requirements for an effective response to an influenza pandemic are less well-defined than those of an anthrax attack response plan. Since the appearance of H5N1 in the early 2000s, a large number of papers have used analytic and simulation models to better understand the spread of influenza globally and locally, as well as the value of potential mitigation strategies. [Germann *et al.*, 2006] used a large-scale, computationally intensive agent-based simulation study of potential pandemics in the United States to explore the value of strategies including antiviral prophylaxis, mass vaccination, school closures, and social distancing through travel restriction, quarantine, or voluntary behavior modification. They found that school closures, combined with antiviral prophylaxis for high-risk individuals, could have a significant impact on the spread of the pandemic. [Ferguson *et al.*, 2006] also used a large-scale computer simulation, originally developed in [Ferguson *et al.*, 2005], to study a potential pandemic in the United States under a variety of mitigation scenarios. They found that travel restrictions were unlikely to be effective, and that school closures would have a minimal effect on the overall pandemic. [Glass *et al.*, 2006] focused on simulating influenza within a single small town and concluded that social distancing methods could be highly effective, even in the absence of antivirals and vaccinations. [Larson, 2007], [Nigmatulina & Larson, 2009], and [Teytelman & Larson, 2012] present analytic models that account for some population heterogeneity and emphasize the importance of social distancing measures, with particular focus on highly social and highly susceptible individuals.

[Milne *et al.*, 2008] used a detailed simulation of a single community, with a model that included detailed social dynamics, to consider the value of non-pharmaceutical mitigation measures. They found that school closures and other social distancing measures are likely to be moderately effective in containing a pandemic if implemented sufficiently early and the influenza attack rate is moderate. However, [Milne *et al.*, 2008] points out that, despite using similar epidemiological data, their results differ significantly from those of several previously mentioned papers, [Germann *et al.*, 2006], [Ferguson *et al.*, 2006], and [Glass *et al.*, 2006]. They emphasize that this indicates the degree to which model parameters affect the outcome of these types of simulation studies. [Riley, 2007] reviews several influenza models, and draws a similar conclusion, as do [Halloran *et al.*, 2008], who present a simulation study showing that a combination of mitigation measures is likely to be useful, but caution that these results must be considered tentative without better data to support the model parameters.

In spite of these concerns, many researchers have continued to use modeling techniques to better understand how an influenza pandemic might be controlled. [Wein & Atkinson, 2009] focused on modeling the transmission of influenza within a household. They identify several key factors for reducing transmission rates, including using separate bedrooms for infected individuals and beginning infection control measures immediately upon the introduction of illness to the household. A number of researchers addressed the potential value of travel restrictions. [Colizza *et al.*, 2007] and [Epstein *et al.*, 2007] used simulation models to show that travel restrictions could be useful in containing the spread of influenza worldwide. [Brownstein *et al.*, 2006] used an empirical study of airline flight data to conclude that travel restrictions could shift the

peak of a pandemic by several weeks, providing more time for vaccine production and dissemination. [Balcan *et al.*, 2009] found that international travel patterns primarily affect the spread of influenza during the first weeks of a pandemic; however, once the disease has been introduced into a region local effects take over.

When the H1N1 pandemic occurred in 2009, a number of researchers immediately began working to determine epidemiological parameters. One essential parameter estimated was the basic reproductive number, R_0 , which indicates the expected number of new cases that would be generated by an infectious individual in a population of susceptible people [Fraser *et al.*, 2009, Yang *et al.*, 2009, White *et al.*, 2009, Tuite *et al.*, 2010, Presanis *et al.*, 2009]. Disease severity measures, including case hospitalization ratios (CHR) and case fatality ratios (CFR), the percentages of infected individuals who are hospitalized and who die, respectively, were also estimated [Fraser *et al.*, 2009, Presanis *et al.*, 2009]. These papers appeared as early as June 2009, based on the first cases of H1N1 that appeared around the world, particularly in Mexico.

Other modeling papers continued to make an effort to guide public health policy. [Medlock & Galvani, 2009] produced a model whose results indicated that the CDC and the Advisory Committee on Immunization Practices (ACIP) were prioritizing the wrong vaccination groups. The ACIP recommended that equal priority be given to people aged six months to 24 years old, as well as caregivers for infants under six months, pregnant women, and immunocompromised individuals of any age [Lee *et al.*, 2010a]. [Medlock & Galvani, 2009] suggested that school children and their parents (adults aged 30-39) should receive vaccination priority to minimize the spread of influenza. [Lee *et al.*, 2010a]

quickly responded by tailoring a large-scale simulation to the H1N1 scenario and showing that, while the ACIP recommendations may be expected to yield a slightly higher overall attack rate, they also result in decreased mortality, morbidity, and economic consequences by protecting groups at risk for severe complications from influenza.

Since 2009, many influenza-focused papers have attempted to accurately represent the H1N1 pandemic and to determine how events might play out under similar pandemic scenarios in the future. [Carrat *et al.*, 2010] used the influenza surveillance data collected in France to estimate the actual H1N1 attack rates for that country. They used these attack rates to show that, if a new strain of H1N1 is reintroduced into the population, the extensiveness of the vaccination campaign will depend on the level of cross-immunity imparted by the 2009 pandemic. [Halder *et al.*, 2010] performed a detailed simulation of a pandemic occurring in a single town with similar epidemiological parameters to those of the H1N1 pandemic. They investigated various mitigation measures and found that, if the pandemic had become more severe, school closures combined with antiviral treatment would be very effective in stemming the spread of the pandemic. [Savachkin & Uribe, 2011] showed that a vaccine and antiviral distribution network that would allow for redistribution of these resources would be valuable during a more severe pandemic scenario. [Bajardi *et al.*, 2011] modeled the impact of the travel restrictions to and from Mexico during 2009. Contradicting earlier papers, they show that the decline in air travel did not slow the spread of the pandemic and conclude that, given the increasing mobility of the global population, travel restrictions are unlikely to be effective containment strategies in the future.

Many people have dedicated time to sifting through the available data sources that give information about the number, location, and severity of H1N1 cases that occurred during the 2009-2010 season. [Nishiura *et al.*, 2010] and [White & Pagano, 2010] have presented statistical methods to help improve estimates of epidemiological parameters during the early stages of future pandemics. [Shaman *et al.*, 2011] have linked the value of basic reproductive number R_0 for H1N1 to decreased absolute humidity conditions. [Shrestha *et al.*, 2011] have used the data collected by CDC, corrected for missing data, and produced the most complete estimates of the actual number of H1N1 cases that occurred in the United States.

However, researchers continue to struggle to make reasonable inferences from the very limited data collected during the 2009 pandemic. [Lipsitch *et al.*, 2011] describe the need for improved influenza surveillance data that should be collected during a pandemic. They emphasize the high potential value that accurate, real-time data would have for quantitative modeling and public health policy support tools. [Schuchat *et al.*, 2011] describe the CDC's response to the pandemic and some of its successes and challenges, specifically mentioning the importance of various modeling techniques. They also identify the need for accurate real-time data to support public health decision making. [Chao *et al.*, 2011] describe the highly successful use of modeling in Los Angeles County's pandemic planning and response efforts. Models were used to predict the local timing of the pandemic peak and to provide justification for keeping schools open by showing that short-term closures would have minimal value in containing the pandemic, while long-term closures would take too great an economic toll.

Our goal throughout this thesis is to develop models that will similarly help public health authorities respond more effectively in emergency scenarios. [Brandeau *et al.*, 2009] offers a detailed review of papers related to public health and medical disaster response modeling and gives recommendations for evaluating these and future models. In particular, they mention the need for stakeholder input, user-friendly and customizable model interfaces, outcomes that address relevant policy needs, and the inclusion of fundamental uncertainties. They also emphasize the need for models that balance simplicity and complexity. They believe models should focus on specific public health needs without relying on a huge number of unknown parameters that can significantly change the results of the model. [Milne *et al.*, 2008] noted that estimating these parameter values was problematic in determining the value of school closures during influenza pandemics.

In the following chapters, we present a set of models that fill some of the gaps in the literature described above, while aiming to satisfy the goals laid out by [Brandeau *et al.*, 2009]. In Chapter 2 we describe a model of the complete supply chain that would move antibiotics from the SNS to the PODs for a mass-dispensing campaign following an anthrax attack. This is a much needed contribution to the literature, since none of the papers above include supply chain logistics in their models, except for [Lee, 2008], who included a distribution warehouse in his model of a more general emergency event. In Chapter 3 we present a POD simulation model that allows dynamic staffing plans and is the first POD model to include nonstationary patient demand patterns. In the same chapter we present a novel simulation model of the emergency response supply chain; no other model has focused on helping policy-makers understand how the full response system will fit together. In Chapter 4 we address the open

question of whether the commercial pharmaceutical supply chain could support a controlled antiviral dispensing campaign during an influenza pandemic.

CHAPTER 2

ANTHRAX RESPONSE NETWORK MODEL

In the previous chapter we explained why it is essential that the public health emergency response network operate effectively to minimize mortality and morbidity. Running frequent realistic exercises of the response network would be a good way to ensure that all parts of the network will perform well, but such exercises would also be extremely expensive and time-consuming. In reality, exercises are only run occasionally, and they are usually carefully planned, thereby removing the elements of surprise and confusion that would be present during an actual emergency response effort. Instead, policy-makers have come to rely on models, like those described in Section 1.1, to help them understand how to design and execute emergency response plans. Planning-oriented models can help policy-makers understand the potential performance of their response plans. Our goal in this section is to present a model that will help planners evaluate the possible outcomes that may occur due to their distribution network designs, supply chain logistic plans, and inventory allocation policies.

Our model includes physical parameters of the system, including transportation capacity, lead times, and the structure of the network. We construct two methods for making inventory allocation decisions within a given system, with the aim of minimizing inventory in the system and patient delay. The next section describes the model notation, constraints, and cost functions. The second section describes the allocation methods that we have constructed, and the third describes the allocation methods currently in general use. The final section presents the results of the simulation study performed to evaluate the allocation methods and demonstrates how these results may help public health

officials better prepare to respond to an anthrax attack.

2.1 Model Description and Notation

We now describe a model of the SNS-RSS-POD network described in the previous chapter. The SNS has strong central control and contractual agreements with competent shipping organizations to ensure that inventory can be moved around as necessary, so we model the SNS as a single location. This single SNS facility will be called location 0 in our model. It has a large stockpile of inventory and may be resupplied over time by suppliers with unlimited inventories. We allow M RSSs, numbered $1, \dots, M$, and N PODs, numbered $M+1, \dots, M+N$, as shown in the Figure 2.1. Let $\mathcal{R} = \{1, \dots, M\}$ be the set of RSSs, $\mathcal{P} = \{M+1, \dots, M+N\}$ be the set of PODs, $\mathcal{L} = \{0, \dots, M+N\}$ be the set of all locations, and $\mathcal{S} = \{0, \dots, M\}$ be the set of all upper echelon locations, which includes both the SNS and the RSSs. Each RSS receives inventory from the SNS and each POD receives inventory from exactly one RSS, so the network has a tree structure. We define $u_{mn} = 1$ if location m serves (i.e., sends inventory to) location n and 0 otherwise, and $\mathcal{P}(m)$ is the set of all PODs served by RSS m , $\mathcal{P}(m) = \{n \in \mathcal{P} : u_{mn} = 1\}$.

Time in this model is divided into T periods numbered $1, 2, \dots, T$; we assume that within each period events always occur in a particular order. Define $\mathcal{T} = \{1, \dots, T\}$ to be the set of all time periods. At the beginning of each time period t , the current state of the system is known. This state consists of all past decisions as well as the current inventory levels and patient queues at the PODs. The number of patients waiting in the queue at POD n in period t is given by q_{nt} . We define a quantity called the “echelon on-hand inventory position” to be the

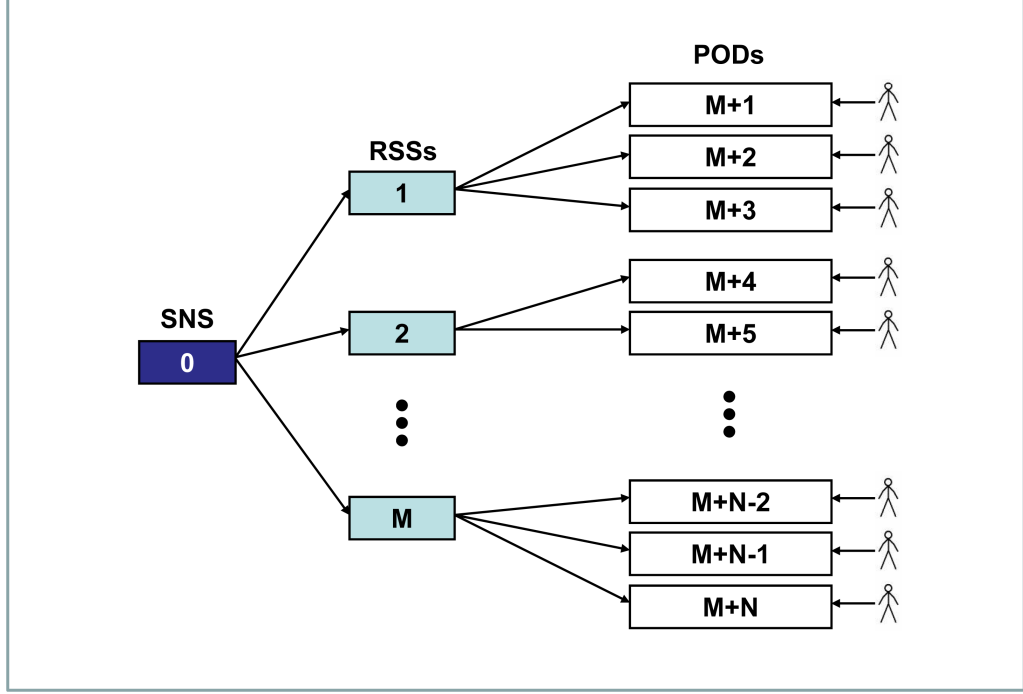


Figure 2.1: Location numbers in our model of the SNS network.

sum of the inventories on-hand at, in-transit to, and downstream from a particular location. Let x_{nt} be the echelon on-hand inventory position at location n at the beginning of period t . Since there are no locations downstream from any POD, at POD n , x_{nt} is simply the amount of inventory on-hand plus the amount in-transit to POD n . At the SNS, which is location 0, x_{0t} is the total amount of inventory on-hand at or in-transit to any location in the system, since all locations in the distribution network are downstream from the SNS.

Given the state variables x_{nt} and q_{nt} for each location, inventory allocation decisions must be made. We choose r_{nt} , the amount of inventory to be shipped to each location n in period t ; the new inventory position for location n then becomes

$$y_{nt} = x_{nt} + r_{nt}.$$

These r_{nt} units of inventory arrive after a lead time of τ_n periods to location n ; notice that the lead time τ_n is assumed to be constant over the time horizon. This is acceptable if we ensure that the time periods are sufficiently coarse (for instance, hours or days rather than seconds or minutes), in which case we can estimate a lead time in terms of periods with accuracy. If the lead time is less than the length of a time period for location n , we let $\tau_n = 0$; in this case, the r_{nt} units shipped in period t arrive in the same time period. We assume that the PODs are relatively close to their respective RSSs, so we set the lead times to zero for the PODs, that is, $\tau_n = 0$ for $n \in \mathcal{P}$. In general, after the allocation decision is made, the shipment sent τ_n periods ago arrives to each location n .

\mathcal{P}	The set of all PODs in the network ($\mathcal{P} = \{M + 1, \dots, M + N\}$)
$\mathcal{P}(m)$	The set of all PODs in the network that are served by RSS m ($\mathcal{P}(m) = \{n \in \mathcal{P} : u_{mn} = 1\}$)
\mathcal{R}	The set of all RSSs in the network ($\mathcal{R} = \{1, \dots, M\}$)
\mathcal{L}	The set of all locations in the network ($\mathcal{L} = \{0, \dots, M + N\}$)
\mathcal{S}	The set that contains the SNS and all RSSs in the network ($\mathcal{S} = \{0, \dots, M\}$)
\mathcal{T}	The set of all time periods ($\mathcal{T} = \{1, \dots, T\}$)
a_{nt}	Service capacity at (or downstream from) location n in period t
a_{n,t_1,t_2}	Service capacity at (or downstream from) location n for periods $t_1, t_1 + 1, \dots, t_2$

$\tilde{C}_{nt}(\cdot)$	Cost function for location n in period t (in the initial formulation of the problem)
$C_{nt}(\cdot)$	Updated cost function for location n in period t , in which all costs associated with decisions made at location n are charged to location n
$C'_{nt}(\cdot)$	Cost function for location n in period t , as a function of inventory position
d_{nt}	Known (observed) demand at (or downstream from) location n in period t
D_{nt}	Random demand at (or downstream from) location n in period t
D_{n,t_1,t_2}	Total demand at (or downstream from) location n in periods $t_1, t_1 + 1, \dots, t_2$ ($D_{n,t_1,t_2} = \sum_{t=t_1}^{t_2} D_{nt}$)
$f_{nt}(\cdot)$	The probability mass function for D_{nt}
$f_{n,t_1,t_2}(\cdot)$	The probability mass function for D_{n,t_1,t_2}
$g_{nt}(\cdot)$	The state transition function for period t for location n
h_{nt}	The holding cost charged for each unit of inventory at or in-transit to location n at the end of period t
h_{nt}^R	The holding cost charged for each unit of inventory at or in-transit to the supplier of location n at the end of period t (so $h_{nt}^R = \sum_{m \in \mathcal{L}} u_{mn} h_{nt}$)
M	Total number of RSSs in the network
N	Total number of PODs in the network
q_{nt}	Number of patients waiting at (or downstream from) location n at the beginning of period t

p_{nt}	Transportation capacity from location n to the locations it serves in period t
p_t^0	Transportation capacity to the SNS in period t
r_{nt}	Inventory shipped to location n in period t
S_{nt}	Number of patients served at (or downstream from) location n in period t
S_{n,t_1,t_2}	Total patients served at (or downstream from) location n in periods t_1, \dots, t_2 ($S_{n,t_1,t_2} = \sum_{t=t_1}^{t_2} S_{nt}$)
\tilde{S}_{n,t_1,t_2}	Upper bound on patients served at (or downstream from) location n in periods t_1, \dots, t_2 , calculated using cumulative values
T	Total number of time periods
u_{mn}	Indicator variable that is 1 if location m serves location n and 0 otherwise
x_{nt}	Echelon on-hand inventory at location n at the beginning of period t
x_{nt}^O	On-hand inventory at location n at the beginning of period t
\bar{x}_{nt}	Inventory position at location n at the beginning of period t ($\bar{x}_{nt} = x_{nt} - q_{nt}$)
y_{nt}	Echelon on-hand inventory at location n after allocation decisions in period t
\bar{y}_{nt}	Inventory position at location n after allocation decisions in period t ($\bar{y}_{nt} = y_{nt} - q_{nt}$)
\mathbf{Z}_t	The vector of all state variables for period t
γ_{nt}	Lagrangian multiplier associated with the relaxation of a service constraint for location n in period t

λ_{nt}	Lagrangian multiplier associated with the relaxation of the non-negative inventory shipment constraint for location n in period t
μ_{mt}	Lagrangian multiplier associated with the relaxation of the transportation constraint for location m in period t ($m \in \mathcal{S}$)
μ_{nt}^U	Lagrangian multiplier associated with the relaxation of the transportation constraint for location m that serves location n in period t ($\mu_{nt}^U = \sum_m u_{mn} \mu_{mt}$)
$\psi_{nt}(\cdot)$	Decomposed dynamic program associated with location n in period t
$\Delta_{mt}(\cdot)$	Penalty function that charges additional costs incurred when location m cannot supply the optimal demands of the locations it serves in period t
τ_n	Lead time to location n from its supplier
τ	Lead time from the SNS to an RSS, assuming that these lead times are identical

Table 2.1: **Table of Model Notation.** Variables names written in boldface, with one or more subscripts suppressed, indicate a vector (e.g., $\mathbf{x}_t = (x_{0t}, x_{1t}, \dots, x_{M+N,t})$). Adding the superscript *past* to a variable indicates the vector of past values of the variable, starting one lead time ago, and, if it is known, the current value (e.g., $\mathbf{x}_{\mathbf{mt}}^{\text{past}} = (x_{m,t-\tau_m}, x_{m,t-\tau_m+1}, \dots, x_{mt})$ and $\mathbf{d}_{\mathbf{nt}}^{\text{past}} = (d_{n,t-\tau_n}, d_{n,t-\tau_n+1}, \dots, d_{n,t-1})$).

The amount of inventory that can be shipped to and from the SNS in each period and the amounts of inventory that can be shipped from the RSSs are limited by transportation constraints. We define p_{mt} to be the maximum number of units that can be shipped from location m at the beginning of period t , for $m \in S$, and we define p_t^0 to be the maximum number of units that can be shipped to the SNS at the beginning of period t . When $p_t^0 = 0$, no inventory can be shipped to the SNS, so we can use this parameter to control when shipments to the SNS are allowed. For most scenarios, we will set $p_t^0 = 0$ for all time periods before 36 or 48 hours have passed since manufacturers could not realistically resupply the SNS any faster. Subsequently, if shipments would be sent once per day or week, then p_t^0 would be 0 for most time periods. Similarly, p_{mt} could be set to 0 to ensure that shipments to the RSSs and PODs are made only at certain times throughout the planning horizon.

After the inventory shipments arrive in period t , patients are served at each POD. The total number of people arriving to location n in period t is called the “patient demand” or just the “demand” and is denoted by D_{nt} . Let $D_{nt_1t_2} = \sum_{t=t_1}^{t_2} D_{nt}$ be the cumulative number of patients who arrive in periods t_1, \dots, t_2 . We assume that the demands are independent by time and location and each D_{nt} is a discrete random variables with a known probability mass function f_{nt} . Let the range of D_{nt} be denoted by \mathcal{D}_{nt} . That is, \mathcal{D}_{nt} is the set such that D_{nt} may only take values in \mathcal{D}_{nt} . The total number of patients who request service at POD n in period t is given by the number waiting at the beginning of the period plus the new demand, $q_{nt} + D_{nt}$.

The number of patients served is limited by each POD’s service capacity for

the current time period. We assume that staffing decisions are made in advance, so at the time of the emergency response, staffing levels at PODs are fixed and deterministic service capacities are known as a function of these plans. Staffing levels may not be modified in response to observed patient demand or inventory availability. Our focus in this model will be on determining the value of a complete response plan and in particular, on the importance of inventory allocation policies. In Chapter 3 we will return to the question of staffing and the value of flexible staffing plans. For now, we define a_{nt} to be the service capacity, or the maximum number of patients who could be served at POD n in period t , and let $a_{nt_1t_2} = \sum_{t=t_1}^{t_2} a_{nt}$ be the cumulative number who could be served in periods t_1, \dots, t_2 . Then the number of people who actually are served at POD n in period t , S_{nt} , can be found. Let $S_{nt_1t_2} = \sum_{t=t_1}^{t_2} S_{nt}$ be the cumulative number served in periods t_1, \dots, t_2 . The number of patients served in each period will be constrained by the inventory available, the number of patients present, and the service capacity.

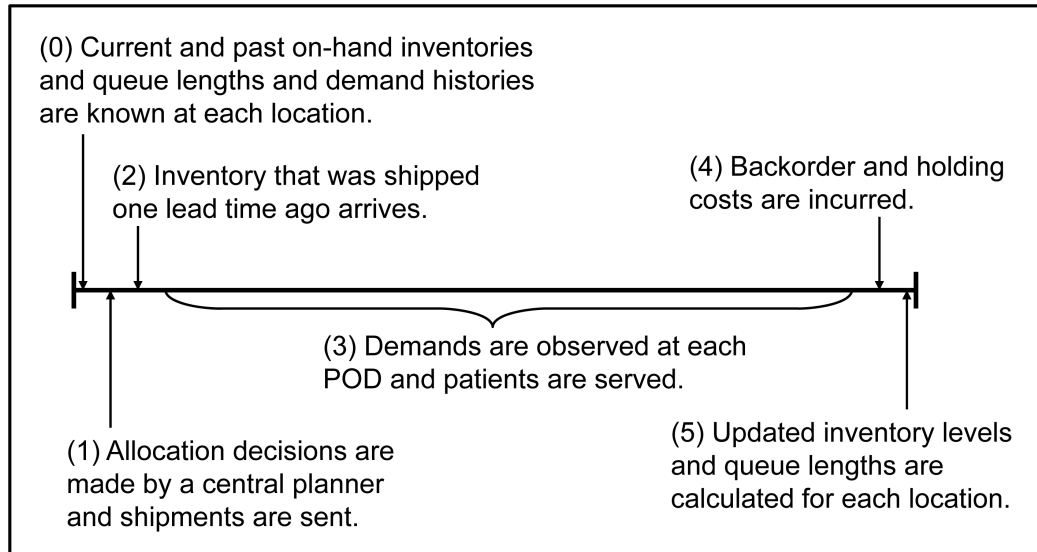


Figure 2.2: Timeline of events within a single time period.

We take an echelon-oriented view of this network. We define echelon demands, service capacities, queue lengths, and patients served at each upper echelon locations, $m \in \mathcal{S}$, to be the sum of those parameters at the locations downstream from m . That is, $D_{mt} = \sum_{n \in \mathcal{L}} u_{mn} D_{nt}$, $a_{mt} = \sum_{n \in \mathcal{L}} u_{mn} a_{nt}$, $q_{mt} = \sum_{n \in \mathcal{L}} u_{mn} q_{nt}$, and $S_{mt} = \sum_{n \in \mathcal{L}} u_{mn} S_{nt}$.

At the end of each time period costs are incurred at every location in the network. Holding costs, h_{nt} , are incurred for each unit of inventory at location n at the end of period t . We will generally assume that holding costs are larger at the RSSs than at the SNS and larger still at the PODs:

$$h_{0t} < h_{mt} < h_{nt} \quad \text{for } m \in \mathcal{R}, n \in \mathcal{P} \text{ and } t \in \mathcal{T}.$$

At the PODs, “backorder” costs are also incurred for any arrived patients who have not been served. These costs are determined by the functions $f_{nt}^B(u)$, where u is the number of unserved patients at POD n at the end of period t . We will assume that $f_{nt}^B(u)$ is convex, and nondecreasing in u , and for $u \leq 0$, $f_{nt}^B(u) = 0$. Thus, there will never be negative backorder costs, but the costs will increase rapidly to help ensure that large queues are not allowed to build up at the PODs due to inventory shortages. Furthermore, since the number of backorders u should always be integer-valued, we can assume without loss of generality that $f_{nt}^B(u)$ is piecewise linear, and that changes in slope only occur at integer-valued points. We will discuss these costs in greater detail in section 2.1.2.

Figure 2.2 shows the sequence of events that we have described. We will refer this figure throughout the rest of this chapter to explain when events occur within a time period. For example, below we will describe how to find the on-

hand inventory at point (0) in a time period, and how to modify this calculation to determine the on-hand inventory at point (2).

At the end of each time period (point (5) in the diagram), the period t demand and numbers of patients served are known for all PODs $n \in \mathcal{P}$, and we can calculate the updated inventory positions and queue lengths, $x_{n,t+1}$ and $q_{n,t+1}$:

$$x_{n,t+1} = y_{nt} - S_{nt} \quad \text{for } n \in \mathcal{L}; \quad (2.1)$$

$$q_{n,t+1} = q_{nt} + D_{nt} - S_{nt} \quad \text{for } n \in \mathcal{L}. \quad (2.2)$$

To make decisions about how much inventory to ship to each location, we need to know how much inventory is on hand and available for distribution at each RSS and the SNS. Inventory located at a location $m \in \mathcal{S}$ at the beginning of period t must have been in-transit to that location no later than period $t - \tau_m$. The echelon on-hand inventory at location m at the beginning of period $t - \tau$ includes all of the inventory in transit to, on hand at, and downstream from m at that time. All of that in-transit inventory will have arrived at m by the beginning of period t . Thus, to determine the on-hand inventory at m in period t , we can use the value $x_{m,t-\tau_m}$ and subtract the current downstream inventory, $\sum_{n \in \mathcal{L}} u_{mn} x_{nt}$ and the inventory that has been used to serve patients in periods $t - \tau_m, \dots, t - 1$, $S_{m,t-\tau_m,t-1}$:

$$\begin{aligned} \text{On-Hand at location } m \text{ at the beginning of } t &= x_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} x_{nt} \\ &\quad - S_{m,t-\tau_m,t-1}. \end{aligned} \quad (2.3)$$

Notice that, for location m and $u < t$, the number of people who were served

in periods $u, \dots, t-1$ is equal to the total number of patients who required service in those periods less the number still waiting at the beginning of period t . This is equal to the number who were already waiting at the beginning of period u , q_{mu} , plus the number who arrived during that time, $d_{m,u,t-1}$, less the number waiting at the beginning of t , q_{mt} :

$$S_{m,u,t-1} = q_{mu} + d_{m,u,t-1} - q_{mt}. \quad (2.4)$$

Substituting $t - \tau_m$ for u , we see that equation (2.3) can be written as

$$\begin{aligned} \text{On-Hand at location } m \text{ at the beginning of } t &= x_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} x_{nt} - q_{m,t-\tau_m} \\ &\quad - d_{m,t-\tau_m,t-1} + q_{mt}. \end{aligned} \quad (2.5)$$

Before patients are served at point (3) in period t , location m also receives an inventory shipment of size $y_{m,t-\tau_m} - x_{m,t-\tau_m}$ and, if it supplies any downstream locations, it ships out a total of $\sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt})$ units of inventory, so the on-hand inventory becomes

$$\begin{aligned} \text{On-Hand at location } m \text{ at point (2) in period } t &= x_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} x_{nt} - q_{m,t-\tau_m} \\ &\quad - d_{m,t-\tau_m,t-1} + q_{mt} \\ &\quad + (y_{m,t-\tau_m} - x_{m,t-\tau_m}) \\ &\quad - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) \\ &= y_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} y_{nt} - q_{m,t-\tau_m} \\ &\quad - d_{m,t-\tau_m,t-1} + q_{mt}. \end{aligned} \quad (2.6)$$

For the SNS and RSSs, equation (2.6) also gives the inventory on-hand at the end of time period t (point (5) in the diagram). At POD n , where the lead time is 0, the amount of inventory available at point (2) is y_{nt} .

Since the lead time to each POD is 0, there is no in-transit inventory for the PODS. For the upper echelon locations, $m \in \mathcal{S}$, we can calculate the amount of inventory in-transit at the beginning of period t by taking the sum of all the shipments that have not arrived by the beginning of period t . These are the shipments that were sent in period $t - \tau_m$ or later. The amount of inventory shipped to m in period r is $y_{mr} - x_{mr}$. Then, using the fact that $x_{m,t+1} = y_{mt} - S_{mt}$, we have

$$\begin{aligned}
\text{In-Transit to location } m \text{ at the beginning of } t &= \sum_{r=t-\tau_m}^{t-1} (y_{mr} - x_{mr}) \\
&= \sum_{r=t-\tau_m}^{t-1} ((x_{m,r+1} + S_{mr}) - x_{mr}) \\
&= x_{mt} - x_{m,t-\tau_m} + S_{m,t-\tau_m,t-1} \\
&= x_{mt} - x_{m,t-\tau_m} + q_{m,t-\tau_m} \\
&\quad + d_{m,t-\tau_m,t-1} - q_{mt}. \tag{2.7}
\end{aligned}$$

During time period t , the in-transit inventory for location m increases when the quantity $y_{mt} - x_{mt}$ is allocated to it, and it decreases when the quantity $y_{m,t-\tau_m} - x_{m,t-\tau_m}$ arrives, so we have

$$\begin{aligned}
\text{In-Transit to location } m \text{ at the end of } t &= x_{mt} - x_{m,t-\tau_m} + S_{m,t-\tau_m,t-1} \\
&\quad + (y_{mt} - x_{mt}) - (y_{m,t-\tau_m} - x_{m,t-\tau_m}) \\
&= y_{mt} - y_{m,t-\tau_m} + q_{m,t-\tau_m} \\
&\quad + d_{m,t-\tau_m,t-1} - q_{mt}. \tag{2.8}
\end{aligned}$$

There is no on-order inventory because allocation decisions are made by a central planner and no inventory will be allocated unless it is available at the supplying location.

The number of patients served in each period is limited by inventory on-hand, patients present, and service capacity. We cannot know how many patients will be served in period t until the patient demands, \mathbf{D}_t , has been observed for that period. At point (4) in the time period, when demand has been observed, we can also calculate the number of patients served at each POD $n \in \mathcal{P}$ in period t :

$$\begin{aligned}
S_{nt} &= \min\{\text{On-Hand at POD at point (2)}, q_{nt} + D_{nt}, a_{nt}\} \\
&= \min\{y_{nt}, q_{nt} + D_{nt}, a_{nt}\} \quad \text{for } n \in \mathcal{P}. \tag{2.9}
\end{aligned}$$

The number of patients served over several periods cannot exceed the cumulative service capacity, patient demand, or inventory available in those periods. Thus, at POD n , the number of patients served in periods $t, t+1, \dots, t+k$ is bounded by

$$S_{n,t,t+k} \leq \min \{a_{n,t,t+k}, \quad q_{nt} + D_{n,t,t+k}, \quad x_{n,t-\tau_n+k} + r_{n,t,t+k}\}. \quad (2.10)$$

Updating the inventory positions for each location using equations (2.1) and (2.2) requires that we know the numbers of patients served over various time intervals. To determine the exact number of patients served at a POD over some length of time, we need to use equation (2.9) for each period, which requires knowing the demands, inventory levels and queue lengths in each period. In a classic inventory model with no limitations on service capacity, equation (2.10) would provide the actual number of people served, rather than an upper bound, and we would only need to know the cumulative demands and inventories to calculate the number served. Our model, however, is more complicated because it is possible to have both inventory available and unserved patients present since service capacity is limited.

2.1.1 Constraints on Inventory Allocation Decisions

Using expressions derived in the previous section, we can construct the constraints that control inventory allocation in each time period. The first constraint ensures that inventory cannot be returned to an upper echelon. We include this constraint because there is no infrastructure in place to allow for returning inventory to upper echelons and returns are impractical in the very short time horizon of an anthrax attack response scenario. We have

$$y_{mt} \geq x_{mt} \quad \text{for all } m \in \mathcal{L} \text{ and } t \in \mathcal{T}. \quad (2.11)$$

The other inventory constraint states that we cannot ship inventory to lower echelons unless it has already been received by the beginning of a time period. The on-hand inventories at the SNS and RSS at the beginning of time period t are given by equation (2.5). Using this expression we obtain the second constraint on allocation decisions:

$$\begin{aligned}
\sum_{n \in \mathcal{L}} u_{mn} r_{nt} &\leq \text{on-hand at location } m \text{ at the beginning of time } t \\
&= x_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} x_{nt} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} + q_{mt} \\
\sum_{n \in \mathcal{L}} u_{mn} y_{nt} &\leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} + q_{mt} \text{ for } m \in \mathcal{S}, t \in \mathcal{T}; \quad (2.12)
\end{aligned}$$

The transportation constraints in period t limit the amount of inventory that can be shipped from the SNS and the RSSs in each period and the total inventory that can be sent to the SNS in each period. These constraints are

$$\sum_{n \in \mathcal{L}} u_{mn} (y_{nt} - x_{nt}) \leq p_{mt} \quad \text{for all } m \in \mathcal{S}, t \in \mathcal{T} \text{ and} \quad (2.13)$$

$$y_{0t} - x_{0t} \leq p_t^0 \quad \text{for all } t \in \mathcal{T}. \quad (2.14)$$

For the SNS or any RSS, $m \in \mathcal{S}$, the outgoing transportation constraint, inequality (2.13), can be rewritten as

$$\sum_{n \in \mathcal{L}} u_{mn} y_{nt} \leq p_{mt} + \sum_{n \in \mathcal{L}} u_{mn} x_{nt},$$

and we see that this constraint will never be active when the on-hand inventory at location m in the beginning of period t is less than p_{mt} :

$$\begin{aligned}
p_{mt} + \sum_{n \in \mathcal{L}} u_{mn} x_{nt} &\leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} + q_{mt} \\
p_{mt} &\leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} + q_{mt} - \sum_{n \in \mathcal{L}} u_{mn} x_{nt}.
\end{aligned}$$

Given the constraints (2.11-2.13), we can characterize the set of feasible solutions at time period t , \mathcal{U}_t . The feasible decisions depend on the state of the network at the beginning of the time period. We need to keep track of the current and previous starting inventory levels, queue lengths, and demands. In time period t we need to know x_{nt} for all $n \in \mathcal{L}$ and for locations $m \in \mathcal{S}$ we need $x_{m,t-\tau_m}$, $q_{m,t-\tau_m} = \sum_{n \in \mathcal{L}} u_{mn} q_{n,t-\tau_m}$, and $d_{m,t-\tau_m,t-1} = \sum_{t'=t-\tau_m}^{t-1} d_{mt'}$. Define

$$\mathbf{x}_{mt}^{\text{past}} = (x_{m,t-\tau_m}, x_{m,t-\tau_m+1}, \dots, x_{mt}); \quad (2.15)$$

$$\mathbf{x}_t^{\text{past}} = (\mathbf{x}_{0t}^{\text{past}}, \mathbf{x}_{1t}^{\text{past}}, \dots, \mathbf{x}_{M+N,t}^{\text{past}}); \quad (2.16)$$

and define $\mathbf{q}_{mt}^{\text{past}}$ and $\mathbf{q}_t^{\text{past}}$ the same way. Define $\mathbf{d}_{mt}^{\text{past}}$, $\mathbf{d}_t^{\text{past}}$, $\mathbf{S}_{mt}^{\text{past}}$, and $\mathbf{S}_t^{\text{past}}$ in a similar manner, except the vector should only include periods $t - \tau_m$ to $t - 1$, since d_{mt} and S_{mt} are unknown at the beginning of time period t . Since the PODs (locations $n \in \mathcal{P}$) have lead times of $\tau_n = 0$, we have $\mathbf{x}_{nt}^{\text{past}} = x_{nt}$, $\mathbf{q}_{nt}^{\text{past}} = q_{nt}$, and $\mathbf{d}_{nt}^{\text{past}} = \emptyset$. In period t , our state variables are the vectors $\mathbf{x}_t^{\text{past}}$, $\mathbf{q}_t^{\text{past}}$, and $\mathbf{d}_t^{\text{past}}$. Each of the first two vectors contains $\tau_0 + \sum_{m \in \mathcal{R}} \tau_m + M + N + 1$ elements, while $\mathbf{d}_t^{\text{past}}$ contains $\tau_0 + \sum_{m \in \mathcal{R}} \tau_m$ elements. For convenience, we will also define a vector of state variables to be $\mathbf{Z}_t = (\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}})$. Using this notation, we can write the feasible set of allocation decisions at time t as

$$\begin{aligned}
\mathcal{U}_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}) = & \left\{ \mathbf{y}_t : y_{mt} \geq x_{mt} \quad \text{for all } m \in \mathcal{L}; \right. \\
& \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - q_{nt}) \leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \quad \text{for } m \in \mathcal{S}; \\
& \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) \leq p_{mt} \quad \text{for all } m \in \mathcal{S}; \\
& \left. y_{0t} - x_{0t} \leq p_t^0 \right\}. \tag{2.17}
\end{aligned}$$

2.1.2 Costs

When allocating inventory, our goal is to minimize the expected cost incurred throughout the system over the time horizon. At the end of each time period, we are charged linear costs for holding inventory at each location in the network and backorder costs for any patients still waiting for service at the PODs.

At the end of time period t , the inventory on-hand at POD n is $y_{nt} - S_{nt}$, and the number of unserved patients waiting for service is $q_{nt} + D_{nt} - S_{nt}$. Thus for $n \in \mathcal{P}$ the total expected cost charged at location n at the end of time period t is

$$\begin{aligned}
\tilde{C}_{nt}(y_{nt}, q_{nt}) = & E \left[h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt})) \right. \\
& \left. + f_{nt}^B(q_{nt} + D_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt})) \right]. \tag{2.18}
\end{aligned}$$

We have written S_{nt} as a function of y_{nt} , q_{nt} , and D_{nt} to emphasize its dependence on these variables. For the SNS and the RSSs, the on-hand inventory at the end of period t is the same as the on-hand inventory at point (2) in period t , which is given in equation (2.6). So for $m \in \mathcal{S}$ the total cost charged at location m at the end of period t is

$$\tilde{C}_{mt}(y_{m,t-\tau_m}, \mathbf{y}_t) = E \left[h_{mt}(y_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} y_{nt} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} + q_{mt}) \right]. \quad (2.19)$$

Notice that by this definition we are charged costs based on decisions made one lead time ago, but we would prefer to express the costs as functions of the current decisions. To do so, we define a modified cost function C by rearranging terms:

$$C_{0t}(y_{0t}, q_{0t}) = h_{0,t+\tau_0}(y_{0t} - q_{0t}) - h_{0,t+\tau_0}E[D_{0,t+\tau_0-1}]; \quad (2.20)$$

$$C_{mt}(y_{mt}, q_{mt}) = (h_{m,t+\tau_m} - h_{0t})(y_{mt} - q_{mt}) - h_{m,t+\tau_m}E[D_{m,t+\tau_m-1}] \quad \text{for } m \in \mathcal{R}; \quad (2.21)$$

$$C_{nt}(y_{nt}, q_{nt}) = E \left[h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt})) + f_{nt}^B(q_{nt} + D_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt})) \right] \\ - \sum_{m \in \mathcal{R}} u_{mn} h_{mt}(y_{nt} - q_{nt}) \quad \text{for } n \in \mathcal{P}; \text{ and} \quad (2.22)$$

$$C_t(\mathbf{y}_t, \mathbf{q}_t) = \sum_{n \in \mathcal{L}} C_{nt}(y_{nt}, q_{nt}), \quad (2.23)$$

where $h_{nt} = 0$ and $D_{nt} = 0$ for $t > T$ and for all n . These definitions are very similar to equations (2.18) and (2.19), but all of the terms in the cost equation for location n have been shifted by τ_n and the terms related to decisions made at location n have been shifted to the cost equation for location n .

It is straightforward to show that, over time, the total cost is equivalent to our original definition, \tilde{C} , less an initial fixed cost. We state this result and the convexity of the cost functions as lemmas.

Lemma 2.1.1. *The total cost calculated using the updated cost functions plus a fixed cost of $\sum_{m \in \mathcal{S}} \sum_{i=1}^{\tau_m} h_{mi} x_{mi}$ is equivalent to the total cost calculated using the original cost functions:*

$$\begin{aligned}
\sum_{t=1}^T C_t(\mathbf{y}_t, \mathbf{q}_t) + \sum_{m \in \mathcal{S}} \sum_{t=1}^{\tau_m} h_{mt}(x_{m1} - q_{m1} - E[D_{m,1,t-1}]) \\
= \sum_{t=1}^T \left[\sum_{m \in \mathcal{S}} \tilde{C}_{mt}(y_{m,t-\tau_m}, \mathbf{y}_t) + \sum_{n \in \mathcal{P}} \tilde{C}_{nt}(y_{n,t-\tau_n}, q_{n,t-\tau_n}) \right].
\end{aligned}$$

Proof. Recall that for all $n \in \mathcal{L}$, we defined $h_{nt} = b_{nt} = 0$ when $t < 1$ or $t > T$. Also, for $t < 1$, $y_{nt} = x_{n1}$ and $q_{nt} = q_{n1}$. We begin with the sum of cost functions defined initially:

$$\sum_{t=1}^T \left[\sum_{m \in \mathcal{S}} \tilde{C}_{mt}(y_{m,t-\tau_m}, \mathbf{y}_t) + \sum_{n \in \mathcal{P}} \tilde{C}_{nt}(y_{n,t-\tau_n}, q_{n,t-\tau_n}) \right].$$

For $m \in \mathcal{S}$ and $t = 1 - \tau_m, \dots, 0$, we substitute x_{m1} for y_{mt} and q_{m1} for q_{mt} :

$$\begin{aligned}
&= \sum_{m \in \mathcal{S}} \sum_{t=1}^{\tau_m} h_{mt}(x_{m1} - q_{m1} - E[D_{m,1,t-1}]) - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - q_{nt}) \\
&\quad + \sum_{m \in \mathcal{S}} \sum_{t=\tau_m+1}^T h_{mt}(y_{m,t-\tau_m} - q_{m,t-\tau_m} - E[D_{m,t-\tau_m,t-1}]) - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - q_{nt}) \\
&\quad + \sum_{n \in \mathcal{P}} \sum_{t=1}^T E[h_{nt}(y_{nt} - S_{nt}) + f_{nt}^B(q_{nt} + D_{nt} - S_{nt})].
\end{aligned}$$

Rearranging terms, changing summation indices, and recalling that $h_{m,t+\tau_m} = 0$ for $t > T - \tau_m$ gives us

$$\begin{aligned}
&= \sum_{m \in \mathcal{S}} \sum_{t=1}^{\tau_m} h_{mt}(x_{m1} - q_{m1} - E[D_{m,1,t-1}]) \\
&\quad + \sum_{t=1}^T h_{0,t+\tau_0}(y_{0t} - q_{0t} - E[D_{0,t+\tau_0-1}]) \\
&\quad + \sum_{m \in \mathcal{R}} \sum_{t=1}^T h_{m,t+\tau_m}(y_{mt} - q_{mt} - E[D_{m,t+\tau_m-1}]) - h_{0t}(y_{mt} - q_{mt}) \\
&\quad + \sum_{n \in \mathcal{P}} \sum_{t=1}^T E[h_{nt}(y_{nt} - S_{nt}) + f_{nt}^B(q_{nt} + D_{nt} - S_{nt})] - \sum_{m \in \mathcal{R}} u_{mn} h_{mt}(y_{nt} - q_{nt}) \\
&= \sum_{m \in \mathcal{S}} \sum_{t=1}^{\tau_m} h_{mt}(x_{m1} - q_{m1} - E[D_{m,1,t-1}]) \\
&\quad + \sum_{t=1}^T \sum_{n \in \mathcal{L}} C_{nt}(y_{nt}, q_{nt}).
\end{aligned}$$

□

Lemma 2.1.2. *The cost functions $C_{nt}(y_{nt}, q_{nt})$ are convex in y_{nt} for fixed q_{nt} .*

Proof. Since the cost functions at the RSS and SNS are linear, they are also trivially convex. To see that the cost function for the PODs is convex, let us rewrite the cost function as

$$\begin{aligned}
C_{nt}(y_{nt}, q_{nt}) &= \sum_{d \in \mathcal{D}_{nt}} f_{nt}(d) \left(h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, d)) + f_{nt}^B(q_{nt} + d - S_{nt}(y_{nt}, q_{nt}, d)) \right) \\
&\quad - \sum_{m \in \mathcal{R}} u_{mn} h_{mt}(y_{nt} - q_{nt}).
\end{aligned}$$

The term $(-\sum_{m \in \mathcal{R}} u_{mn} h_{mt}(y_{nt} - q_{nt}))$ is linear in y_{nt} and therefore convex. We know $f_{nt}(d) \geq 0$, so if the holding cost term $h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, d))$ and backorder cost term $f_{nt}^B(q_{nt} + d - S_{nt}(y_{nt}, q_{nt}, d))$ are convex in y_{nt} for each d and fixed q_{nt} , then

the full cost function is also convex, since nonnegative linear combinations of convex functions are convex.

First consider $h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, d))$. Suppose that $a_{nt} \leq q_{nt} + d$. Then

$$\begin{aligned} h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, d)) &= h_{nt}(y_{nt} - \min(y_{nt}, a_{nt}, q_{nt} + d)) \\ &= h_{nt}(\max(0, y_{nt} - a_{nt}, y_{nt} - q_{nt} - d)) \\ &= h_{nt} \max(0, y_{nt} - a_{nt}). \end{aligned}$$

The maximum of 0 and a linear function of y_{nt} is a convex function of y_{nt} , and multiplying the maximum by the nonnegative constant h_{nt} maintains convexity. Hence, the holding cost term of the cost function is convex in this case. When $a_{nt} > q_{nt} + d$, then

$$\begin{aligned} h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, d)) &= h_{nt}(y_{nt} - \min(y_{nt}, a_{nt}, q_{nt} + d)) \\ &= h_{nt}(\max(0, y_{nt} - a_{nt}, y_{nt} - q_{nt} - d)) \\ &= h_{nt} \max(0, y_{nt} - q_{nt} - d). \end{aligned}$$

For the same reasons as above, this expression is convex in y_{nt} , so the holding cost term is convex for all fixed values of a_{nt} , q_{nt} , and d .

Next, consider $f_{nt}^B(q_{nt} + d - S_{nt}(y_{nt}, q_{nt}, d))$. As with the holding cost term, we can rewrite this as

$$f_{nt}^B(q_{nt} + d - S_{nt}(y_{nt}, q_{nt}, d)) = f_{nt}^B(\max(q_{nt} + d - y_{nt}, q_{nt} + d - a_{nt}, 0)).$$

As discussed above, we know that the maximization is a convex function of y_{nt} . Since $f_{nt}^B(\cdot)$ is a convex, nondecreasing function, we see that the backorder cost term is convex. \square

2.1.3 Dynamic Programming Formulation

Our goal in period t is to make inventory allocation decisions that not only minimize the cost $C_t(\mathbf{y}_t, \mathbf{q}_t)$, but that also minimize the expected cost over all future time periods. In particular, we want to minimize

$$\text{Expected Current and Future Cost} = \sum_{r=t}^T C_r(\mathbf{y}_r, \mathbf{q}_r), \quad (2.24)$$

subject to the appropriate constraints. Let us define the value function

$$\begin{aligned} V_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}) &= \min C_t(\mathbf{y}_t, \mathbf{q}_t) + E[V_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}})] \\ \text{such that } \mathbf{y}_t &\in \mathcal{U}_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}) \end{aligned} \quad (2.25)$$

where the transition functions for the state variable vectors are

$$\begin{aligned}
\mathbf{x}_{n,t+1}^{\text{past}} &= y_{nt} - S_{nt} \quad \text{for } n \in \mathcal{P}; \\
\mathbf{x}_{m,t+1}^{\text{past}} &= \left(x_{m,t-\tau_m+1}, x_{m,t-\tau_m+2}, \dots, x_{mt}, y_{mt} - \sum_{n \in \mathcal{P}} u_{mn} S_{nt} \right) \quad \text{for } m \in \mathcal{S}; \\
\mathbf{q}_{n,t+1}^{\text{past}} &= q_{nt} + D_{nt} - S_{nt} \quad \text{for } n \in \mathcal{P}; \\
\mathbf{q}_{m,t+1}^{\text{past}} &= \left(q_{m,t-\tau_m+1}, q_{m,t-\tau_m+2}, \dots, q_{mt}, q_{mt} + \sum_{n \in \mathcal{P}} u_{mn} (D_{nt} - S_{nt}) \right) \quad \text{for } m \in \mathcal{S}; \\
\mathbf{d}_{m,t+1}^{\text{past}} &= \left(d_{m,t-\tau_m+1}, d_{m,t-\tau_m+2}, \dots, d_{m,t-1}, \sum_{n \in \mathcal{P}} u_{mn} D_{nt} \right) \quad \text{for } m \in \mathcal{S};
\end{aligned}$$

and g_t is the transition function defined so that $\mathbf{Z}_{t+1} = g_t(\mathbf{Z}_t, \mathbf{y}_t, \mathbf{D}_t)$.

From Bellman's Equation, we know that $V_t(\mathbf{x}_t, \mathbf{q}_t)$ is equal to expression (2.24). This dynamic program is computationally intractable, due to its large, complex state space. In the following section we will present several methods for constructing approximate solutions.

2.2 Inventory Allocation Strategies

In this section we will focus on several different mathematical techniques that we use to approximately solve the dynamic program (2.25). We begin by defining the Wait-and-See value, which is the cost that we could achieve if the patient demands were known exactly at the beginning of the time horizon. This value is a lower bound on the cost that could be achieved by any allocation policy. We will then describe a myopic heuristic, which makes decisions based on estimated costs for several periods into the future. Finally, we will present a novel method for relaxing and decomposing the dynamic program.

2.2.1 Wait-and-See Solution

Suppose that we could always wait to make our decisions until we had perfect information about all of the patient demands for the whole time horizon. That is, suppose we could “wait and see” before making our decisions. We could then make optimal inventory allocations by solving a single linear program. The expected cost of implementing this is called the “Wait-and-See” (WS) value, as discussed by Birge and Louveaux [Birge & Louveaux, 1997]. The cost associated with these decisions is a lower bound on the total cost incurred when perfect information about patient demands is unavailable.

Let $\mathbf{d}_t' = (\mathbf{d}_t, \dots, \mathbf{d}_T)$ be the actual demands incurred at all PODs from period t through the end of the time horizon. Let $\mathbf{y}_t' = (\mathbf{y}_t, \dots, \mathbf{y}_T)$ be a complete solution for all locations and all time periods. Recall that we defined the period t state vector $\mathbf{Z}_t = (\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}})$ with the transition function $g_t(\cdot)$. Define the feasible set \mathcal{V}_t to be the set of all feasible decisions for periods t, \dots, T :

$$\mathcal{V}_t(\mathbf{Z}_t, \mathbf{d}_t') = \left\{ \mathbf{y}_t' : \mathbf{y}_t \in \mathcal{U}_t(\mathbf{Z}_t), \mathbf{Z}_{t+1} = g_t(\mathbf{Z}_t, \mathbf{y}_t, \mathbf{d}_t) \right\}. \quad (2.26)$$

Define $\hat{C}_t(\mathbf{y}_t, \mathbf{q}_t, \mathbf{d}_t)$ to be the actual cost incurred as a consequence of decisions made in time period t given the true patient demands. That is, $\hat{C}_t(\mathbf{y}_t, \mathbf{q}_t, \mathbf{d}_t) = C_t(\mathbf{y}_t, \mathbf{q}_t)$ when $Pr(D_{nt'} = d_{nt'}) = 1$ for all $n \in \mathcal{P}$ and $t' = t, \dots, T$. The value of the Wait-and-See solution for periods t, \dots, T is given by

$$WS_t(\mathbf{Z}_t) = E \left[\min_{\mathbf{y}_t' \in \mathcal{V}_t(\mathbf{Z}_t, \mathbf{d}_t')} \sum_{t'=t}^T \hat{C}_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}, \mathbf{d}_{t'}) \right]. \quad (2.27)$$

We see that $WS_t(\mathbf{Z}_t)$ is a lower bound on $DP_t(\mathbf{Z}_t) = V_t(\mathbf{Z}_t)$ by applying Jensen's inequality, since the Wait-and-See solution takes the expectation of the minimum of the total cost, while the dynamic program minimizes the expected total cost.

Lemma 2.2.1. $WS_t(\mathbf{Z}_t) \leq DP_t(\mathbf{Z}_t)$.

Notice that, for fixed sets of patient demands \mathbf{d}'_t , the problem

$$\min_{\mathbf{y}'_t \in \mathcal{V}_t(\mathbf{Z}_t, \mathbf{d}'_t)} \sum_{t'=t}^T \hat{C}_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}, \mathbf{d}_{t'}) \quad (2.28)$$

can be written as a linear program, since $\hat{C}_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}, \mathbf{d}_{t'})$ is convex and piecewise linear. An optimal solution to problem (2.28) must exist, since the objective function is convex and there is always a trivial feasible solution of $y_{nt} = x_{nt}$ for all locations $n \in \mathcal{L}$ and time periods $t \in \mathcal{T}$. Furthermore, we observe that there must always exist an optimal integer solution. Recall that we mentioned earlier that the backorder cost function is piecewise linear whose slope changes only at integer values, so there must exist an unconstrained minimum which is integer-valued. Since all of the parameter values are also integer, bounds on the decision variables will be integer, so there must exist an integer-valued minimum solution. Other non-integer minimum solutions may also exist.

2.2.2 Expected Value Solution

In an actual emergency, we will always need to make logistics decisions before the patient demands are known, so the Wait-and-See solution may be a poor

estimate of the cost that we would actually incur. Instead, assume patient demands equal their expected values. We call this the Expected Value (EV) solution in accordance with Birge and Louveaux [Birge & Louveaux, 1997]. Let

$$EV_t(\mathbf{Z}_t) = \min_{\mathbf{y}_t' \in \mathcal{V}_t(\mathbf{Z}_t, E[\mathbf{D}_t])} \left[\sum_{t'=t}^T \hat{C}_{t'}(\mathbf{y}_t, \mathbf{q}_t, E[\mathbf{D}_t]) \right]. \quad (2.29)$$

The EV problem also has an optimal integer solution, as long as the expected demands, $E[\mathbf{D}_t]$, are integer, since it is simply a special case of the WS problem. We state this as a corollary.

Corollary 2.2.2. *If the initial state vectors \mathbf{x}_1 and \mathbf{q}_1 are integer-valued and the service capacities \mathbf{a}_t and transportation capacities \mathbf{p}_t and p_t^0 are integer for all time periods $t \in \mathcal{T}$, then for any set of integer expected values $E[\mathbf{D}]$ there exists an optimal integer solution to problem (2.29).*

Let \mathbf{y}_t^E be the optimal inventory position decisions for $EV_t(\mathbf{Z}_t)$; let \mathbf{x}_t^E be the inventory position state variables generated when using these allocations. Define

$$r_{nt}^E = y_{nt}^E - x_{nt}^E \quad \text{for all } n \in \mathcal{L} \text{ and } t \in \mathcal{T}$$

to be the optimal $EV_t(\mathbf{Z}_t)$ allocation amounts. We know $y_{nt}^E \geq x_{nt}^E$, so $r_{nt}^E \geq 0$. We can use the vector of \mathbf{r} values to construct a feasible solution \mathbf{y} for any set of demands, since the feasibility of allocation decisions depends only on inventories at the SNS and RSSs and the transportation capacities. So we can find the expected cost of implementing the EV solution, called the Expectation of the

Expected Value solution (EEV) as in Birge and Louveaux [Birge & Louveaux, 1997], which is given by

$$\begin{aligned}
EEV_t(\mathbf{Z}_t) &= E\left[\sum_{\tau=1}^T C_\tau(\mathbf{x}_\tau + \mathbf{r}_\tau^E, \mathbf{q}_\tau, \mathbf{D}_\tau)\right]. \\
\text{where } x_{n,t+1} &= x_{nt} + r_{nt}^E - S_{nt} \quad \text{for all } n \in \mathcal{L} \text{ and } t = 1, \dots, T-1 \\
q_{n,t+1} &= q_{nt} + D_{nt} - S_{nt} \quad \text{for all } n \in \mathcal{P} \text{ and } t = 1, \dots, T-1 \\
S_{nt} &= \min\{x_{nt} + r_{nt}^E, q_{nt} + D_{nt}, a_{nt}\} \quad \text{for all } n \in \mathcal{P} \text{ and } t \in \mathcal{T} \\
S_{mt} &= \sum_{n \in \mathcal{L}} u_{mn} S_{nt} \quad \text{for all } m \in \mathcal{S} \text{ and } t \in \mathcal{T}.
\end{aligned}$$

Note that $DP_t(\mathbf{Z}_t) \leq EEV_t(\mathbf{Z}_t)$ because the $EEV_t(\mathbf{Z}_t)$ is the expected cost of using some feasible solution to the same problem for which $DP_t(\mathbf{Z}_t)$ finds the minimum. From Jensen's inequality know that $EV_t(\mathbf{Z}_t) \leq WS_t(\mathbf{Z}_t)$. We state this result as part of the proposition below.

Proposition 2.2.3. $EV_t(\mathbf{Z}_t) \leq WS_t(\mathbf{Z}_t) \leq DP_t(\mathbf{Z}_t) \leq EEV_t(\mathbf{Z}_t)$.

2.2.3 Truncated Cumulative Approximation

The EV solution presented above is feasible for problem (2.25), but Proposition 2.2.3 shows that the expected cost of using this solution is larger than the cost of solving the dynamic program. In practice, the gap is quite large unless the demand random variables have extremely low variances. In this section, our goal is to construct a much better method for making feasible decisions. We shift our focus from thinking about the total cost that would be incurred over the time horizon to considering how the best decisions may be made in each period.

In the first hours after an anthrax attack, patient demands at the PODs will be highly unpredictable and inventory shortages are possible. Because of this uncertainty, a good inventory policy should call for small, frequent shipments to be sent to the PODs. This would help ensure that no inventory is wasted at a POD that sees lower-than-expected demands so additional supplies may be sent quickly to assist PODs with higher demands. Each shipment would only be expected to cover a small portion of a location's demand, so it is not unreasonable to consider a myopic allocation policy. In a myopic allocation policy, we make allocation decisions in each time period by optimizing inventory decisions over just a few time periods, rather than considering the full horizon. We can repeat this process in a rolling horizon manner, re-solving the problem in each time period to review our decisions based on further information.

We also need to ensure that we can efficiently solve the problem in each time period. As we discussed earlier, the presence of service capacities increases the complexity of our problem by preventing us from writing it in terms of cumulative quantities. We can significantly simplify the problem if we approximate the number of people served over some time horizon by its upper bound, as given in inequality 2.10. We will call this rolling horizon, myopic approximation approach of constructing a solution the Truncated Cumulative Approximation (TCA).

Let us assume that the length of our time horizon will be $k + 1$ time periods. For the best performance, one must look over the full time required for inventory to be shipped to the SNS and travel to a POD, so we will always set $k \geq \tau_0 + \max \tau_m : m \in \mathcal{R} + 2$. So, in each period t , we will calculate a solution that approximately minimizes the expected costs for periods $t, t+1, \dots, t+k$. We define

the approximate number served in periods $t, \dots, t + j$ to be

$$\tilde{S}_{n,t,t+j} = \min \left\{ a_{n,t,t+j}, q_{nt} + D_{n,t,t+j}, x_{n,t-\tau_n+j} + r_{n,t,t+j} \right\}$$

for $j \leq k$. Since $\tilde{S}_{n,t,t+j}$ is a function of $D_{n,t,t+j}$, it is a random variable. Notice that for $j = 0$, $\tilde{S}_{n,t,t} = S_{nt}$. We now rewrite problem (2.25) with these two modifications

$$\begin{aligned} \hat{V}_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}) &= \min \sum_{t'=t}^{t+k} C_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}) & (2.30) \\ \text{such that} \quad y_{nt'} &\geq x_{nt'} \quad \text{for all } n \in \mathcal{L}, t' = t, \dots, t+k \\ \sum_{n \in \mathcal{L}} u_{mn}(y_{nt'} - q_{nt'}) &\leq x_{m,t'-\tau_m} - q_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} \\ &\quad \text{for all } m \in \mathcal{S}, t' = t, \dots, t+k \\ \sum_{n \in \mathcal{L}} u_{mn}(y_{nt'} - x_{nt'}) &\leq p_{mt'} \quad \text{for all } m \in \mathcal{S}, t' = t, \dots, t+k \\ y_{0t'} - x_{0t'} &\leq p_{t'}^0 \quad \text{for all } t' = t, \dots, t+k \\ x_{n,t'+1} &= y_{nt'} - \tilde{S}_{nt'} \quad \text{for all } n \in \mathcal{L}, t' = t, \dots, t+k-1 \\ q_{n,t'+1} &= q_{nt'} + D_{nt} - \tilde{S}_{nt'} \quad \text{for all } n \in \mathcal{L}, t' = t, \dots, t+k-1. \end{aligned}$$

This formulation includes constraints that rely on $D_{nt'}$ and $\tilde{S}_{nt'}$, whose values are unknown for $t' = t, \dots, t+k$, and on $x_{nt'}$ and $q_{nt'}$, whose values are unknown for $t' = t+1, \dots, t+k$. Hence, the problem cannot be solved in its current form. But we can rewrite the problem in terms of on-hand inventory at the upper echelons and use shipment quantities as our decision variables.

Recall that $D_{nt'}$ is the cumulative demand random variable for location n for periods t through t' , $f_{nt'}$ is the associated probability mass function, $\mathcal{D}_{nt'}$ is the

range of $D_{ntt'}$, and r_{nt} is the amount of inventory shipped to location n in period t , so $y_{nt} = x_{nt} + r_{nt}$. Let $r_{ntt'} = r_{nt} + \dots + r_{nt'}$ be the total amount shipped in periods t through t' , and let $a_{ntt'} = a_{nt} + \dots + a_{nt'}$ be the total service capacity for periods t through t' . Let x_{mt}^O be the inventory on-hand at location $m \in \mathcal{S}$ at the beginning of time period t .

For the SNS or an RSS $m \in \mathcal{S}$ in period t we showed in equation (2.3) that the on-hand inventory is given by

$$x_{mt}^O = x_{m,t-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn}(x_{nt} + S_{n,t-\tau_m,t-1}).$$

To update this value in subsequent periods, we can use the following equation for $t' = t, \dots, t+k-1$ and $m \in \mathcal{S}$:

$$x_{m,t'+1}^O = x_{mt'}^O + r_{m,t'-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn}r_{nt'}.$$

Using the on-hand inventory and shipment decision variables, we can rewrite the constraints above as follows

$$\begin{aligned} y_{nt'} &\geq x_{nt'} \Rightarrow r_{nt'} \geq 0 \quad n \in \mathcal{L}, t' = t, \dots, t+k \\ \sum_{n \in \mathcal{L}} u_{mn}(y_{nt'} - q_{nt'}) &\leq x_{m,t'-\tau_m} - q_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} \\ &\Rightarrow \sum_{n \in \mathcal{L}} u_{mn}r_{nt'} \leq x_{mt'}^O \quad m \in \mathcal{S}, t' = t, \dots, t+k \\ \sum_{n \in \mathcal{L}} u_{mn}(y_{nt'} - x_{nt'}) &\leq p_{mt'} \Rightarrow \sum_{n \in \mathcal{L}} u_{mn}r_{nt'} \leq p_{mt'} \quad m \in \mathcal{S}, t' = t, \dots, t+k \\ y_{0t'} - x_{0t'} &\leq p_{t'}^0 \Rightarrow r_{0t'} \leq p_{t'}^0 \quad t' = t, \dots, t+k \end{aligned}$$

These constraints no longer rely on any unknown variables. We have not included a replacement for the final constraint, which showed how to update the values of $q_{nt'}$, nor have we included a method for updating $x_{nt'}$, because the new formulation does not rely on these variables for $t' > t$. Thus, we must modify the cost function so that it relies only on the state variables for period t , the on-hand inventories upstream, and the shipment decision variables. The cost functions rely on the $(y_{nt} - q_{nt})$ for all $n \in \mathcal{L}$. We can rewrite this expression for the PODs, $n \in \mathcal{P}$, as

$$y_{nt'} - q_{nt'} = (x_{nt}^O + r_{ntt'} - S_{nt,t'-1}) - (q_{nt} + D_{nt,t'-1} - S_{nt,t'-1}) = x_{nt}^O + r_{ntt'} - q_{nt} - D_{nt,t'-1}. \quad (2.31)$$

For the upper echelon locations, $m \in \mathcal{S}$, we can write

$$\begin{aligned} y_{mt'} - q_{mt'} &= \left(x_{mt'}^O + r_{m,t'-\tau_m,t'} + \sum_{n \in \mathcal{L}} u_{mn} y_{nt'} \right) - \sum_{n \in \mathcal{L}} u_{mn} q_{nt'} \\ &= x_{mt'}^O + r_{m,t'-\tau_m,t'} + \sum_{n \in \mathcal{L}} u_{mn} (y_{nt'} - q_{nt'}) \end{aligned} \quad (2.32)$$

$$= x_{mt'}^O + r_{m,t'-\tau_m,t'} + \sum_{n \in \mathcal{L}} u_{mn} (x_{nt}^O + r_{ntt'} - q_{nt} - D_{nt,t'-1}) \quad (2.33)$$

For periods $t' > t$, we rewrite the cost function in terms of the on-hand inventories and shipments:

$$\begin{aligned}
C_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}) &= h_{0,t'+\tau_0}(y_{0t'} - q_{0t'}) - h_{0,t'+\tau_0}E[D_{0,t',t'+\tau_0-1}] \\
&+ \sum_{m \in \mathcal{R}} \left((h_{m,t'+\tau_m} - h_{0t'}) (y_{mt'} - q_{mt'}) - h_{m,t'+\tau_m} E[D_{m,t',t'+\tau_m-1}] \right) \\
&+ \sum_{n \in \mathcal{P}} \left(E[h_{nt'}(y_{nt'} - S_{nt'}) + f_{nt'}^B(q_{nt'} + D_{nt'} - S_{nt'})] \right) \\
&- \sum_{m \in \mathcal{R}} u_{mn} h_{mt'} (y_{nt'} - q_{nt'})).
\end{aligned}$$

Applying equation (2.32) for the SNS, grouping the RSS terms, setting $y_{nt'} - S_{nt'} = x_{nt}^O + r_{ntt'} - \tilde{S}_{ntt'}$ for $n \in \mathcal{P}$, and setting $q_{nt'} + D_{nt'} = q_{nt} + D_{ntt'} - \tilde{S}_{ntt'}$ for $n \in \mathcal{P}$:

$$\begin{aligned}
C_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}) &= h_{0,t'+\tau_0}(x_{0t'}^O + r_{0,t'-\tau_0,t'} - E[D_{0,t',t'+\tau_0-1}]) \\
&+ \sum_{m \in \mathcal{R}} \left((h_{m,t'+\tau_m} + h_{0,t'+\tau_0} - h_{0t'}) (y_{mt'} - q_{mt'}) - h_{m,t'+\tau_m} E[D_{m,t',t'+\tau_m-1}] \right) \\
&+ \sum_{n \in \mathcal{P}} \left(E[h_{nt'}(x_{nt}^O + r_{ntt'} - \tilde{S}_{ntt'}) + f_{nt'}^B(q_{nt} + D_{ntt'} - \tilde{S}_{ntt'})] \right) \\
&- \sum_{m \in \mathcal{R}} u_{mn} h_{mt'} (y_{nt'} - q_{nt'})).
\end{aligned}$$

Applying equation (2.32) to the RSS terms, grouping the POD terms, and applying equation (2.31):

$$\begin{aligned}
C_{t'}(\mathbf{y}_{t'}, \mathbf{q}_{t'}) &= h_{0,t'+\tau_0}(x_{0t'}^O + r_{0,t'-\tau_0,t'} - E[D_{0,t',t'+\tau_0-1}]) \\
&+ \sum_{m \in \mathcal{R}} \left((h_{m,t'+\tau_m} + h_{0,t'+\tau_0} - h_{0t'}) (x_{mt'}^O + r_{m,t'-\tau_m,t'}) - h_{m,t'+\tau_m} E[D_{m,t',t'+\tau_m-1}] \right) \\
&+ \sum_{n \in \mathcal{P}} \left(E[h_{nt'}(x_{nt}^O + r_{ntt'} - \tilde{S}_{ntt'}) + f_{nt'}^B(q_{nt} + D_{ntt'} - \tilde{S}_{ntt'})] \right) \\
&+ (h_{0,t+\tau_0} - h_{0t} + \sum_{m \in \mathcal{R}} u_{mn} (h_{m,t+\tau_m} - h_{mt})) (x_{nt}^O + r_{ntt'} - q_{nt} - E[D_{nt,t'-1}]) \\
&= \bar{C}_{t'}((x_{0t'}^O, x_{1t'}^O, \dots, x_{Mt'}^O), (x_{M+1,t}^O, \dots, x_{M+N,t}^O), \mathbf{q}_t, \mathbf{r}_{t'}^{\text{past}}).
\end{aligned}$$

The expectations included in the period t' cost function are taken over the cumulative demand random variables $D_{ntt'}$. If we used the exact cumula-

tive numbers of people served $S_{nt'}$, we would need to take expectations over $D_{nt}, D_{n,t+1}, \dots, D_{nt'}$, for a total of $t' + 1 - t$ nested expectations for each POD $n \in \mathcal{P}$. Replacing $S_{nt'}$ with $\tilde{S}_{nt'}$ allows us to consider only one expectation for each POD $n \in \mathcal{P}$, significantly reducing the computational complexity of the cost calculation.

We use the newly defined cost function $\bar{C}_t(\cdot)$ in formulating the truncated cumulative approximation problem:

$$\begin{aligned} \hat{V}_t(\mathbf{Z}_t) &= \min \sum_{t'=t}^{t+k} \bar{C}_{t'}((x_{0t'}^O, x_{1t'}^O, \dots, x_{Mt'}^O), (x_{M+1,t}^O, \dots, x_{M+N,t}^O), \mathbf{q}_t, \mathbf{r}_t^{\text{past}}) \quad (2.34) \\ \text{such that} \quad &r_{nt'} \geq 0 \quad \text{for all } n \in \mathcal{L} \text{ and } t' = t, \dots, t+k \\ &\sum_{n \in \mathcal{L}} u_{mn} r_{nt'} \leq x_{mt'}^O \quad \text{for all } m \in \mathcal{S}; \quad t' = t, \dots, t+k \\ &\sum_{n \in \mathcal{L}} u_{mn} r_{nt'} \leq p_{mt'} \quad \text{for all } m \in \mathcal{S}; \quad t' = t, \dots, t+k \\ &r_{0t'} \leq p_{t'}^0, \quad \text{for all } t' = t, \dots, t+k \\ &x_{m,t'+1}^O = x_{mt'}^O + r_{m,t'-\tau_m} - \sum_{n \in \mathcal{L}} u_{mn} r_{nt'} \quad \text{for } m \in \mathcal{S}, t' = t, \dots, t+k-1. \end{aligned}$$

To make decisions using the TCA model, we must re-solve problem (2.34) during each time period in a rolling horizon manner, as described by Algorithm 1 below.

Algorithm 1:

1. Initialize the state variables $\mathbf{Z}_1 = (\mathbf{x}_1, \mathbf{q}_1)$.
2. For $t \in \mathcal{T}$:
 - a. Solve problem (2.34) to find $\hat{V}_t(\mathbf{Z}_t)$ and obtain allocation decisions \mathbf{y}_t .

- b. Observe the simulated demand: $d_{nt} = D_{nt}$ for $n \in \mathcal{P}$.
 - c. Calculate the number of patients served, $S_{nt} = S_{nt}(y_{nt}, q_{nt}, d_{nt})$ for $n \in \mathcal{P}$.
 - d. Calculate the costs incurred in time period t from making decisions \mathbf{y}_t : $\bar{C}_t(\mathbf{y}_t, \mathbf{q}_t, \mathbf{d}_t)$.
 - e. Update state variables: $\mathbf{Z}_{t+1} = g_t(\mathbf{Z}_t, \mathbf{y}_t, \mathbf{d}_t)$.
3. Return $TCA(\mathbf{Z}_1, \mathbf{d}) = \sum_{t=1}^T \bar{C}_t(\mathbf{y}_t, \mathbf{q}_t, \mathbf{d}_t)$.

The expected cost of using the approximate myopic solution to make allocation decisions for the full time horizon can be estimated by calculating $TCA(\mathbf{Z}_1, \mathbf{D}_i)$ for $i = 1, \dots, I$, where I is a large number and setting

$$ETCA(\mathbf{Z}_1) = \frac{1}{I} \sum_{i=1}^I TCA(\mathbf{Z}_1, \mathbf{D}_i). \quad (2.35)$$

It is clear that $ETCA(\mathbf{Z}_1) \geq DP_1(\mathbf{Z}_1)$ since the decisions made by the TCA solution method are feasible, but not necessarily optimal for the original problem (2.25). However, we expect to find that $ETCA(\mathbf{Z}_1) < EEV_1(\mathbf{Z}_1)$, since the TCA approach responds to the current inventory position over time and, as we discussed at the beginning of this section, the modifications that we made to problem (2.25) are reasonable ones. In the following section, we present a different method for estimating $DP_1(\mathbf{Z}_1)$ and constructing near-optimal solutions to problem (2.25).

2.2.4 Lagrangian Relaxation Method

The SNS emergency response network is a three-echelon divergent supply chain. Many people have studied inventory allocation in multi-echelon sup-

ply chains; we review a few of the relevant papers here and discuss how we can build on some of these methods to construct a Lagrangian Relaxation inventory policy. [Clark & Scarf, 1960] published one of the first papers on multi-echelon inventory theory. They studied an N -echelon serial system under periodic review, with discounted holding and backorder costs. They showed that it is possible to decompose this system into a set of small dynamic programs which can be solved for an optimal solution. They also introduced the balance assumption, which states that it will never be desirable to redistribute inventory among the retailers; that is, the inventory will never become “imbalanced.” However, [Dogru *et al.*, 2005] showed that there are a number of scenarios when the balance assumption is inadequate.

[Eppen & Schrage, 1981] studied a two echelon divergent inventory system, and, under a number of assumptions, they derived an approximately optimal policy for inventory allocation. [Federgruen & Zipkin, 1984a] and [Federgruen & Zipkin, 1984b] extended Eppen and Schrage’s model by relaxing many of the restrictive assumptions. They showed that a myopic allocation policy is optimal at all of the lower-echelon locations when these locations are allowed to return inventory to the upper echelons.

In recent years, [Kunnumkal & Topaloglu, 2008] and [Kunnumkal & Topaloglu, 2010] have improved on the work of Federgruen and Zipkin. They allow negative inventory allocations, but they include Lagrange multipliers to penalize negative allocation decisions. In [Kunnumkal & Topaloglu, 2008], they present a method for obtaining the Lagrange multipliers by solving convex optimization problems, and in [Kunnumkal & Topaloglu, 2010] they present a much faster linear programming method that produced lower expected cost policies

for a set of their problems.

Most multi-echelon supply chain research has focused on serial systems or two echelon divergent networks. However, [Diks & de Kok, 1998] and [Diks & de Kok, 1999] have studied general N -echelon divergent systems in which all of the locations in the network can hold stock. Under the balance assumption, they show that order-up-to policies are optimal at all locations in the system, and they show how to find the optimal levels and allocation policies for each location [Diks & de Kok, 1998]. [Diks & de Kok, 1999] presents a faster, easier method for making allocation decisions under similar assumptions. [Dogru *et al.*, 2004] perform similar studies using discrete demand distributions. [Graves & Willems, 2008] determine optimal base stock levels for general complex supply chains under nonstationary demand and [Neale & Willems, 2009] extends this work for practical applications, but both papers assume that the demands have known upper bounds.

Imbalance may be a very significant problem in an emergency response scenario, since demand is highly uncertain, so the work by Diks, de Kok, and Dogru *et al.* does not translate directly to this problem. Instead, we build on the work from Kunnumkal and Topaloglu, who include Lagrange multipliers to discourage negative inventory shipments. They have shown that their method produces tight lower bounds on the value function and good simulated performance in a two-echelon uncapacitated distribution network. In this section, we extend their work to the three-echelon capacitated SNS distribution network. Our goals are twofold: to calculate a lower bound that is tighter than the Wait-and-See bound and to construct an inventory policy with lower expected cost than the TCA allocation method.

In this subsection, we will use a four stage process to relax problem (2.25) and decompose it into a set of smaller, tractable subproblems. The four steps are: (1) use Lagrangian relaxation to replace the transportation constraints and some of the nonnegative shipment constraints with costs; (2) rewrite the problem in terms of inventory position instead of on-hand inventory and patient queues; (3) use Lagrangian relaxation to relax some of the service constraints; and (4) decompose the problem by location into a set of single variable dynamic programs.

Before we propose the first relaxation, observe that in the dynamic program (2.25) it is never advantageous to ship inventory from the SNS to RSS m after time period $T - \tau_m - 1$ since such shipments will not reach the RSS in time to be sent on to a POD for use before the end of the time horizon. Similarly, it is never useful to ship inventory to the SNS after time period $T - \tau_0 - \min_{m \in \mathcal{R}} \tau_m - 2$. Therefore, without loss of generality, we can add the following constraints to the dynamic program (2.25):

$$y_{mt} = x_{mt} \quad \text{for } m \in \mathcal{R} \text{ and } t = T - \tau_m, \dots, T, \quad (2.36)$$

$$y_{0t} = x_{0t} \quad \text{for } t = T - \tau_0 - \min_{m \in \mathcal{R}} \tau_m - 1, \dots, T. \quad (2.37)$$

We will include these two constraints in subsequent statements of the dynamic program (2.25).

Let us now define two sets of Lagrange multipliers:

λ_{nt} = the Lagrange multiplier associated with the constraint $y_{nt} \geq x_{nt}$ for

$$n \in \mathcal{R} \cup \mathcal{P} \text{ and } t = 1, \dots, T - \tau_n - 1, \text{ and}$$

μ_{nt} = the Lagrange multiplier associated with the constraint

$$\sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) \leq p_{mt} \text{ for } m \in \mathcal{S} \text{ and } t = 1, \dots, T - \tau_n.$$

For notational convenience, let $\lambda_{nt} = 0$ for $n \in \mathcal{R}$ and $t = 1, \dots, T - \tau_n - 1$, $\lambda = \{\lambda_{nt} : n \in \mathcal{R} \cup \mathcal{P}, t \in \mathcal{T}\}$, and $\mu = \{\mu_{mt} : m \in \mathcal{S}, t \in \mathcal{T}\}$. For $n \in \mathcal{R} \cup \mathcal{P}$, let $\mu_{nt}^U = \sum_{m=0}^M u_{mn} \mu_{mt}$ and $h_{nt}^U = \sum_{m=0}^M u_{mn} h_{mt}$ be the parameters associated with the upper echelon location that supplies location n . Let $\tau = \min_{m \in \mathcal{R}} \tau_m$.

We now relax problem (2.25) by removing the constraints associated with the multipliers defined above and adding Lagrange terms to the cost function. We also include the additional constraints stated above in equations (2.36)-(2.37):

$$\begin{aligned} V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) = & \min \left\{ \sum_{n \in \mathcal{L}} C_{nt}(y_{nt}, q_{nt}) + E[V'_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \right. \\ & - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt}(y_{nt} - x_{nt}) \\ & \left. - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) \right) \right\} \end{aligned} \quad (2.38)$$

such that

$$\begin{aligned} x_{0t} &\leq y_{0t} \leq x_{0t} + p_t^0 \\ \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - q_{nt}) &\leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \quad (2.39) \\ &\text{for all } m \in \mathcal{S} \end{aligned}$$

$$y_{mt} = x_{mt} \quad \text{for } m \in \mathcal{R} \text{ if } t = T - \tau_m, \dots, T$$

$$y_{0t} = x_{0t} \quad \text{if } t = T - \tau_0 - \tau - 1, \dots, T.$$

As in Kunnumkal and Topaloglu [Kunnumkal & Topaloglu, 2008], we can show

that problem (2.38) provides a lower bound on the problem (2.25).

Lemma 2.2.4. *Suppose the Lagrange multipliers λ and μ are nonnegative. Then*

$$V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) \leq V_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}).$$

Proof. Let \mathbf{y}_t^* be the optimal solution to $V_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}})$. Then we know $\mathbf{y}_t^* \geq \mathbf{x}_t$ and $p_{mt} \geq \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt})$ for $m \in \mathcal{S}$. So for any nonnegative λ_{nt} and μ_{mt} , we see that

$$\begin{aligned} \lambda_{nt}(y_{nt}^* - x_{nt}) &\geq 0 \quad \text{for } n \in \mathcal{R} \cup \mathcal{P} \text{ and} \\ \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt}^* - x_{nt}) \right) &\geq 0 \quad \text{for } m \in \mathcal{R}. \end{aligned}$$

Consequently

$$- \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt}(y_{nt} - x_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) \right) \leq 0. \quad (2.40)$$

Note also that \mathbf{y}_t^* is feasible for $V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$.

In time period T we see that

$$\begin{aligned}
V'_T(\mathbf{x}_T^{\text{past}}, \mathbf{q}_T^{\text{past}}, \mathbf{d}_T^{\text{past}} | \lambda, \mu) &= \min \left\{ C_T(\mathbf{y}_T, \mathbf{q}_T) - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (y_{nT} - x_{nT}) \right. \\
&\quad \left. - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mT} - \sum_{n \in \mathcal{L}} u_{mn} (y_{nT} - x_{nT}) \right) \right\} \\
&\leq C_T(\mathbf{y}_T^*, \mathbf{q}_T) - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (y_{nT}^* - x_{nT}) \\
&\quad - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mT} - \sum_{n \in \mathcal{L}} u_{mn} (y_{nT}^* - x_{nT}) \right) \\
&\leq C_T(\mathbf{y}_T^*, \mathbf{q}_T) \\
&= V_T(\mathbf{x}_T^{\text{past}}, \mathbf{q}_T^{\text{past}}, \mathbf{d}_T^{\text{past}}),
\end{aligned}$$

where the first inequality follows from the definition of feasibility and the second from inequality (2.40).

Let us assume that $V'_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu) \leq V_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}})$. Then following the same reasoning,

$$\begin{aligned}
V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) &= \min \left\{ C_t(\mathbf{y}_t, \mathbf{q}_t) - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (y_{nt} - x_{nt}) \right. \\
&\quad \left. - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (y_{nt} - x_{nt}) \right) \right. \\
&\quad \left. + E[V'_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \right\} \\
&\leq C_t(\mathbf{y}_t^*, \mathbf{q}_t) - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (y_{nt}^* - x_{nt}) \\
&\quad - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (y_{nt}^* - x_{nt}) \right) \\
&\quad + E[V'_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \\
&\leq C_t(\mathbf{y}_t^*, \mathbf{q}_t) + E[V_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \\
&= V_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}).
\end{aligned}$$

□

A good allocation policy should never set $y_{nt} > a_{nt}$ for any POD n in any period t , unless this is required by another constraint in the problem. Inventory above the value of a_{nt} cannot be used at a POD; setting $y_{nt} > a_{nt}$ guarantees additional holding cost. Since a_{nt} is known, the lead time from the RSSs to the PODs is 0, and holding costs are lower at the RSSs than at the PODs, it would always be better to hold inventory above a_{nt} units back at the RSS when possible. Now that the constraint $y_{nt} \geq x_{nt}$ has been removed, there are no constraints to require y_{nt} to be larger than a_{nt} ; the only reason why an optimal solution might set $y_{nt} > a_{nt}$ would be if the newly introduced Lagrangian cost terms made such a decision cost-effective. Suppose that, for POD n in period t , $y_{nt} \geq a_{nt}$. The cost function terms that include y_{nt} are

$$E\left[h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt})) + f_{nt}^B(q_{nt} + D_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt}))\right] - h_{nt}^U y_{nt} \\ + \lambda_{nt} y_{nt} + \mu_{nt}^U y_{nt} + V_t'(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu).$$

The number of patients served is given by $S_{nt}(y_{nt}, q_{nt}, D_{nt}) = \min(a_{nt}, q_{nt} + D_{nt})$, so we can remove these terms:

$$h_{nt} y_{nt} - h_{nt}^U y_{nt} + \lambda_{nt} y_{nt} + \mu_{nt}^U y_{nt} + V_t'(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu).$$

The only impact that y_{nt} will have on future costs is through $x_{n,t+1} = y_{nt} - S_{nt}$. The cost terms containing $x_{n,t+1}$ or y_{nt} are

$$h_{nt} y_{nt} - h_{nt}^U y_{nt} + \lambda_{nt} y_{nt} + \mu_{nt}^U y_{nt} + \lambda_{n,t+1} x_{n,t+1} - \mu_{nt}^U x_{n,t+1} \\ = h_{nt} y_{nt} - h_{nt}^U y_{nt} + \lambda_{nt} y_{nt} + \mu_{nt}^U y_{nt} + \lambda_{n,t+1} (y_{nt} - S_{nt}) - \mu_{nt}^U (y_{nt} - S_{nt}).$$

Dropping the S_{nt} terms leaves us with

$$(h_{nt} - h_{nt}^U + \lambda_{nt} + \mu_{nt}^U + \lambda_{n,t+1} - \mu_{nt}^U)y_{nt}.$$

So, if

$$h_{nt} - h_{nt}^U - \lambda_{nt} + \mu_{nt}^U + \lambda_{n,t+1} - \mu_{n,t+1}^U \geq 0, \quad (2.41)$$

then any optimal solution will force y_{nt} to be as small as the constraints allow. Since we assumed $y_{nt} \geq a_{nt}$, an optimal solution would set $y_{nt} = a_{nt}$. Since $y_{nt} \geq a_{nt}$ implies that $y_{nt} = a_{nt}$, we see that in general $y_{nt} \leq a_{nt}$ when equation (2.41) holds. Thus, we can add the constraint $y_{nt} \leq a_{nt}$ to problem (2.38) without any loss of generality to get the new dynamic program

$$\begin{aligned} V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) = & \min \left\{ \sum_{n \in \mathcal{L}} C_{nt}(y_{nt}, q_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) \right) \right. \\ & - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt}(y_{nt} - x_{nt}) \\ & \left. + E[V'_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \right\} \end{aligned} \quad (2.42)$$

such that

$$x_{0t} \leq y_{0t} \leq x_{0t} + p_t^0$$

$$\sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - q_{nt}) \leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \text{ for } m \in \mathcal{S}$$

$$y_{nt} \leq a_{nt} \text{ for } n \in \mathcal{P}$$

$$y_{mt} = x_{mt} \text{ for } m \in \mathcal{R} \text{ if } t = T - \tau_m, \dots, T$$

$$y_{0t} = x_{0t} \text{ if } t = T - \tau_0 - \tau - 1, \dots, T.$$

We make two observations about problem (2.42) in the following lemma:

Lemma 2.2.5. Let $V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$ be defined as in problem (2.42) and suppose that for all PODs $n \in \mathcal{P}$, $h_{nt} - h_{nt}^U - \lambda_{nt} + \mu_{nt}^U + \lambda_{n,t+1} - \mu_{n,t+1}^U \geq 0$. Then

1. $S_{nt}(y_{nt}, q_{nt}, D_{nt}) = \min(y_{nt}, q_{nt} + D_{nt})$ for all $n \in \mathcal{P}$, and
2. in any optimal solution to problem (2.42), for each $n \in \mathcal{P}$, at most one of x_{nt} and q_{nt} may be positive (i.e., $x_{nt} > 0 \Rightarrow q_{nt} = 0$ and $q_{nt} > 0 \Rightarrow x_{nt} = 0$).

Proof. Consider problem (2.42). Since $y_{nt} \leq a_{nt}$, the number of people served can be written as

$$S_{nt}(y_{nt}, q_{nt}, D_{nt}) = \min(y_{nt}, q_{nt} + D_{nt}, a_{nt}) = \min(y_{nt}, q_{nt} + D_{nt}),$$

which proves (1).

Now, assume that $q_{n1} = 0$ or $x_{n1} = 0$ for all PODs n . In any optimal solution, in periods $t = 2, \dots, T$ we have

$$x_{nt} = y_{n,t-1} - S_{n,t-1}(y_{n,t-1}, q_{n,t-1}, D_{n,t-1}) = \max(0, y_{n,t-1} - q_{n,t-1} - D_{n,t-1})$$

and similarly

$$q_{nt} = q_{n,t-1} + D_{n,t-1} - S_{n,t-1}(y_{n,t-1}, q_{n,t-1}, D_{n,t-1}) = \max(0, -(y_{n,t-1} - q_{n,t-1} - D_{n,t-1})),$$

so at least one of the x_{nt} and q_{nt} must be 0. This proves (2). \square

We now use Lemma 2.2.5 to reduce the number of state variables required by problem (2.42). Define the inventory position variables for points (0) and (2) in each time period to be

$$\bar{\mathbf{x}}_t = \mathbf{x}_t - \mathbf{q}_t \quad \text{and}$$

$$\bar{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{q}_t,$$

respectively, with $\bar{\mathbf{x}}_t^{\text{past}}$ defined as in section 2.1.3. Let the transition functions be given by

$$\bar{\mathbf{x}}_{m,t+1}^{\text{past}} = \bar{\mathbf{y}}_{mt} - D_{mt} \quad \text{for } m \in \mathcal{P}; \quad (2.43)$$

$$\bar{\mathbf{x}}_{m,t+1}^{\text{past}} = \left(\bar{x}_{m,t-\tau_m+1}, \bar{x}_{m,t-\tau_m+2}, \dots, \bar{x}_{mt}, \bar{\mathbf{y}}_{mt} - \sum_{n \in \mathcal{P}} u_{mn} D_{nt} \right) \quad \text{for } m \in \mathcal{S}. \quad (2.44)$$

The vector $\bar{\mathbf{x}}_t^{\text{past}}$, which gives the current and past inventory positions at the beginning of each time period, will replace $\mathbf{x}_t^{\text{past}}$ and $\mathbf{q}_t^{\text{past}}$ in the vector of state variables in our new problem formulation. The new decision variable vector will be $\bar{\mathbf{y}}_t$, the inventory positions after inventory has been shipped in period t , but before patients have been served at the PODs.

We will define modified cost functions in terms of the new variables $\bar{\mathbf{x}}_t^{\text{past}}$ and $\bar{\mathbf{y}}_t$. Let

$$\begin{aligned}
C'_{nt}(\bar{y}_{nt}) &= \sum_{d \in \mathcal{D}_{nt}} f_{nt}(d)[h_{nt}(\bar{y}_{nt} - d)^+ + f_{nt}^B(d - \bar{y}_{nt})] \\
&\quad - \sum_{m \in \mathcal{R}} u_{mn} h_{mt} \bar{y}_{nt} \text{ for } n \in \mathcal{P}; \\
C'_{nt}(\bar{y}_{nt}) &= (h_{n,t+\tau_n} - h_{0t}) \bar{y}_{nt} - h_{n,t+\tau_n} E[D_{n,t,t+\tau_n-1}] \text{ for } n \in \mathcal{S}; \\
C'_{0t}(\bar{y}_{0t}) &= h_{0,t+\tau_0} \bar{y}_{0t} - h_{0,t+\tau_0} E[D_{0,t,t+\tau_0-1}]; \text{ and} \\
C'_t(\bar{\mathbf{y}}_t) &= \sum_{n \in \mathcal{L}} C'_{nt}(\bar{y}_{nt}).
\end{aligned}$$

We can now state a modified dynamic program in terms of the new variables, with the new value function, $\bar{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$:

$$\begin{aligned}
\bar{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) &= \min \left\{ C'_t(\bar{\mathbf{y}}_t) + E[\bar{V}_{t+1}(\bar{\mathbf{x}}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (\bar{y}_{nt} - \bar{x}_{nt}) \right. \\
&\quad \left. - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (\bar{y}_{nt} - \bar{x}_{nt}) \right) \right\} \quad (2.45)
\end{aligned}$$

$$\begin{aligned}
\text{such that} \quad & \sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nt} \leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \text{ for all } m \in \mathcal{S} \\
& \bar{x}_{0t} \leq \bar{y}_{0t} \leq p_t^0 + \bar{x}_{0t} \\
& \bar{y}_{nt} \leq a_{nt} + \bar{x}_{nt} \text{ for all } n \in \mathcal{P} \\
& \bar{y}_{nt} \leq a_{nt} \text{ for all } n \in \mathcal{P} \\
& \bar{y}_{mt} = \bar{x}_{mt} \text{ for } m \in \mathcal{R} \text{ if } t = T - \tau_m, \dots, T \\
& \bar{y}_{0t} = \bar{x}_{0t} \text{ if } t = T - \tau_0 - \tau - 1, \dots, T.
\end{aligned}$$

We can prove that the new problem (2.45) is equivalent to the earlier relaxed problem (2.42):

Proposition 2.2.6. *Let $V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$ be defined as in problem (2.42) and let $\bar{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$ be defined as in problem (2.45). Suppose that the Lagrange*

multipliers λ_t and μ_t are nonnegative and $h_{nt} - h_{nt}^U - \lambda_{nt} + \mu_{nt}^U + \lambda_{n,t+1} - \mu_{n,t+1}^U \geq 0$. Then for all time periods $t \in \mathcal{T}$,

$$V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) = \bar{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$$

Proof. Let $\bar{x}_{nt} = x_{nt} - q_{nt}$ and $\bar{y}_{nt} = y_{nt} - q_{nt}$ for all n and t . We begin by writing the cost function at each POD n for period t , $C_{nt}(y_{nt}, q_{nt})$ in terms of \bar{y}_{nt} . Recall the cost function definition for the PODs, $n \in \mathcal{P}$, from equation (2.22),

$$\begin{aligned} C_{nt}(y_{nt}, q_{nt}) &= E[h_{nt}(y_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt})) + f_{nt}^B(D_{nt} + q_{nt} - S_{nt}(y_{nt}, q_{nt}, D_{nt}))] \\ &\quad - \sum_{m \in \mathcal{R}} u_{mn} h_{mt}(y_{nt} - q_{nt}). \end{aligned}$$

Applying Lemma 2.2.5, we substitute $\min(y_{nt}, q_{nt} + D_{nt})$ for $S_{nt}(y_{nt}, q_{nt}, D_{nt})$ to get

$$\begin{aligned} C_{nt}(y_{nt}, q_{nt}) &= E[h_{nt}(y_{nt} - q_{nt} - D_{nt})^+ + f_{nt}^B(D_{nt} + q_{nt} - y_{nt})] \\ &\quad - \sum_{m \in \mathcal{R}} u_{mn} h_{mt}(y_{nt} - q_{nt}) \\ &= \sum_{d \in \mathcal{D}_{nt}} f_{nt}(d) [h_{nt}(\bar{y}_{nt} - d)^+ + f_{nt}^B(d - \bar{y}_{nt})] \\ &\quad - \sum_{m \in \mathcal{R}} u_{mn} h_{mt} \bar{y}_{nt} \\ &= C'_{nt}(\bar{y}_{nt}). \end{aligned}$$

The cost functions at the SNS and RSSs can also be written in terms of \bar{y}_{nt} :

$$\begin{aligned}
C_{0t}(y_{0t}, q_{0t}) &= h_{0,t+\tau_0}(y_{0t} - q_{0t}) - h_{0,t+\tau_0}E[D_{0,t,t+\tau_0-1}] \\
&= h_{0,t+\tau_0}\bar{y}_{0t} - h_{0,t+\tau_0}E[D_{0,t,t+\tau_0-1}] \\
&= C'_{0t}(\bar{y}_{0t}) \quad \text{and} \\
C_{mt}(y_{mt}, q_{mt}) &= (h_{m,t+\tau_m} - h_{0t})(y_{mt} - q_{mt}) - h_{m,t+\tau_m}E[D_{m,t,t+\tau_m-1}] \quad \text{for } m \in \mathcal{R} \\
&= (h_{m,t+\tau_m} - h_{0t})\bar{y}_{mt} - h_{m,t+\tau_m}E[D_{m,t,t+\tau_m-1}] \\
&= C'_{mt}(\bar{y}_{mt}).
\end{aligned}$$

Thus $C_t(\mathbf{y}_t, \mathbf{q}_t) = C'_t(\mathbf{y}_t - \mathbf{q}_t) = C'_t(\bar{\mathbf{y}}_t)$.

We now show that the feasible regions of both dynamic programs are equivalent. We first write problem (2.42) in terms of $\bar{\mathbf{y}}_t$ and $\bar{\mathbf{x}}_t$. Notice that for all $n \in \mathcal{L}$

$$y_{nt} - x_{nt} = (y_{nt} - q_{nt}) - (x_{nt} - q_{nt}) = \bar{y}_{nt} - \bar{x}_{nt},$$

and

$$y_{nt} = x_{nt} \Leftrightarrow y_{nt} - q_{nt} = x_{nt} - q_{nt} \Leftrightarrow \bar{y}_{nt} = \bar{x}_{nt}.$$

Thus, substituting into problem (2.45),

$$\begin{aligned}
V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu) = & \min \left\{ \sum_{n \in \mathcal{L}} C'_{nt}(\bar{y}_{nt}) + E[V'_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \right. \\
& - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt}(\bar{y}_{nt} - \bar{x}_{nt}) \\
& \left. - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn}(\bar{y}_{nt} - \bar{x}_{nt}) \right) \right\} \quad (2.46)
\end{aligned}$$

such that

$$\begin{aligned}
& \sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nt} \leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \text{ for all } m \in \mathcal{S} \\
& \bar{x}_{0t} \leq \bar{y}_{0t} \leq p_t^0 + \bar{x}_{0t} \\
& y_{nt} \leq a_{nt} \text{ for all } n \in \mathcal{P} \\
& \bar{y}_{mt} = \bar{x}_{mt} \text{ for } m \in \mathcal{R} \text{ if } t = T - \tau_m, \dots, T \\
& \bar{y}_{0t} = \bar{x}_{0t} \text{ if } t = T - \tau_0 - \min_{m \in \mathcal{R}} \tau_m - 1, \dots, T.
\end{aligned}$$

Only the one constraint still relies on y_{nt} instead of \bar{y}_{nt} . If q_{nt} is subtracted from both sides of the inequality, we get $\bar{y}_{nt} \leq a_{nt} - q_{nt}$. From Lemma 2.2.5 we know that if $q_{nt} > 0$ then $x_{nt} = 0$ so $\bar{x}_{nt} = x_{nt} - q_{nt} = -q_{nt}$. So we see that

$$a_{nt} - q_{nt} = \begin{cases} a_{nt} & \text{if } q_{nt} = 0 \\ a_{nt} + \bar{x}_{nt} & \text{if } q_{nt} > 0 \end{cases}.$$

Thus, we can replace the constraint $y_{nt} \leq a_{nt}$ in problem (2.46) with the two constraints

$$\begin{aligned}
\bar{y}_{nt} & \leq a_{nt} \\
\bar{y}_{nt} & \leq a_{nt} + \bar{x}_{nt}.
\end{aligned}$$

It remains only to show that the state transition functions are equivalent for the two problems. We see that

$$\begin{aligned}
\bar{x}_{n,t+1} &= \bar{y}_{nt} - D_{nt} \\
&= y_{nt} - q_{nt} - D_{nt} \\
&= (y_{nt} - S_{nt}) - (q_{nt} + D_{nt} - S_{nt}) \\
&= x_{n,t+1} - q_{n,t+1}.
\end{aligned}$$

Thus, problem (2.42) now has the same cost function, constraints, and state transition function as problem (2.45), so the two are equivalent. \square

Problem (2.45) has fewer state variables and a simpler cost function than our original problem (2.25), but we need to relax one more set of constraints before we can decompose the problem into a set of single variable problems. We now define the Lagrange multipliers

γ_{nt} = the Lagrange multiplier associated with the constraint $\bar{y}_{nt} \leq a_{nt} + \bar{x}_{nt}$ for $n \in \mathcal{P}$ and $t \in \mathcal{T}$.

Relaxing the constraint $\bar{y}_{nt} \leq a_{nt} + \bar{x}_{nt}$ gives us

$$\begin{aligned}
\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma) = & \min \left\{ C'_t(\bar{\mathbf{y}}_t) + E[\tilde{V}_{t+1}(\bar{\mathbf{x}}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu, \gamma)] \right. \\
& - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (\bar{y}_{nt} - \bar{x}_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} (p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (\bar{y}_{nt} - \bar{x}_{nt})) \\
& \left. - \sum_{n \in \mathcal{P}} \gamma_{nt} (a_{nt} + \bar{x}_{nt} - \bar{y}_{nt}) \right\} \tag{2.47}
\end{aligned}$$

such that

$$\begin{aligned}
\sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nt} &\leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \text{ for all } m \in \mathcal{S} \\
\bar{x}_{0t} &\leq \bar{y}_{0t} \leq p_t^0 + \bar{x}_{0t} \\
\bar{y}_{nt} &\leq a_{nt} \text{ for all } n \in \mathcal{P} \\
\bar{y}_{mt} &= \bar{x}_{mt} \text{ for } m \in \mathcal{R} \text{ if } t = T - \tau_m, \dots, T \\
\bar{y}_{0t} &= \bar{x}_{0t} \text{ if } t = T - \tau_0 - \min_{m \in \mathcal{R}} \tau_m - 1, \dots, T.
\end{aligned}$$

If the optimal solution of $\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma)$ has $\bar{x}_{nt} \geq 0$ for all $n \in \mathcal{P}$ and $t \in \mathcal{T}$, then the optimal solution of problem (2.47) will equal that of problem (2.45) since the constraint $\bar{y}_{nt} \leq a_{nt} + \bar{x}_{nt}$ is completely dominated by the constraint $\bar{y}_{nt} \leq a_{nt}$. However, if some PODs have positive queues for some time periods, then $\bar{x}_{nt} < 0$ and the problems are not equivalent. The new problem (2.47) allows a service capacity of $a_{nt} + (\text{queue length}) = a_{nt} - (\bar{x}_{nt})^-$ for POD n in period t , which means that more patients may be served than the true capacity allows. For large queues this relaxation could be problematic, although the convex increasing backorder costs will help to discourage large queues by charging high penalties. The Lagrange multipliers will help to ensure that the system does not allow too many patients to be served over the planned service capacity.

We now state a modified version of Lemma 2.2.4 to show that problem (2.47) provides a lower bound on our first relaxed problem (2.38).

Lemma 2.2.7. *Suppose the Lagrange multipliers λ, μ and γ are nonnegative, $h_{nt} - h_{nt}^U - \lambda_{nt} + \mu_{nt}^U + \lambda_{n,t+1} - \mu_{n,t+1}^U \geq 0$. Then*

$$\tilde{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma) \leq V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}} \mathbf{d}_t^{\text{past}} | \lambda, \mu) \leq V_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}}).$$

Proof. The right-most inequality is a re-statement of Lemma 2.2.4, which is true since we still satisfy the same nonnegative multiplier condition. We only need to prove the left inequality. We showed that $V'_t(\mathbf{x}_t^{\text{past}}, \mathbf{q}_t^{\text{past}} \mathbf{d}_t^{\text{past}} | \lambda, \mu) = \bar{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$ in Proposition 2.2.6. We will show that

$$\tilde{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma) \leq \bar{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu). \quad (2.48)$$

We complete the proof using an induction argument. Inequality (2.48) is easy to show for time period T . Suppose now that it is true for periods $t = t + 1, \dots, T$. We will prove that it also holds in period t . Define $\bar{\mathbf{x}}_t^{\text{past}} = \mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}$. Let $\bar{\mathbf{y}}_t^*$ be the optimal solution to $\bar{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu)$. Then certainly $\bar{\mathbf{y}}_t^*$ is feasible for $\tilde{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma)$, which means that $\bar{y}_{nt}^* \leq a_{nt} + \bar{x}_{nt}$ for all PODs $n \in \mathcal{P}$. So for any POD n and any $\gamma_{nt} \geq 0$, we see that

$$-\gamma_{nt}(a_{nt} + \bar{x}_{nt} - \bar{y}_{nt}^*) \leq 0. \quad (2.49)$$

So

$$\begin{aligned}
\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma) &= \min \left\{ C'_t(\bar{\mathbf{y}}_t) + E[\tilde{V}_{t+1}(\bar{\mathbf{x}}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu, \gamma)] \right. \\
&\quad - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (\bar{y}_{nt} - \bar{x}_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (\bar{y}_{nt} - \bar{x}_{nt}) \right) \\
&\quad \left. - \sum_{n \in \mathcal{P}} \gamma_{nt} (a_{nt} + \bar{x}_{nt} - \bar{y}_{nt}) \right\} \\
&\leq C'_t(\bar{\mathbf{y}}_t^*) + E[\tilde{V}_{t+1}(\bar{\mathbf{x}}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu, \gamma)] \\
&\quad - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (\bar{y}_{nt}^* - \bar{x}_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (\bar{y}_{nt}^* - \bar{x}_{nt}) \right) \\
&\quad - \sum_{n \in \mathcal{P}} \gamma_{nt} (a_{nt} + \bar{x}_{nt} - \bar{y}_{nt}^*) \\
&\leq C'_t(\bar{\mathbf{y}}_t^*) + E[\tilde{V}_{t+1}(\bar{\mathbf{x}}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu)] \\
&\quad - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt} (\bar{y}_{nt}^* - \bar{x}_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn} (\bar{y}_{nt}^* - \bar{x}_{nt}) \right) \\
&= \bar{V}_t(\mathbf{x}_t^{\text{past}} - \mathbf{q}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu),
\end{aligned}$$

where the first inequality follows from the definition of feasibility, and the second follows from equation (2.49) and our inductive assumption. \square

We have now relaxed the problem sufficiently to decompose it into single variable subproblems, which will be presented in the next subsection.

Decomposition

We first define the subproblems and then state and prove the main decomposition theorem of this section. We will also present a simple algorithm for solving the decomposed problem.

For PODs $n \in \mathcal{P}$, let

$$\begin{aligned}
\psi_{nt}(\bar{x}_{nt}|\lambda_{nt}, \gamma_{nt}) &= \min \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt}a_{nt} \right. \\
&\quad \left. + E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})] \right\} \\
\text{such that} \quad &\bar{y}_{nt} \leq a_{nt}.
\end{aligned} \tag{2.50}$$

Recall that $\mu_{nt}^U = \sum_{m \in \mathcal{R}} u_{mn} \mu_{mt}$. The following lemma shows that solving problem (2.50) is straightforward.

Lemma 2.2.8. *For $n \in \mathcal{P}$ the optimal solution to $\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma)$ is given by*

$$\bar{y}_{nt}^* = \operatorname{argmin}_{\bar{y}_{nt} \leq a_{nt}} \left\{ ([\gamma_{nt} - \lambda_{nt} + \mu_{nt}^U] - [\gamma_{n,t+1} - \lambda_{n,t+1} + \mu_{n,t+1}^U])\bar{y}_{nt} + C'_{nt}(\bar{y}_{nt}) \right\} \tag{2.51}$$

and

$$\begin{aligned}
\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) &= (\gamma_{nt} - \lambda_{nt} + \mu_{nt}^U)(\bar{y}_{nt}^* - \bar{x}_{nt}) + C'_{nt}(\bar{y}_{nt}^*) - \gamma_{nt}a_{nt} + \sum_{t'=t+1}^T \left[C'_{nt'}(\bar{y}_{nt'}^*) \right. \\
&\quad \left. - \gamma_{nt'}a_{nt'} + (\gamma_{nt'} - \lambda_{nt'} + \mu_{nt'}^U)(\bar{y}_{nt'}^* - \bar{y}_{n,t'-1}^* + E[D_{n,t'-1}]) \right]. \tag{2.52}
\end{aligned}$$

Proof. This proof is based on a similar proof in Kunnumkal and Topaloglu [Kunnumkal & Topaloglu, 2010].

We begin by showing that equation (2.51) holds.

$$\begin{aligned}
\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) &= \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt}a_{nt} \right. \\
&\quad \left. + E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})] \right\}.
\end{aligned}$$

Substituting for $\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})$ gives us

$$\begin{aligned} \psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) = & \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt}a_{nt} \right. \\ & + E \left[\min_{\bar{y}_{n,t+1} \leq a_{n,t+1}} \left\{ C'_{n,t+1}(\bar{y}_{n,t+1}) - \gamma_{n,t+1}a_{n,t+1} \right. \right. \\ & + (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})(\bar{y}_{n,t+1} - \bar{x}_{n,t+1}) \\ & \left. \left. + E[\psi_{n,t+2}(\bar{y}_{n,t+1} - D_{n,t+1})] \right\} \right] \Bigg\}. \end{aligned}$$

Since $\bar{x}_{n,t+1} = \bar{y}_{nt} - D_{nt}$ we get

$$\begin{aligned} \psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) = & \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt}a_{nt} \right. \\ & + E \left[\min_{\bar{y}_{n,t+1} \leq a_{n,t+1}} \left\{ C'_{n,t+1}(\bar{y}_{n,t+1}) - \gamma_{n,t+1}a_{n,t+1} \right. \right. \\ & + (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})(\bar{y}_{n,t+1} - (\bar{y}_{nt} - D_{nt})) \\ & \left. \left. + E[\psi_{n,t+2}(\bar{y}_{n,t+1} - D_{n,t+1})] \right\} \right] \Bigg\}. \end{aligned}$$

Rearranging terms gives us

$$\begin{aligned} \psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) = & -(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})\bar{x}_{nt} - \gamma_{nt}a_{nt} \\ & + E \left[\min_{\bar{y}_{n,t+1} \leq a_{n,t+1}} \left\{ C'_{n,t+1}(\bar{y}_{n,t+1}) - \gamma_{n,t+1}a_{n,t+1} \right. \right. \\ & \left. \left. + (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})(\bar{y}_{n,t+1} + D_{nt}) + E[\psi_{n,t+2}(\bar{y}_{n,t+1} - D_{n,t+1})] \right\} \right] \\ & + \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})\bar{y}_{nt} - (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})\bar{y}_{nt} \right\}. \end{aligned}$$

This proves equation (2.51), since only the last line of the final equation depends on y_{nt} ; the first terms are independent of y_{nt} .

We now prove that equation (2.52) holds using induction. In time period T we see that

$$\begin{aligned}
\psi_{nT}(\bar{x}_{nT}|\lambda, \mu, \gamma) &= \min_{\bar{y}_{nT} \leq a_{nT}} \left\{ C'_{nT}(\bar{y}_{nT}) + (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})(\bar{y}_{nT} - \bar{x}_{nT}) - \gamma_{nT}a_{nT} \right. \\
&= \min_{\bar{y}_{nT} \leq a_{nT}} \left\{ C'_{nT}(\bar{y}_{nT}) + (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})\bar{y}_{nT} \right\} \\
&\quad \left. - (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})\bar{x}_{nT} - \gamma_{nT}a_{nT} \right\}.
\end{aligned}$$

From the definition of \bar{y}_{nt}^* , we see that

$$\begin{aligned}
\psi_{nT}(\bar{x}_{nT}|\lambda, \mu, \gamma) &= C'_{nT}(\bar{y}_{nT}^*) + (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})\bar{y}_{nT}^* \\
&\quad - (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})\bar{x}_{nT} - \gamma_{nT}a_{nT}.
\end{aligned}$$

Now suppose that equation (2.52) holds in all periods $t + 1, \dots, T$. We will show that it also holds in period t by substituting for $\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})$.

$$\begin{aligned}
\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) &= \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt}a_{nt} \right. \\
&\quad \left. + E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})] \right\}.
\end{aligned}$$

Substituting for $\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})$ and letting $\bar{x}_{n,t+1} = \bar{y}_{nt} - D_{nt}$ gives us

$$\begin{aligned}
\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) &= \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt}a_{nt} \right. \\
&\quad \left. + E \left[(\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})(\bar{y}_{n,t+1}^* - (\bar{y}_{nt} - D_{nt})) + C'_{n,t+1}(\bar{y}_{n,t+1}^*) \right. \right. \\
&\quad \left. \left. - \gamma_{n,t+1}a_{n,t+1} + \sum_{t'=t+2}^T [C'_{nt'}(\bar{y}_{nt'}^*) - \gamma_{nt'}a_{nt'} \right. \right. \\
&\quad \left. \left. + (\mu_{nt'}^U + \gamma_{nt'} - \lambda_{nt'})\bar{y}_{nt'}^* - \bar{y}_{n,t'-1}^* + E[D_{n,t'-1}]] \right] \right\}.
\end{aligned}$$

Rearranging terms, we see that

$$\begin{aligned}\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) = & -(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})\bar{x}_{nt} - \gamma_{nt}a_{nt} - \sum_{t'=t+2}^T (\mu_{nt'}^U + \gamma_{nt'} - \lambda_{nt'})\bar{y}_{n,t'-1}^* \\ & + \sum_{t'=t+1}^T \left[C'_{nt'}(\bar{y}_{nt'}^*) - \gamma_{nt'}a_{nt'} + (\mu_{nt'}^U + \gamma_{nt'} - \lambda_{nt'})\bar{y}_{nt'}^* + E[D_{n,t'-1}] \right] \\ & \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})\bar{y}_{nt} - (\gamma_{n,t+1} - \lambda_{n,t+1} + \mu_{n,t+1}^U)\bar{y}_{nt} \right\}.\end{aligned}$$

Finally, we use the definition to substitute for \bar{y}_{nt}^* :

$$\begin{aligned}\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma) = & -(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})\bar{x}_{nt} - \gamma_{nt}a_{nt} - \sum_{t'=t+2}^T (\mu_{nt'}^U + \gamma_{nt'} - \lambda_{nt'})\bar{y}_{n,t'-1}^* \\ & + \sum_{t'=t+1}^T \left[C'_{nt'}(\bar{y}_{nt'}^*) - \gamma_{nt'}a_{nt'} + (\mu_{nt'}^U + \gamma_{nt'} - \lambda_{nt'})\bar{y}_{nt'}^* + E[D_{n,t'-1}] \right] \\ & C'_{nt}(\bar{y}_{nt}^*) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})\bar{y}_{nt}^* - (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})\bar{y}_{nt}^*.\end{aligned}$$

If we rearrange these terms slightly, we see that we have satisfied equation (2.52), so by induction, the equation holds for all time periods. \square

As we showed in Section 2.1.1, the echelon inventory position at RSS m at the beginning of period t minus any in-transit inventory is given by $\bar{x}_{m,t-\tau_m} - D_{m,t-\tau_m,t-1}$, and we refer to this as the RSS's echelon on-hand inventory. For RSS m to provide all of its PODs with their desired inventories would require a total of $\sum_{n \in \mathcal{P}(m)} \bar{y}_{nt}^*$ units of inventory. If the RSS had an on-hand echelon inventory of $\bar{x}'_{mt} = \bar{x}_{m,t-\tau_m} - D_{m,t-\tau_m,t-1}$ and $\bar{x}'_{mt} < \sum_{n \in \mathcal{P}(m)} \bar{y}_{mt}^*$, then the additional cost incurred by the PODs served by RSS m in period t would be

$$\begin{aligned} \Delta_{mt}(\bar{x}'_{mt}|\lambda, \mu, \gamma) = & \min \sum_{n \in \mathcal{P}} u_{mn} \left[(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt}) - (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1}) \right] (\bar{y}_{nt} - \bar{y}_{nt}^*) \\ & + C'_{nt}(\bar{y}_{nt}) - C'_{nt}(\bar{y}_{nt}^*) \end{aligned} \quad (2.53)$$

$$\begin{aligned} \text{such that } & \sum_{n \in \mathcal{P}} u_{mn} \bar{y}_{nt} \leq \bar{x}'_{mt} \\ & \bar{y}_{nt} \leq a_{nt} \text{ for } n \in \mathcal{P} \end{aligned}$$

and $\Delta_{mt}(\bar{x}'_{mt}|\lambda, \mu, \gamma) = 0$ for $t > T$. We can use this expression to define the single variable problems for the RSSs. For $m \in \mathcal{R}$ let

$$\begin{aligned} \psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma) = & \min \left\{ C'_{mt}(\bar{y}_{mt}) + E[\Delta_{m,t+\tau_m}(\bar{x}_{mt} - D_{m,t+\tau_m-1}|\lambda, \mu, \gamma)] \right. \\ & \left. - \lambda_{mt}(\bar{y}_{mt} - \bar{x}_{mt}) - \mu_{mt} p_{mt} + E[\psi_{m,t+1}(\bar{y}_{mt} - D_{mt}|\lambda, \mu, \gamma)] \right\}. \end{aligned} \quad (2.54)$$

$$\bar{y}_{mt} = \bar{x}_{mt} \quad \text{if } t = T - \tau_m, \dots, T.$$

Notice that, for RSS m in periods $t = 1, \dots, T - \tau_m - 1$, the optimal solution $\bar{y}_{mt} = \bar{y}_{mt}^*$ is independent of \bar{x}_{mt} . However, in periods $t = T - \tau_m, \dots, T$, no shipments should be sent to the RSS so the optimal solution is $\bar{y}_{mt}^* = \bar{x}_{mt}$. Since this solution is known, we analytically calculate the costs for periods $t = T - \tau_m, \dots, T$ in the following lemma:

Lemma 2.2.9. *For periods $t = T - \tau_m + 1, \dots, T$,*

$$\psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma) = -h_{0,t,T} \bar{x}_{mt} + \sum_{t'=t}^T (h_{0,t'+1,T} E[D_{mt'}] - \mu_{mt'} p_{mt'}). \quad (2.55)$$

In period $T - \tau_m$,

$$\begin{aligned}
\psi_{m,T-\tau_m}(\bar{x}_{m,T-\tau_m}|\lambda,\mu,\gamma) &= (h_{mT} - h_{0,T-\tau_m,T})\bar{x}_{m,T-\tau_m} + E[\Delta_{mT}(\bar{x}_{m,T-\tau_m} - D_{m,T-\tau_m,T-1})] \\
&\quad - h_{mT}E[D_{m,T-\tau_m,T-1}] \\
&\quad + \sum_{t'=t}^T (h_{0,t'+1,T}E[D_{mt'}] - \mu_{mt'}p_{mt'}). \tag{2.56}
\end{aligned}$$

Proof. By induction. □

The next step is to define a penalty function $\Delta_{0t}(\cdot)$ that accounts for the cost incurred by the system when the SNS does not have sufficient inventory to give the RSSs all of the inventory that they require in period t . To ensure that this function can be calculated efficiently, we must assume that the lead times to the RSSs are identical, that is, there exists $\tau \geq 0$ such that $\tau_m = \tau$ for all $m \in \mathcal{R}$. We can, however, maintain some control over the times when shipments are sent by choosing the transportation capacities carefully. For example, if we set $p_{mt} = 0$ and $p_{m,t+1} > 0$ for some $m \in \mathcal{R}$ and some period t , then we will try to avoid shipping to RSS m in period $t - \tau$ and instead wait to ship until period $t + 1 - \tau$.

We noted above that $\bar{y}_{mt}^* = \bar{x}_{mt}$ for $t = T - \tau, \dots, T$ and \bar{y}_{mt}^* is independent of \bar{x}_{mt} for $t = 1, \dots, T - \tau - 1$. So, for periods $t = 1, \dots, T - \tau - 1$, let the echelon inventory minus the in-transit inventory at the SNS in period t be $\bar{x}'_{0t} = \bar{x}_{0,t-\tau_0} + D_{0,t-\tau_0,t-1}$, and define

$$\begin{aligned}
\Delta_{0t}(\bar{x}'_{0t}|\lambda, \mu, \gamma) &= \min \sum_{m \in \mathcal{R}} \left(C'_{mt}(\bar{y}_{mt}) - C'_{mt}(\bar{y}_{mt}^*) - \lambda_{mt}(\bar{y}_{mt} - \bar{y}_{mt}^*) \right. \\
&\quad \left. + E[\psi_{m,t+1}(\bar{y}_{mt} - D_{mt}|\lambda, \mu, \gamma) \right. \\
&\quad \left. - \psi_{m,t+1}(\bar{y}_{mt}^* - D_{mt}|\lambda, \mu, \gamma)] \right) \tag{2.57}
\end{aligned}$$

such that $\sum_{m \in \mathcal{R}} \bar{y}_{mt} \leq \bar{x}'_{0t}.$

For $t \geq T - \tau$ we will not incur any penalty, since no new allocations are made to the RSS after this time, and we have $\Delta_{0t}(\bar{x}'_{0t}|\lambda, \mu, \gamma) = 0.$

We use the penalty function $\Delta_{0t}(\bar{x}'_{0t}|\lambda, \mu, \gamma)$ to state our final single variable problem for the SNS. Let

$$\begin{aligned}
\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma) &= \min \left\{ C'_{0t}(\bar{y}_{0t}) + E[\Delta_{0,t+\tau_0}(\bar{x}_{0t} - D_{0,t+\tau_0-1}|\lambda, \mu, \gamma)] - \mu_{0t}p_{0t} \right. \\
&\quad \left. + E[\psi_{0,t+1}(\bar{y}_{0t} - D_{0t}|\lambda, \mu, \gamma)] \right\} \tag{2.58}
\end{aligned}$$

such that $\bar{x}_{0t} \leq \bar{y}_{0t} \leq \bar{x}_{0t} + p_t^0$

$\bar{y}_{0t} = \bar{x}_{0t}$ for $t = T - \tau_0 - \tau - 1, \dots, T.$

We can state a lemma similar to Lemma (2.2.9) that provides an analytical solution for $\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma)$ for periods $t = T - \tau_0 - \tau - 1, \dots, T:$

Lemma 2.2.10. *For periods $t = T - \tau_0 - \tau, \dots, T,$*

$$\begin{aligned}
\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma) &= h_{0,t+\tau_0,T}\bar{x}_{0t} - \sum_{t'=t}^T (\mu_{0t'}p_{0t'} \\
&\quad + h_{0,t'+\tau_0}E[D_{0,t',t'+\tau_0-1}] + h_{0,t'+\tau_0+1,T}E[D_{0t'}]). \tag{2.59}
\end{aligned}$$

For period $t = T - \tau_0 - \tau - 1$,

$$\begin{aligned} \psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma) &= h_{0,t+\tau_0,T}\bar{x}_{0t} + E[\Delta_{0,t+\tau_0}(\bar{x}_{0t} - D_{0t,t+\tau_0-1}|\lambda, \mu, \gamma)] - \sum_{t'=t}^T (\mu_{0t'} p_{0t'} \\ &\quad + h_{0,t'+\tau_0} E[D_{0,t',t'+\tau_0-1}] + h_{0,t'+\tau_0+1,T} E[D_{0t'}]). \end{aligned} \quad (2.60)$$

Proof. By induction. □

Finally, we will define the total expected costs incurred in periods t through T by the $\psi_{nt}(\cdot)$ functions in period t to be

$$\psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma) = \sum_{n \in \mathcal{L}} \psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma). \quad (2.61)$$

Before stating the main proposition, we make several observations about equation (2.61). First, notice that the new function $\psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma)$ only relies on current inventory positions, $\bar{\mathbf{x}}_t$, while the value function $\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma)$ defined in problem (2.47) relies on both current and past inventory positions as well as previous demands. This difference exists because the times at which costs are charged have been modified. In the definition of the new function $\psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma)$, the expected cost of not having enough inventory at the SNS or an RSS, $m \in \mathcal{S}$, in period $t + \tau_m$ is charged in period t . The value function $\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma)$ charges this cost in period $t + \tau_m$. Also, the new function $\psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma)$ does not charge any costs for failing to have sufficient inventory at the SNS or an RSS for periods $t = 1, \dots, \tau_m$. These costs are unavoidable consequences of the initial conditions, but the original relaxed problem (2.47) includes all of the costs incurred at each location n starting in period τ_n even if these costs are unavoidable given the initial conditions of the problem.

Proposition 2.2.11. Suppose that $b_{nt} = b_t$ and $h_{nt} = h_t^P$ for all PODs n , and that $h_{mt} = h_t^R$ for all RSSs m , with $h_t^P > h_t^R > h_{0t}$. When $\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma)$ is defined as in problem (2.47), then

$$\begin{aligned} \tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma) &= \psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma) \\ &\quad + \sum_{m \in S} \sum_{t'=t}^{t+\tau_m-1} E[\Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1}|\lambda, \mu, \gamma)]. \end{aligned} \quad (2.62)$$

Proof. We will prove that equation (2.62) holds using induction. Consider time period T . We know that $\bar{y}_{mT} = \bar{x}_{mT}$ for $m \in S$ from the equality constraint and $\psi_{n,T+1}(\bar{x}_{n,T+1}) = 0$ for all locations $n \in \mathcal{L}$, so we substitute the definitions to see that

$$\begin{aligned} &\sum_{n=0}^{N+M} \psi_{nT}(\bar{x}_{nT}) + \sum_{m \in S} \sum_{t'=T}^{T+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1}|\lambda, \mu, \gamma) \\ &= \sum_{n \in \mathcal{P}} \min_{\bar{y}_{nT} \leq a_{nT}} \left\{ C'_{nT}(\bar{y}_{nT}) + (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})(\bar{y}_{nT} - \bar{x}_{nT}) - \gamma_{nT} a_{nT} \right\} \\ &\quad + \sum_{m \in \mathcal{R}} \left\{ C'_{mT}(\bar{x}_{mT}) + E[\Delta_{m,T+\tau_m}(\bar{x}_{mT} - D_{mT,T+\tau_m-1}|\lambda, \mu, \gamma)] \right. \\ &\quad \left. - \lambda_{mT}(\bar{x}_{mT} - \bar{x}_{mT}) - \mu_{mT} p_{mT} \right\} \\ &\quad + \left\{ C'_{0T}(\bar{x}_{0T}) + E[\Delta_{0,T+\tau_0}(\bar{x}_{0T} - D_{0T,T+\tau_0-1}|\lambda, \mu, \gamma)] - \mu_{0T} p_{0T} \right\} \\ &\quad + \sum_{m \in S} \sum_{t'=T}^{T+\tau_m-1} E[\Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1}|\lambda, \mu, \gamma)]. \end{aligned} \quad (2.63)$$

For $m \in \mathcal{S}$ and $t' > T$ we know that $\Delta_{mt'}(\cdot) = 0$, so the only Δ terms that remain are the Δ_{mT} terms. Further, $E[\Delta_{mt}(\bar{x}_{m,T-\tau_m} - d_{m,T-\tau_m,T-1})] = \Delta_{mt}(\bar{x}_{m,T-\tau_m} - d_{m,T-\tau_m,T-1})$. Canceling the other Δ terms and substituting y_{nT}^* into the first line leaves

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nT}(\bar{x}_{nT}) + \sum_{m \in \mathcal{S}} \sum_{t'=T}^{T+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma) \\
&= \sum_{n \in \mathcal{P}} \left\{ C'_{nT}(\bar{y}_{nT}^*) + (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})(\bar{y}_{nT}^* - \bar{x}_{nT}) - \gamma_{nT} a_{nT} \right\} \\
&+ \sum_{m \in \mathcal{R}} \left\{ C'_{mT}(\bar{x}_{mT}) - \lambda_{mT}(\bar{x}_{mT} - \bar{x}_{mT}) - \mu_{mT} p_{mT} \right\} \\
&+ \left\{ C'_{0T}(\bar{x}_{0T}) - \mu_{0T} p_{0T} \right\} \\
&+ \sum_{m \in \mathcal{S}} \Delta_{mT}(\bar{x}_{m,T-\tau_m} - d_{m,T-\tau_m,T-1} | \lambda, \mu, \gamma). \tag{2.64}
\end{aligned}$$

Noting that $\Delta_{0T}(\cdot) = 0$ and expanding the $\Delta_{mT}(\cdot)$ terms for $m \in \mathcal{S}$ gives

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nT}(\bar{x}_{nT}) + \sum_{m \in \mathcal{S}} \sum_{t'=T}^{T+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma) \\
&= \sum_{n \in \mathcal{P}} \left\{ C'_{nT}(\bar{y}_{nT}^*) + (\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})(\bar{y}_{nT}^* - \bar{x}_{nT}) - \gamma_{nT} a_{nT} \right\} \\
&+ \sum_{m \in \mathcal{R}} \left\{ C'_{mT}(\bar{x}_{mT}) - \lambda_{mT}(\bar{x}_{mT} - \bar{x}_{mT}) - \mu_{mT} p_{mT} \right\} \\
&+ \left\{ C'_{0T}(\bar{x}_{0T}) - \mu_{0T} p_{0T} \right\} \\
&+ \sum_{m \in \mathcal{R}} \left[\min_{\bar{y}_{nT} \leq a_{nT}; \sum_n u_{mn} \bar{y}_{nT} \leq \bar{x}_{m,T-\tau_m} - d_{m,T-\tau_m,T-1}} \sum_{n \in \mathcal{P}} u_{mn} (C'_{nT}(\bar{y}_{nT}) - C'_{nT}(\bar{y}_{nT}^*)) \right. \\
&\left. + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{y}_{nt}^*) \right].
\end{aligned}$$

Canceling the y_{nt}^* terms leaves

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nT}(\bar{x}_{nT}) + \sum_{m \in \mathcal{S}} \sum_{t'=T}^{T+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma) \\
&= \min \left\{ \sum_{n \in \mathcal{L}} C'_{nT}(\bar{y}_{nT}) + \sum_{n \in \mathcal{P}} [(\mu_{nT}^U + \gamma_{nT} - \lambda_{nT})(\bar{y}_{nT} - \bar{x}_{nT}) - \gamma_{nT} a_{nT}] \right. \\
&\quad \left. - \sum_{m \in \mathcal{R}} [\lambda_{mT}(\bar{y}_{mT} - \bar{x}_{mT}) + \mu_{mT} p_{mT}] - \mu_{0T} p_{0T} \right\} \\
&\quad \text{s.t. } \bar{y}_{mT} = \bar{x}_{mT} \quad \text{for } m \in \mathcal{S}; \\
&\quad \bar{y}_{nT} \leq a_{nT} \quad \text{for } n \in \mathcal{P}; \\
&\quad \sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nT} \leq \bar{x}_{m,T-\tau_m} - d_{m,T-\tau_m,T-1} \quad \text{for } m \in \mathcal{S} \\
&= \tilde{V}_T(\bar{\mathbf{x}}_T^{\text{past}}, \mathbf{d}_T^{\text{past}} | \lambda, \mu, \gamma). \tag{2.65}
\end{aligned}$$

We now assume that the proposition holds in time period $t+1$. We will show that it also holds in time period t . Suppose that $t < T - \tau_0 - \tau - 1$. We first group all of the $\Delta_{mt}(\cdot)$ terms together in the last expression:

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nt}(\bar{x}_{nt}) + \sum_{m \in \mathcal{S}} \sum_{t'=t}^{t+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma) \\
&= \sum_{n \in \mathcal{P}} \min_{\bar{y}_{nt} \leq a_{nt}} \left\{ C'_{nt}(\bar{y}_{nt}) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt} a_{nt} + E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt} | \lambda, \mu, \gamma)] \right\} \\
&\quad + \sum_{m \in \mathcal{R}} \min \left\{ C'_{mt}(\bar{y}_{mt}) - \lambda_{mt}(\bar{y}_{mt} - \bar{x}_{mt}) - \mu_{mt} p_{mt} + E[\psi_{m,t+1}(\bar{y}_{mt} - D_{mt} | \lambda, \mu, \gamma)] \right\} \\
&\quad + \min_{\bar{x}_{0t} \leq \bar{y}_{0t} \leq \bar{x}_{0t} + p_t^0} \left\{ C'_{0t}(\bar{y}_{0t}) - \mu_{0t} p_{0t} + E[\psi_{0,t+1}(\bar{y}_{0t} - D_{0t} | \lambda, \mu, \gamma)] \right\} \\
&\quad + \sum_{m \in \mathcal{S}} \sum_{t'=t}^{t+\tau_m} E[\Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma)].
\end{aligned}$$

We replace the Δ_{mt} terms with their definitions and note that, as in the case of $t = T$, we do not need the $E[\cdot]$ for these terms. We also substitute for y_{nt}^* for $n \in \mathcal{R} \cup \mathcal{P}$ to get

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nt}(\bar{x}_{nt}) + \sum_{m \in \mathcal{S}} \sum_{t'=t}^{t+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1}|\lambda, \mu, \gamma) \\
&= \sum_{n \in \mathcal{P}} \left\{ C'_{nt}(\bar{y}_{nt}^*) + (\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt}^* - \bar{x}_{nt}) - \gamma_{nt}a_{nt} + E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt}|\lambda, \mu, \gamma)] \right\} \\
&+ \sum_{m \in \mathcal{R}} \left\{ C'_{mt}(\bar{y}_{mt}^*) - \lambda_{mt}(\bar{y}_{mt}^* - \bar{x}_{mt}) - \mu_{mt}p_{mt} + E[\psi_{m,t+1}(\bar{y}_{mt}^* - D_{mt}|\lambda, \mu, \gamma)] \right\} \\
&+ \min_{\bar{x}_{0t} \leq \bar{y}_{0t} \leq \bar{x}_{0t} + p_t^0} \left\{ C'_{0t}(\bar{y}_{0t}) - \mu_{0t}p_{0t} + E[\psi_{0,t+1}(\bar{y}_{0t} - D_{0t}|\lambda, \mu, \gamma)] \right\} \\
&+ \sum_{m \in \mathcal{R}} \left[\min_{\bar{y}_{nt} \leq a_{nt}; \sum_n u_{mn}\bar{y}_{nt} \leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,t-1}} \sum_{n \in \mathcal{P}} u_{mn} (C'_{nt}(\bar{y}_{nt}) - C'_{nt}(\bar{y}_{nt}^*)) \right. \\
&+ \left. [(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt}) - (\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})](\bar{y}_{nt} - \bar{y}_{nt}^*) \right] \\
&+ \left[\min_{\sum_m \bar{y}_{mt} \leq \bar{x}_{0,t-\tau_0} - d_{0,t-\tau_0,t-1}} \sum_{m \in \mathcal{R}} (C'_{mt}(\bar{y}_{mt}) - C'_{mt}(\bar{y}_{mt}^*)) \right. \\
&+ \left. E[\psi_{m,t+1}(\bar{y}_{mt} - D_{mt}|\lambda, \mu, \gamma) - \psi_{m,t+1}(\bar{y}_{mt}^* - D_{mt}|\lambda, \mu, \gamma)] \right] \\
&+ \sum_{m \in \mathcal{S}} \sum_{t'=t+1}^{t+\tau_m} E[\Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1}|\lambda, \mu, \gamma)].
\end{aligned}$$

Notice that for $n \in \mathcal{P}$, $E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt})] - E[\psi_{n,t+1}(\bar{y}_{nt}^* - D_{nt})] = -(\mu_{n,t+1}^U + \gamma_{n,t+1} - \lambda_{n,t+1})(\bar{y}_{nt} - \bar{y}_{nt}^*)$. Substituting and canceling terms leaves

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nt}(\bar{x}_{nt}) + \sum_{m \in \mathcal{S}} \sum_{t'=t}^{t+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma) \\
&= \min \left\{ \sum_{n \in \mathcal{L}} C'_{nt}(\bar{y}_{nt}) + \sum_{n \in \mathcal{P}} [(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt} a_{nt}] \right. \\
&\quad - \sum_{m \in \mathcal{R}} [\lambda_{mt}(\bar{y}_{mt} - \bar{x}_{mt}) + \mu_{mt} p_{mt}] - \mu_{0t} p_{0t} \\
&\quad + \sum_{n \in \mathcal{L}} E[\psi_{n,t+1}(\bar{y}_{nt} - D_{nt} | \lambda, \mu, \gamma)] \\
&\quad \left. + \sum_{m \in \mathcal{S}} \sum_{t'=t+1}^{t+\tau_m} E[\Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma)] \right\} \\
&\text{s.t. } \bar{y}_{nt} \leq a_{nt} \quad \text{for } n \in \mathcal{P}; \\
&\quad \sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nt} \leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,T-1} \quad \text{for } m \in \mathcal{S}.
\end{aligned}$$

Applying the inductive hypothesis gives

$$\begin{aligned}
& \sum_{n=0}^{N+M} \psi_{nt}(\bar{x}_{nt}) + \sum_{m \in \mathcal{S}} \sum_{t'=t}^{t+\tau_m-1} \Delta_{mt'}(\bar{x}_{m,t'-\tau_m} - d_{m,t'-\tau_m,t'-1} | \lambda, \mu, \gamma) \\
&= \min \left\{ \sum_{n \in \mathcal{L}} C'_{nt}(\bar{y}_{nt}) + \sum_{n \in \mathcal{P}} [(\mu_{nt}^U + \gamma_{nt} - \lambda_{nt})(\bar{y}_{nt} - \bar{x}_{nt}) - \gamma_{nt} a_{nt}] \right. \\
&\quad - \sum_{m \in \mathcal{R}} [\lambda_{mt}(\bar{y}_{mt} - \bar{x}_{mt}) + \mu_{mt} p_{mt}] - \mu_{0t} p_{0t} \\
&\quad \left. + \sum_{n \in \mathcal{L}} E[\tilde{V}_{t+1}(\bar{y}_t - D_t | \lambda, \mu, \gamma)] \right\} \\
&\text{s.t. } \bar{y}_{nt} \leq a_{nt} \quad \text{for } n \in \mathcal{P}; \\
&\quad \sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nt} \leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,T-1} \quad \text{for } m \in \mathcal{S} \\
&= \tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}} | \lambda, \mu, \gamma). \tag{2.66}
\end{aligned}$$

For $T - \tau_0 - \tau - 1 \leq t < T - \tau$, the same argument holds, but we have $\bar{y}_{0t} = \bar{x}_{0t}$. For $t \geq T - \tau$, we also have set $\bar{y}_{mt} = \bar{x}_{mt}$.

□

Solving the Decomposition

We now describe a method for calculating $\psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma)$. Let \underline{M}_{nt} be the minimum value that may be taken by \bar{x}_{nt} and let \overline{M}_{nt} be the maximum value that may be taken by \bar{x}_{nt} . We can then use the following algorithm:

Algorithm 2:

1. Calculate $\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma)$ for the PODS:
 - (a) Solve the LP (2.51) to find \bar{y}_{nt}^* and $C'_{nt}(\bar{y}_{nt}^*)$ for $n \in \mathcal{P}$ and $t \in \mathcal{T}$.
 - (b) Use equation (2.52) to calculate $\psi_{nt}(\bar{x}_{nt}|\lambda, \mu, \gamma)$ for $n \in \mathcal{P}$, $t = T, \dots, 1$, and $\bar{x}_{nt} = \underline{M}_{nt}, \dots, \overline{M}_{nt}$.
2. Calculate $\Delta_{mt}(\bar{x}'_{mt}|\lambda, \mu, \gamma)$ for the RSSs:
 - (a) The smallest possible value for \bar{x}'_{mt} is $\underline{M}'_{mt} = \underline{M}_{m,t-\tau} - \max d : d \in \mathcal{D}_{m,t-\tau,t-1}$. Use marginal analysis to calculate $\Delta_{mt}(\bar{x}'_{mt}|\lambda, \mu, \gamma)$ for $m \in \mathcal{R}$, $t = \tau + 1, \dots, T$, and $\bar{x}'_{mt} = \underline{M}'_{mt}, \dots, \sum_{n \in \mathcal{L}} u_{mn} y_{nt}^*$.
 - (b) Calculate

$$E[\Delta_{m,t+\tau}(\bar{x}_{mt} - D_{m,t,t+\tau-1})] = \sum_{d \in \mathcal{D}_{m,t,t+\tau-1}} Pr(D_{m,t,t+\tau-1} = d) \Delta_{m,t+\tau}(\bar{x}_{mt} - d)$$

for $m \in \mathcal{R}$, $t = 1, \dots, T - \tau$, and $\bar{x}'_{mt} = \underline{M}'_{mt}, \dots, \sum_{n \in \mathcal{L}} u_{mn} y_{nt}^*$.

3. Calculate $\psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma)$ for the RSSs:

- (a) Use equation (2.55) to calculate $\psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma)$ for $m \in \mathcal{R}$, $t = T - \tau + 1, \dots, T$, and $\bar{x}_{mt} = \underline{M}_{mt}, \dots, \overline{M}_{mt}$.
- (b) Use equation (2.56) to calculate $\psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma)$ for $m \in \mathcal{R}$, $t = T - \tau$, and $\bar{x}_{mt} = \underline{M}_{mt}, \dots, \overline{M}_{mt}$.
- (c) For $t = T - \tau - 1, \dots, 1$ and $m \in \mathcal{R}$: (1) Solve the MIP (2.54) without the $\Delta(\cdot)$ term or the \bar{x}_{mt} terms to find \bar{y}_{mt}^* and $C'_{mt}(\bar{y}_{mt}^*) + E[\psi_{m,t+1}(\bar{y}_{mt}^* - D_{mt}|\lambda, \mu, \gamma)]$. Then, (2) calculate

$$\begin{aligned} \psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma) = & C'_{mt}(\bar{y}_{mt}^*) + E[\psi_{m,t+1}(\bar{y}_{mt}^* - D_{mt}|\lambda, \mu, \gamma)] - \lambda_{mt}(\bar{y}_{mt} + \\ & E[\Delta_{m,t+\tau_m}(\bar{x}_{mt} - D_{m,t+\tau_m-1}|\lambda, \mu, \gamma)] - \bar{x}_{mt}) - \mu_{mt}p_{mt} \end{aligned}$$

$$\text{for } \bar{x}_{mt} = \underline{M}_{mt}, \dots, \overline{M}_{mt}.$$

4. Calculate $\Delta_{0t}(\bar{x}'_{0t}|\lambda, \mu, \gamma)$:

- (a) The smallest possible value for \bar{x}'_{0t} is $\underline{M}'_{0t} = \underline{M}_{0,t-\tau_0} - \max d : d \in \mathcal{D}_{0,t-\tau_0,t-1}$. Use marginal analysis to calculate $\Delta_{0t}(\bar{x}'_{0t}|\lambda, \mu, \gamma)$ for $t = \tau_0 + 1, \dots, T - \tau - 1$ and $\bar{x}'_{0t} = \underline{M}'_{0t}, \dots, \sum_{m \in \mathcal{R}} \bar{y}_{mt}^*$.
- (b) Calculate

$$E[\Delta_{0,t+\tau_0}(\bar{x}_{0t} - D_{0,t,t+\tau_0-1})] = \sum_{d \in \mathcal{D}_{0,t,t-\tau_0-1}} Pr(D_{0,t,t-\tau_0-1} = d) \Delta_{0,t+\tau_0}(\bar{x}_{0t} - d)$$

$$\text{for } t = 1, \dots, T - \tau_0, \text{ and } \bar{x}'_{0t} = \underline{M}_{0t}, \dots, \sum_{m \in \mathcal{R}} \bar{y}_{mt}^*.$$

5. Calculate $\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma)$ for the SNS:

- (a) Use equation (2.59) to calculate $\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma)$ for $t = T - \tau_0 - \tau, \dots, T$, and $\bar{x}_{0t} = \underline{M}_{0t}, \dots, \overline{M}_{0t}$.
- (b) Use equation (2.60) to calculate $\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma)$ $t = T - \tau_0 - \tau - 1$ and $\bar{x}_{0t} = \underline{M}_{0t}, \dots, \overline{M}_{0t}$.

(c) Solve the MIP (2.58) for $\psi_{0t}(\bar{x}_{0t}|\lambda, \mu, \gamma)$ for $t = T - \tau_0 - \tau - 2, \dots, 1$ and

$$\bar{x}_{0t} = \underline{M}_{0t}, \dots, \overline{M}_{0t}.$$

6. Calculate $\psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma) = \sum_{m \in \mathcal{L}} \psi_{mt}(\bar{x}_{mt}|\lambda, \mu, \gamma)$.

To estimate $\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma)$ in time period t , we also need to add in the additional penalty terms to account for the gap between the costs counted by the decomposition and the costs accounted for by the original problem. Given the historical inventory positions $\bar{\mathbf{x}}_t^{\text{past}}$ and demands $\mathbf{d}_t^{\text{past}}$, we can do the following in time period t :

1. For $m \in \mathcal{S}$, $t' = t - \tau_m, \dots, t - 1$, and $d \in \mathcal{D}_{m,t',t'+\tau_m-1}$, use marginal analysis to calculate $\Delta_{m,t'+\tau_m}(\bar{x}_{mt'} - d|\lambda, \mu, \gamma)$.
2. For $m \in \mathcal{S}$ and $t' = t - \tau_m, \dots, t - 1$, calculate $E[\Delta_{m,t'+\tau_m}(\bar{x}_{mt'} - d_{m,t',t'+\tau_m-1}|\lambda, \mu, \gamma)] = \sum_{d \in \mathcal{D}_{m,t',t'+\tau_m-1}} \text{Pr}(D_{m,t',t'+\tau_m-1} = d) \Delta_{m,t'+\tau_m}(\bar{x}_{mt'} - d|\lambda, \mu, \gamma)$.
3. Calculate

$$\begin{aligned} \tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma) &= \psi_t(\bar{\mathbf{x}}_t|\lambda, \mu, \gamma) \\ &+ \sum_{t'=t}^{t+\tau_m-1} \sum_{m \in \mathcal{S}} E[\Delta_{mt'}(\bar{x}_{mt'} - d_{m,t'-\tau_m,t'-1}|\lambda, \mu, \gamma)]. \end{aligned}$$

We showed in Lemma 2.2.7 that $\tilde{V}_1(\bar{\mathbf{x}}_1, \mathbf{d}_1|\lambda, \mu, \gamma)$ is a lower bound on the total cost of the dynamic program. We can also use the relaxed value functions $\tilde{V}_t(\bar{\mathbf{x}}_t^{\text{past}}, \mathbf{d}_t^{\text{past}}|\lambda, \mu, \gamma)$ to make inventory decisions in a rolling horizon manner, as in the TCA approach. In time period t , we can make inventory decisions by solving the original, unrelaxed problem (2.25), replacing $V_{t+1}(\mathbf{x}_{t+1}^{\text{past}}, \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}})$ with $\tilde{V}_{t+1}(\mathbf{x}_{t+1}^{\text{past}} - \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}}|\lambda, \mu, \gamma)$. We will refer to this as the Lagrangian Relaxation (LR) inventory policy. Further, we can use a simple Monte Carlo simulation,

like the one described in Algorithm 1, to estimate the cost of implementing the LR policy. Since the relaxed future cost functions, $\tilde{V}_{t+1}(\mathbf{x}_{t+1}^{\text{past}} - \mathbf{q}_{t+1}^{\text{past}}, \mathbf{d}_{t+1}^{\text{past}} | \lambda, \mu, \gamma)$, are approximations of the original ones, the expected cost given by the Monte Carlo simulation is an upper bound on the true cost that could be incurred. In Section 2.4 we will explore the gap between these bounds. First, however, we must show how the Lagrange multipliers are calculated.

Calculating Lagrange Multiplier Values

We showed in Proposition 2.2.7 that the Lagrangian relaxation (2.47) is a lower bound on the original problem (2.25). Hence, we would like to find sets of Lagrange multipliers λ, μ , and γ that maximize this lower bound, given that the multipliers must be nonnegative and satisfy inequality (2.41). That is, we wish to solve

$$\max_{\lambda, \mu, \gamma} \tilde{V}_1(\tilde{\mathbf{x}}_1 | \lambda, \mu, \gamma) \quad (2.67)$$

such that $\lambda_{nt}, \mu_{nt}, \gamma_{nt} \geq 0$ for all $n \in \mathcal{L}; t \in \mathcal{T}$

$$h_{nt} - \lambda_{nt} + \lambda_{n,t+1} + \sum_{m \in \mathcal{R}} u_{mn}(\mu_{mt} - h_{mt} - \mu_{m,t+1}) \geq 0 \text{ for } n \in \mathcal{P}; t \in \mathcal{T}.$$

However, solving this problem is quite difficult, so instead we will use a method for estimating the multiplier values using dual variables presented by Kunnumkal and Topaloglu [Kunnumkal & Topaloglu, 2010]. First, let us make several observations about the multipliers. Each Lagrange multiplier is associated with a constraint. The multiplier's value is approximately equal to the cost that could be saved by relaxing the constraint by one unit in the original problem (2.25).

To estimate the values of the Lagrange multipliers, we average dual variables associated with deterministic versions of the original problems, in which we minimize the expected cost in each time period, but step forward in time using known demand values. We use a two stage method. First, we will construct the λ and μ multipliers from the unrelaxed deterministic problem. Then we will obtain the γ multipliers from the relaxed problem, in which the λ and μ multipliers are used in the cost function. The first problem is given by

$$G_1(\mathbf{d}) = \min \sum_{t \in \mathcal{T}} C_t(\mathbf{y}_t, \mathbf{q}_t) \quad (2.68)$$

$$\text{such that } y_{nt} \geq x_{nt} \quad \text{for all } n \in \mathcal{L} \quad (2.69)$$

$$\begin{aligned} \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - q_{nt}) &\leq x_{m,t-\tau_m} - q_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \quad \text{for all } m \in \mathcal{S}; t \in \mathcal{T} \\ \sum_{n \in \mathcal{L}} u_{mn}(y_{nt} - x_{nt}) &\leq p_{mt} \quad \text{for all } m \in \mathcal{S}; t \in \mathcal{T} \\ y_{0t} - x_{0t} &\leq p_t^0 \quad \text{for all } t \in \mathcal{T} \end{aligned} \quad (2.70)$$

$$\begin{aligned} x_{n,t+1} &= y_{nt} - S_{nt}(y_{nt}, a_{nt}, q_{nt}, d_{nt}) \\ &\quad \text{for all } n \in \mathcal{P}; t = 1, \dots, T-1. \end{aligned}$$

$$\begin{aligned} q_{n,t+1} &= q_{nt} + d_{nt} - S_{nt}(y_{nt}, a_{nt}, q_{nt}, d_{nt}) \\ &\quad \text{for all } n \in \mathcal{P}; t = 1, \dots, T-1. \end{aligned}$$

We λ_{nt} for $n = 1, \dots, M+N$ and $t \in \mathcal{T}$ from the dual variables associated with constraint (2.69). We find μ_{nt} for $m \in \mathcal{S}$ and $t \in \mathcal{T}$ from the dual variables associated with constraint (2.70). We use Algorithm 3, given below, to estimate values for λ and μ . We choose the number of iterations, I , to be a large number.

Algorithm 3:

1. For $i = 1, \dots, I$:

- a. Randomly draw a set of demands $\mathbf{d} = \{d_{nt} : n \in \mathcal{P}, t \in \mathcal{T}\}$.
- b. Solve the linear program (2.68).
- c. For $n \in \mathcal{R} \cup \mathcal{P}$ and $t \in \mathcal{T}$, set $\lambda_{nt}(i)$ to be the dual variables associated with constraint (2.69).
For $m \in \mathcal{S}$ and $t \in \mathcal{T}$, set $\mu_{mt}(i)$ to be the dual variables associated with constraint (2.70).

2. For the appropriate values of n and $t \in \mathcal{T}$, return

$$\lambda_{nt} = \frac{1}{I} \sum_{i=1}^I \lambda_{nt}(i) \quad \text{and} \quad \mu_{mt} = \frac{1}{I} \sum_{i=1}^I \mu_{mt}(i).$$

Using these λ and μ multipliers, we can calculate the values of γ in a similar manner. The next deterministic problem is given by

$$G_2(\mathbf{d}) = \min \sum_{t \in \mathcal{T}} \left\{ C'_t(\bar{\mathbf{y}}_t) - \sum_{n \in \mathcal{P} \cup \mathcal{R}} \lambda_{nt}(\bar{y}_{nt} - \bar{x}_{nt}) - \sum_{m \in \mathcal{S}} \mu_{mt} \left(p_{mt} - \sum_{n \in \mathcal{L}} u_{mn}(\bar{y}_{nt} - \bar{x}_{nt}) \right) \right\} \quad (2.71)$$

$$\text{such that} \quad \sum_{n \in \mathcal{L}} u_{mn} \bar{y}_{nt} \leq \bar{x}_{m,t-\tau_m} - d_{m,t-\tau_m,t-1} \quad \text{for all } m \in \mathcal{S}; t \in \mathcal{T}$$

$$\bar{y}_{0t} - \bar{x}_{0t} \leq p_t^0 \quad t \in \mathcal{T}$$

$$\bar{y}_{nt} \leq a_{nt} + \bar{x}_{nt} \quad \text{for all } n = M+1, \dots, M+N; t \in \mathcal{T} \quad (2.72)$$

$$\bar{y}_{nt} \leq a_{nt} \quad \text{for all } n \in \mathcal{P}; t \in \mathcal{T}$$

$$\bar{x}_{n,t+1} = \bar{y}_{nt} - d_{nt} \quad \text{for all } n \in \mathcal{P}; t = 1, \dots, T-1.$$

We can find γ using a modified version of Algorithm 3, in which we solve problem (2.71) in step (1b) and we let γ_{nt} be the dual variables associated with constraint (2.72) for $n \in \mathcal{P}$ and $t \in \mathcal{T}$ in step (1c).

To satisfy all previously stated propositions and lemmas, the calculated Lagrange multipliers must satisfy several requirements. First, the multipliers must satisfy the constraints $\lambda_{nt}, \mu_{nt}, \gamma_{nt} \geq 0$ for all $n \in \mathcal{L}$ and $t \in \mathcal{T}$ and

$$h_{nt} - \lambda_{nt} + \lambda_{n,t+1} + \sum_{m \in \mathcal{R}} u_{mn}(\mu_{mt} - h_{mt} - \mu_{m,t+1}) \geq 0 \quad \text{for all } n \in \mathcal{P} \text{ and } t \in \mathcal{T}.$$

Also, we know that $\gamma_{nt} = 0$ if $\bar{x}_{nt} > 0$, since in this case the constraint $\bar{y}_{nt} \leq a_{nt} + \bar{x}_{nt}$ will always be slack because it is dominated by the constraint $\bar{y}_{nt} \leq a_{nt}$. So if $\gamma_{n1} > 0$ we must have $\bar{x}_{n1} \leq 0$, and we should have

$$\gamma_{n1} \bar{x}_{n1} \leq 0 \quad \text{for all } n \in \mathcal{P}.$$

Finally, if $\lambda_{nt} > 0$, then $\bar{y}_{nt} = \bar{x}_{nt}$, so $\bar{y}_{nt} - \bar{x}_{nt} \leq 0$ and hence the constraint $\bar{y}_{nt} \leq \bar{x}_{nt} + a_{nt}$ will be satisfied and its associated multiplier, γ_{nt} , will be 0. The contrapositive is also true ($\gamma_{nt} > 0$ implies $\lambda_{nt} = 0$), so we have

$$\lambda_{nt} \gamma_{nt} = 0 \quad \text{for all } n \in \mathcal{P} \text{ and } t \in \mathcal{T}.$$

2.3 Decentralized Allocation Methods

Most existing state and local emergency response plans call for inventory decisions to be made in one of two ways. The most common allocation strategy calls for all locations in the network to receive a “fair share” of the available inventory. Upper echelon locations “push” inventory downstream so that each location receives inventory proportional to the total population that it is expected to serve. Inventory is shipped without regard to the state of the receiving locations, so no information needs to be passed upstream for allocation decisions to be made. The second allocation strategy in use allows each location in the distribution network to place inventory orders, which are filled as long as sufficient inventory is available at the supplying locations. This “independent ordering” policy is slightly more sophisticated than the Fair Share policy, since some information is passed upstream, in the form of inventory orders.

We will implement both the Fair Share and Independent Ordering policies to establish a baseline performance of the distribution network to allow us to assess the value of the TCA and Lagrangian Relaxation policies described in the previous sections. Since neither of the two decentralized policies is well-defined mathematically, we must first establish detailed decision-making rules for each allocation method.

2.3.1 Fair Share Method

For the planning model, we have consistently assumed that the expected demands are known in advance for all time periods, and we will make the same

assumption here. We will assume that for POD n , the proportion of the total inventory that will be shipped to POD n is

$$\text{Fair Share}_n = \frac{\sum_{t \in \mathcal{T}} E[D_{nt}]}{\sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{P}} E[D_{nt}]}.$$

The question remains of how much inventory would be shipped in each period. The likely answer would involve an initial large push followed by smaller subsequent pushes. However, to make the comparison between the Fair Share method and the other policies more reasonable, we will assume shipments are sent out more frequently, taking advantage of information regarding the expected demands. The total expected demand in period t is $\sum_{n \in \mathcal{P}} E[D_{nt}]$. We will suppose that the policy accounts for some safety stock, and the total inventory shipped out to PODs in time t will be the expected total demand plus k standard deviations of the total demand. If the POD demands are Poisson distributed, then this is

$$\sum_{n \in \mathcal{P}} E[D_{nt}] + k \sqrt{\sum_{n \in \mathcal{P}} E[D_{nt}]},$$

and in period t POD n receives a shipment of size

$$\left\lceil (\text{Fair Share}_n) \left(\sum_{n \in \mathcal{P}} E[D_{nt}] + k \sqrt{\sum_{n \in \mathcal{P}} E[D_{nt}]} \right) \right\rceil.$$

Given these shipment values, we can calculate the desired size of the shipments that should to be sent to the RSSs and the SNS to make this possible. However, shipment constraints may prevent material from being sent in the period during which it is required. We will resolve this by shipping any un-sent inventory

as soon as excess capacity becomes available. And, of course, each POD and RSS receives a “fair share” of the available shipping capacity, as well. We will use a Monte Carlo simulation like Algorithm 1 to estimate the expected cost of implementing the Fair Share method.

2.3.2 Independent Ordering Method

To implement the Independent Ordering allocation method, we need to define the policy that each location in the network will use to determine its order quantities. Most regions that use an Independent Ordering system provide PODs with little or no guidance in determining order quantities, but we will optimistically assume that each location in the network will order its myopic optimal quantity.

Each POD determines its inventory without regard for system capacities, so we define the problems to be solved at the PODs without any linking constraints related to inventory or transportation. With these constraints removed, we know that we can write the problem in terms of the inventory positions \bar{y}_{nt} and \bar{x}_{nt} , so in period t , the optimal inventory level \tilde{y}_{nt} for POD n is found by solving

$$\tilde{y}_{nt} = \operatorname{argmin}_{\bar{y}_{nt} \geq \bar{x}_{nt}} C'_{nt}(\bar{y}_{nt}).$$

The optimal solution is $\tilde{y}_{nt} = \max\{\bar{x}_{nt}, \operatorname{argmin} C'_{nt}(\bar{y}_{nt})\}$, so the PODs use an order-up-to policy. Let I_{nt} be smallest value of \bar{y}_{nt} that minimizes $C'_{nt}(\bar{y}_{nt})$, so $\tilde{y}_{nt} = \max\{\bar{x}_{nt}, I_{nt}\}$. Since $C'_{nt}(\bar{y}_{nt})$ is convex in \bar{y}_{nt} , I_{nt} is the smallest value for which

$C'_{nt}(\bar{y}_{nt} + 1) - C'_{nt}(\bar{y}_{nt}) \geq 0$. Notice that if $\bar{y}_{nt} \geq a_{nt}$, then $S_{nt}(y_{nt}) = S_{nt}(y_{nt} + 1) = \min(a_{nt}, D_{nt})$ and

$$\begin{aligned} C'_{nt}(\bar{y}_{nt} + 1) - C'_{nt}(\bar{y}_{nt}) &= \sum_{d \in \mathcal{D}_{nt}} f_{nt}(d) \left[h_{nt}(\bar{y}_{nt} + 1 - \min(a_{nt}, d)) + f_{nt}^B(d - \min(a_{nt}, d)) \right. \\ &\quad \left. - h_{nt}(\bar{y}_{nt} - \min(a_{nt}, d)) - f_{nt}^B(d - \min(a_{nt}, d)) \right] \\ &= \sum_{d \in \mathcal{D}_{nt}} f_{nt}(d) [h_{nt}(1)] \\ &= h_{nt}. \end{aligned}$$

So the cost function is increasing for all $\bar{y}_{nt} \geq a_{nt}$. Thus, the optimal myopic solution will never exceed a_{nt} . This is reasonable, since there is no point in accumulating inventory if there is not enough service capacity to use it. We will only have $\tilde{y}_{nt} > a_{nt}$ if $\bar{x}_{nt} > a_{nt}$.

Thus, we wish to find

$$I_{nt} = \operatorname{argmin}_{\bar{y}_{nt} \leq a_{nt}} \sum_{d \in \mathcal{D}_{nt}} f_{nt}(d) [h_{nt}(\bar{y}_{nt} - \min(\bar{y}_{nt}, d)) + f_{nt}^B(d - \min(\bar{y}_{nt}, d))].$$

If f_{nt}^B is a simple linear function of the form $f_{nt}^B(u) = b_{nt}u$, then the optimal solution is

$$I_{nt} = \min \left\{ a_{nt}, \left\lceil F_{nt}^{-1} \left(\frac{b_{nt}}{b_{nt} + h_{nt}} \right) \right\rceil \right\},$$

which is a classic inventory management result [Muckstadt & A., 2010]. There is not a simple analytic solution when $f_{nt}^B(\cdot)$ is nonlinear, but since the function is piecewise linear, the problem can be solved by a linear program.

A more sophisticated mechanism for calculating order quantities for the PODs would be to solve a dynamic program for each POD, minimizing costs over the full time horizon rather than just a single time period. The problem that we would wish to solve could be written as the following value function

$$v_{nt}(\bar{x}_{nt}) = \min_{\bar{x}_{nt} \leq \bar{y}_{nt}} C'_{nt}(\bar{y}_{nt}) + E[v_{n,t+1}(\bar{y}_{nt} - D_{nt})].$$

As in the case of the myopic problem, the objective function is convex, so the optimal ordering policy is an order-up-to policy, I'_{nt} . If the random demands are stochastically nondecreasing, then the myopic optimal order-up-to levels, I_{nt} , are optimal for the dynamic program, as well [Muckstadt & A., 2010].

We must also define ordering and allocation policies for the RSSs and the SNS. We optimistically assume that all locations independently calculate the optimal inventory levels for the PODs, according to the myopic or dynamic programming minimization problems. When shortages arise, we assume that the SNS and RSSs will send equal proportions of the amount ordered to each of the lower echelon locations served, so that all of the on-hand inventory is consumed. Then RSS m places an order in period $t - \tau_m - 1$ for $\sum_{n \in \mathcal{P}} u_{mn} I_{nt}$ units of inventory (or I'_{nt} , if we are using the dynamic programming order-up-to values). If there is not sufficient shipping capacity in period $t - \tau_m$, then the inventory is shipped when it becomes available. Similarly, the SNS orders $\sum_{n \in \mathcal{P}} I_{nt}$ (or I'_{nt}) units of inventory in period $t - \tau - \tau_0 - 2$. This policy provides reasonable and clear rules for placing orders at all locations, and we will use it in our simulations. Note that we are effectively assuming that all of the locations in the network share information about order-up-to levels at the beginning of the time horizon, so that the SNSs and RSSs may place orders for the quantities described here.

2.4 Computational Results and Discussion

We now compare the performance of the Truncated Cumulative Approximation (TCA), Lagrangian Relaxation (LR), Individual Ordering (Order), and Fair Share inventory allocation policies discussed in the previous section. Our goals are to evaluate how well these allocation methods perform under a variety of conditions and to identify some insights that can help public health planners improve emergency response plans.

We tested the four allocation methods, using them to make decisions and calculating costs incurred, patient waiting times, and inventory required for simulated patient demands in each period of a time horizon. We repeated this process for a number of iterations to estimate the expected cost of applying each method for each set of simulation parameters. We ran 23 simulation experiments with varying parameter values, which are described in Table 2.2. In all of the simulations we modeled a distribution network with two RSSs, each of which served an equal number of the PODs. Time periods were assumed to be four hours long. The lead time from the manufacturer to the SNS was set to three periods, and the lead times from the SNS to the RSSs were each one period. Patient demands were assumed to be independent and Poisson distributed, but the means varied by time and location. The numbers of PODs and the service capacity at each POD in each period also varied. The per unit holding costs were 0.01 at the SNS, 0.1 at the RSSs, and 1 at the PODs, and the backorder costs at each POD were given by a linearized quadratic function.

Data Set	Number of PODs	Time Periods	Cumulative Service Capacity (per POD)	Cumulative Expected Demand (per POD)	Demand Rate Description	Iters.
1.11	10	12	21,600	14,400	Constant	50
1.12	10	12	21,600	14,400	Increasing	20
1.13	10	12	21,600	14,400	Decreasing	20
1.14	10	12	21,600	14,400	Alternating high and low	20
1.15	10	12	21,600	14,400	High then low	20
1.21	10	12	43,200	14,400	Constant	20
1.22	10	12	43,200	14,400	Increasing	20
1.24	10	12	43,200	14,400	Alternating high and low	20
2.11	10	10	10,000	10,000	Constant	30
2.12	10	10	13,000	10,000	Constant	30
2.13	10	10	1600	10,000	Constant	30
2.21	10	10	13,000	10,000	Increasing	30
2.22	10	10	13,100 (varying)	10,000	Increasing	30
2.23	10	10	16,200 (varying)	10,000	Increasing	30

2.31	10	10	13,000	10,000	Decreasing	30
2.32	10	10	13,100 (varying)	10,000	Decreasing	30
2.33	10	10	16,200 (varying)	10,000	Decreasing	30
3.11	2	10	9,600	6,400	Constant	50
3.12	4	10	9,600	6,400	Constant	50
3.13	8	10	9,600	6,400	Constant	50
3.14	16	10	9,600	6,400	Constant	50
3.23*	8	10	19,200	6,400	Constant	50
3.24*	16	10	19,200	6,400	Constant	50

Table 2.2: Simulation Parameters. *All simulations began with initial RSS inventories set at two periods of expected demand and initial SNS inventories set at five periods of expected demand, except for simulations 3.23 and 3.24, which were initialized with five periods of expected demand at the RSSs and nine periods of expected system demand at the SNS.

We ran three main sets of simulations, which we will refer to as experiments. The first experiment (simulations 1.11-1.24) explored the consequences of varying expected patient demand rates and service capacities in a ten POD network over twelve time periods. In simulations 1.11-1.15, all of the simulation param-

eters, except for the demand patterns, were identical. Simulations 1.21, 1.22, and 1.24 were identical to simulations 1.11, 1.12, and 1.14, respectively, except the service capacity in the former simulations is double that of the latter. The goal of these simulations was to measure the effect of demand on system performance, when other capacities were ample. Note that the first digit of each simulation number indicates the experiment of which it was a part.

The second experiment (simulations 2.11-2.33) was a more thorough investigation of the importance of varying service capacity in a ten POD network over ten time periods. Simulations 2.11-2.13 used a constant expected patient arrival rate of 1,000 people per POD per time period, but the service capacity for each POD increased from 1,000 to 1,300 to 1,600 for the three simulations. Simulations 2.21-2.23 assumed an expected patient demand rate that increased from 600 to 1,400 people per hour over the course of the day. The service capacities for simulation 2.21 remained a constant rate of 1,300 people served per period. For simulation 2.22, the service capacity was dynamic, changing over time with the demand rate. For each period, service capacity was set to the expected patient demand plus one standard deviation of the demand. For simulation 2.23, we also set service capacity in a dynamic manner, increasing the per period capacity to the expected patient demand plus two standard deviations. Simulations 2.31-2.33 used the same rules for setting service capacity as were used in simulations 2.21-2.23, but the expected demand pattern for these simulations decreased steadily over time from 1,400 to 600 patient arrivals per period.

The third experiment (simulations 3.11-3.24) explored the impact of modifying the network structure given a constant set of resources. In simulations 3.11-3.14 the total expected patient demand and total service capacity remained

constant across the network in each time period. Both the expected demand and the service capacity were divided evenly among the PODs. There were 2, 4, 8, and 16 PODs in simulations 3.11, 3.12, 3.13, and 3.14, respectively, so the networks became more dispersed as the simulation number increased. Simulations 3.23 and 3.24 were identical to simulations 3.13 and 3.14, but increased both the initial inventory in the system and the available service capacity in each period to measure the consequences of more dispersed POD networks under generous constraints.

The next section presents a comparison of the costs incurred in each simulation under each allocation method. We also discuss the quality of the Wait-and-See (WS) and Lagrangian Relaxation (LR) lower bounds. The following section presents further results from these simulations in terms of patient delay and inventory use, and we show how these may help public health officials better improve their emergency preparedness plans.

2.4.1 Cost Comparison of Allocation Methods

Figure 2.3 shows the average cost incurred by each allocation method in all of the simulations done for each of the three experiments. We observe that there is very significant variance between the costs incurred under different simulations. In particular, we can see that in the top graph of Figure 2.3, all of the allocation methods accumulated costs of almost 200,000 in simulations 1.13 and 450,000 in simulation 1.15, while the costs for each allocation method in simulations 1.11, 1.12, and 1.14 generally remained well below 25,000. The reason for this significant disparity is the difference in demand patterns; the distribution

network is overwhelmed in simulations 1.13 and 1.15; huge backorder costs are incurred no matter which allocation method is used because the service and transportation capacities are insufficient for the patient demand.

For most simulations, however, we see that TCA and LR methods both perform very well. The Order method generally performs almost as well, while the Fair Share method incurs significantly larger costs. There are, of course, some exceptions to this pattern, which we will discuss below. However, it is difficult to compare the different simulation runs simply from examining the costs; to understand the actual performance of each method, we scale these average costs by the two lower bounds that we have defined. The scaled costs are presented in Table 2.3.

Table 2.3 displays the average cost from each allocation method in each simulation divided by the Wait-and-See (WS) and Lagrangian Relaxation (LR) lower bounds. Recall that the Wait-and-See lower bound is the expected cost that would be incurred if the patient demands were known perfectly in advance. The LR lower bound is found by calculating the decomposed, relaxed dynamic program, as in Algorithm 2. We observe that the Wait-and-See lower bound is tighter than the LR lower bound in about 70% of the simulations that we ran. The WS bound is large (and therefore tight) when there are significant “unavoidable” costs; that is, costs that result from limitations of the system parameters rather than poor allocation decisions. For example, in simulations 1.13 and 1.15 the costs are dominated by unavoidable costs, so the WS bound is very good. However, when there is ample capacity in a simulated distribution network, the WS bound becomes quite small, and the LR lower bound becomes more useful, as in simulations 3.23 and 3.24. In networks with ample capacity,

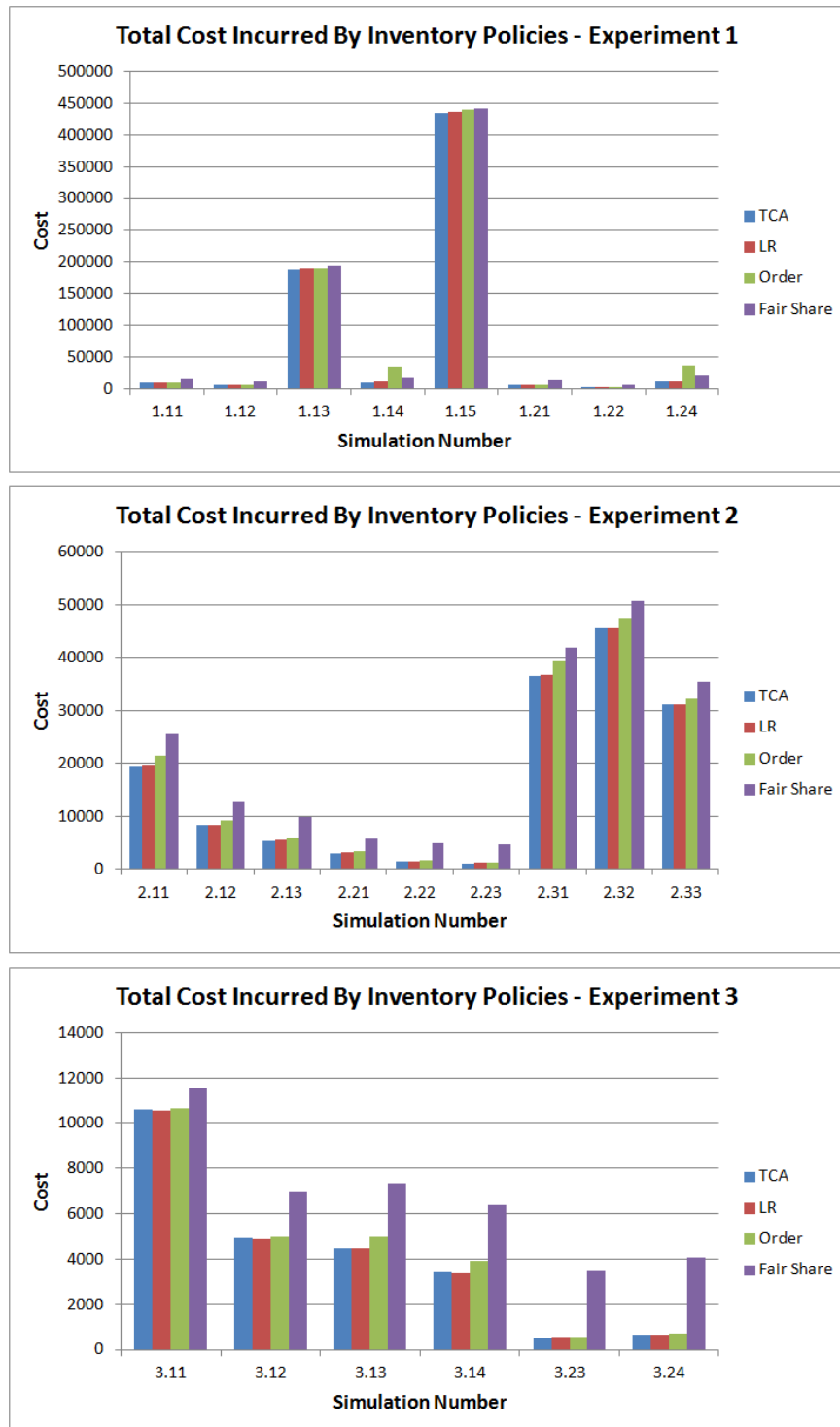


Figure 2.3: Total cost incurred for all simulations.

the constraints that were relaxed to obtain the LR bound are not very important to the problem, so the cost estimated by the relaxed problem will be close to the true cost. Thus, the WS and LR lower bounds complement one another, allowing us to guarantee an optimality gap of at most 20% in more than half of the simulations that we ran. Unfortunately, since neither the WS nor the LR captures the cost of imbalance very well, neither bound is very good when significant imbalance arises, as in simulations 3.13 and 3.14. In the following section, we discuss some of the challenges of managing more dispersed POD networks in which imbalance is more likely.

We also see from Table 2.3 that the TCA solution is the smallest in just over half of the simulations, while the LR solution is smallest in just under half. The Order solution gives the minimum in one simulation, and the Fair Share solution is never minimal. However, in addition to considering these averages, we can examine the performance of each method during each iteration. Table 2.4 below shows the number of iterations that each allocation method provided the minimum cost solution and the second smallest solution. We see that the TCA method provided the smallest or second-smallest solution in 87% of the iterations, while the LR method was smallest or second smallest in 75% of the iterations. Somewhat surprisingly, the Order method performs best or second best in over 30% of cases.

The performance of the Order method seems too good to be true for an allocation method that shares so little information between the different parts of the supply chain. However, recall that the method that we used for calculating order-up-to levels at the RSSs and SNS allows these locations to predict the PODs' orders with great accuracy, effectively implementing collaboration be-

tween the different members of the supply chain. We also assumed that the PODs would order exactly their optimal value in every period, which is overly optimistic, since most POD managers are given little or no training in determining their order quantities. To explore the robustness of the Order policy, we explored what would happen if the PODs sometimes ordered too much or too little inventory. We re-ran the simulations, allowing order-up-to levels at the PODs to range uniformly from 0.7 to 1.3 times their optimal quantities. We refer to this new method as the “Imperfect Order” policy. Table 2.5 shows the number of iterations for which each allocation method provided the minimum cost solution and the second smallest solution, when the Order policy has been replaced by the Imperfect Order policy. When compared to the Imperfect Order policy, the TCA method provides the best or second-best solution in about 95% of simulation iterations, with the LR method a close second at about 93%.

This comparison may seem rather unfair, since we have only modified the calculations of the Order policy, while continuing to allow the TCA and LR methods to operate correctly. In fact, we argue that the Imperfect Order policy is a more accurate representation of how a dispensing campaign would perform when each location in the network makes independent decisions. To guarantee that PODs would order accurately and ensure that the RSSs and SNS are also aware of these future orders would require an information infrastructure system that could calculate near-optimal order quantities for the PODs and share these values throughout the network. However, if such a system were in place, then there would be no reason to make decisions for individual locations without accounting for the state of the complete network. A centralized policy like the TCA or LR methods could be implemented to provide even better performance.

In the next section, we continue discussing the implications that these simulations have for the design and operation of an emergency response network. We specifically address several questions that have arisen during conversations with public health authorities in New York City [Starr, 2012].

Sim.	TCA		LR		Order		Fair Share	
	WS LB	LR LB	WS LB	LR LB	WS LB	LR LB	WS LB	LR LB
1.11	1.431	1.634	1.423	1.625	1.575	1.798	2.465	2.814
1.12	1.094	8.931	1.135	9.263	1.155	9.433	2.047	16.715
1.13	1.074	2.208	1.089	2.240	1.089	2.240	1.122	2.307
1.14	1.130	2.609	1.166	2.693	3.865	8.928	1.839	4.248
1.15	1.035	1.992	1.040	2.001	1.050	2.020	1.052	2.025
1.21	1.344	1.045	1.343	1.044	1.408	1.095	3.069	2.386
1.22	3.393	1.374	3.602	1.459	3.385	1.370	19.519	7.903
1.24	1.137	1.973	1.171	2.033	3.787	6.572	2.067	3.587
2.11	1.140	4.413	1.160	4.491	1.261	4.880	1.499	5.802
2.12	1.557	2.194	1.553	2.188	1.699	2.395	2.386	3.362
2.13	1.550	1.406	1.558	1.413	1.680	1.524	2.784	2.526
2.21	1.536	3.379	1.578	3.472	1.723	3.791	2.936	6.459
2.22	2.072	1.527	2.136	1.575	2.291	1.689	7.032	5.185
2.23	3.793	1.202	4.051	1.284	4.033	1.278	16.598	5.259
2.31	1.176	2.263	1.181	2.273	1.265	2.436	1.353	2.604
2.32	1.195	2.902	1.195	2.900	1.247	3.028	1.332	3.232
2.33	1.243	2.003	1.240	1.998	1.280	2.061	1.410	2.271
3.11	1.145	1.619	1.138	1.609	1.149	1.625	1.249	1.766
3.12	1.324	1.361	1.312	1.348	1.336	1.372	1.876	1.928
3.13	1.590	2.021	1.581	2.010	1.770	2.250	2.599	3.304
3.14	1.978	2.303	1.935	2.253	2.248	2.617	3.673	4.276
3.23	3.964	1.135	4.059	1.162	4.227	1.210	26.509	7.588
3.24	5.151	1.374	5.039	1.344	5.493	1.465	31.156	8.311

Table 2.3: Average total costs scaled by the Wait-and-See (WS) and Lagrangian Relaxation (LR) lower bounds. The minimum value in each row is indicated by bold font.

Simulation	TCA	LR	Order	Fair Share
1.11	16 / 25	25 / 18	9 / 7	0 / 0
1.12	13 / 7	3 / 10	4 / 3	0 / 0
1.13	12 / 6	0 / 8	7 / 5	1 / 1
1.14	18 / 2	2 / 18	0 / 0	0 / 0
1.15	15 / 4	3 / 12	1 / 3	1 / 1
1.21	7 / 7	9 / 8	4 / 5	0 / 0
1.22	11 / 9	0 / 7	9 / 4	0 / 0
1.24	33 / 17	17 / 32	0 / 0	0 / 1
2.11	19 / 11	9 / 17	2 / 1	0 / 1
2.12	11 / 15	13 / 13	5 / 1	1 / 1
2.13	13 / 14	11 / 12	6 / 4	0 / 0
2.21	11 / 17	7 / 8	11 / 5	1 / 0
2.22	13 / 16	3 / 11	14 / 3	0 / 0
2.23	11 / 19	4 / 7	15 / 4	0 / 0
2.31	18 / 10	9 / 19	2 / 1	1 / 0
2.32	14 / 13	10 / 15	5 / 2	1 / 0
2.33	8 / 15	10 / 10	8 / 4	4 / 1
3.11	16 / 17	14 / 21	15 / 6	5 / 6
3.12	10 / 19	17 / 19	20 / 12	3 / 0
3.13	22 / 22	23 / 16	5 / 12	0 / 0
3.14	15 / 29	33 / 15	2 / 6	0 / 0
3.23	31 / 18	4 / 28	15 / 4	0 / 0
3.24	7 / 25	28 / 22	15 / 3	0 / 0
Totals:	344 / 337	254 / 346	174 / 95	18 / 12
Fraction in 1 st :	0.435	0.322	0.220	0.023
Fraction in 2 nd :	0.427	0.438	0.120	0.015
Fraction in 1st or 2nd:	0.862	0.759	0.341	0.038

Table 2.4: Number of iterations for which each solution method gives the smallest / second smallest cost.

Simulation	TCA	LR	Imperfect Order	Fair Share
1.11	20 / 28	28 / 21	0 / 0	2 / 1
1.12	17 / 3	3 / 17	0 / 0	0 / 0
1.13	14 / 5	1 / 13	2 / 0	3 / 2
1.14	18 / 2	2 / 18	0 / 0	0 / 0
1.15	16 / 3	3 / 14	1 / 3	0 / 0
1.21	9 / 11	11 / 9	0 / 0	0 / 0
1.22	20 / 0	0 / 20	0 / 0	0 / 0
1.24	21 / 19	7 / 30	0 / 0	22 / 1
2.11	21 / 9	9 / 20	0 / 1	0 / 0
2.12	13 / 14	14 / 15	2 / 1	1 / 0
2.13	18 / 12	12 / 18	0 / 0	0 / 0
2.21	22 / 8	7 / 22	1 / 0	0 / 0
2.22	26 / 4	4 / 26	0 / 0	0 / 0
2.23	26 / 4	4 / 26	0 / 0	0 / 0
2.31	19 / 10	10 / 20	1 / 0	0 / 0
2.32	14 / 14	11 / 16	1 / 0	4 / 0
2.33	11 / 15	11 / 12	5 / 0	3 / 3
3.11	20 / 22	22 / 19	6 / 8	2 / 1
3.12	19 / 28	27 / 22	3 / 0	1 / 0
3.13	24 / 25	25 / 21	0 / 0	1 / 4
3.14	16 / 32	34 / 16	0 / 0	0 / 2
3.23	45 / 5	5 / 45	0 / 0	0 / 0
3.24	7 / 43	43 / 7	0 / 0	0 / 0
Totals:	436 / 316	293 / 395	447 / 13	39 / 14
Fraction in 1 st :	0.552	0.371	0.028	0.049
Fraction in 2 nd :	0.400	0.566	0.016	0.018
Fraction in 1st or 2nd:	0.952	0.937	0.044	0.067

Table 2.5: Number of Minima and Second Place (Imperfect Ordering).

2.4.2 Policy Implications

Our goal in this section is to demonstrate how simulations like these can help public health authorities prepare better for emergencies. We will address three questions that have been asked during meetings with New York City public health officials. First, why is a command and control system necessary? How could it help reduce patient waiting times or inventory use? Second, how much staffing capacity is necessary at PODs, and how can we best use limited staff? Third, is opening additional PODs a good idea? The last question was a direct consequence of complaints from the public during the H1N1 vaccination campaign in 2009-2010. New York City opened PODs to give vaccines, and some people felt that more PODs should have been opened to reduce the travel burden for patients. Since then, the NYC Department of Health and Mental Hygiene (DOHMH) Office of Emergency Preparedness and Response has been asked to consider opening more PODs, but they are concerned that this could strain their resources. In this section, we show how these questions can be addressed quantitatively.

As discussed earlier, the costs in this model are proxies for our actual goals of reducing inventory use and patient delay. In this section, we present simulation results in terms of average patient waiting time and average inventory required per person, which is the total amount of inventory sent out from the SNS divided by the number of patients served over the time horizon. To determine the value of a command and control system, we begin by considering the patient delays and inventory requirements under our four allocation methods for a variety of simulations. Figure 2.4 shows the average waiting time per person for experiments 1 and 2. It is clear that the TCA and LR methods provide

the lowest patient delays; decisions are made centrally for both of these methods, so we could not implement them without a good command and control system that shares inventory and patient demand information in every time period. However, the Order policy performs almost as well in most simulations, and the Fair Share policy only performs significantly worse in simulations 2.13 and 2.21-2.23. But we must consider inventory use, as well.

Figure 2.5 shows the average units of inventory required per person, which allows us to quantify the degree of waste that results from each allocation policy. Notice that that in simulation 1.11, about 1.1 units of inventory are used for each patient served under the TCA policy, but almost 1.5 units are required for each patient served under the Fair Share policy. So almost 40% more inventory would be required to serve the same number of patients if the Fair Share policy were implemented instead of the TCA policy. We see that the situation is even worse in many of the simulations. This means that if we implemented the simple Fair Share policy, which does not require any information sharing or central decision-making, we may require 40% or more additional units of inventory to provide patients with longer waiting times. Any money that was saved by not creating a command and control system would quickly be consumed in inventory costs.

We discussed some problems with the Order policy in the previous section. To obtain the high performance that we observe in Figures 2.4 and 2.5 would require significant infrastructure support. If PODs truly operated individually, it is unlikely that they would order near-optimal quantities. Instead, we presented the Imperfect Order policy to be a more accurate representation of independent POD performance. Figure 2.6 shows the average patient waiting times

and inventory use per person that would result from implementing the Imperfect Order method. We see that the patient delay under the Imperfect Order policy increases significantly, in some cases more than doubling compared to the original Order policy delays, and the inventory required per person increases as well.

There are also two exceptions to the good performance displayed by the unmodified Order method: simulations 1.14 and 1.24. In the top graph in Figure 2.4, we see that the average patient delay under the Order policy in these simulations is more than double the delay under the TCA and LR policies. In the top graph in Figure 2.5, we see that the ratio of inventory distributed to patients served is only slightly larger than one; this indicates that there was almost no excess inventory remaining in the system at the end of the time horizon. Recall that simulations 1.14 and 1.24 were identical, except for the service capacity, which was doubled for 1.24. But we see from these results that this increase had virtually no impact on the performance of the network. The demand pattern for both simulations 1.14 and 1.24 was an oscillating one. The Order policy, with its myopic perspective, significantly underestimated the inventory that would be required. All other demand rates simulated changed much more gradually, and the Order policy performed significantly better. However, in a public health emergency, highly unpredictable demands may occur, and the Order policy would perform very poorly in such a scenario. Thus, we see that the policy of allowing individual PODs to place orders is not robust for unpredictable demand patterns or imperfect ordering strategies, while the centralized TCA and LR methods perform well in both cases. This provides further evidence of the value of a strong command and control system which would allow centralized allocation methods to be implemented easily.

To address questions related to staffing and service capacity, we first compare the outcomes of simulations 1.11, 1.12, and 1.14 with those of 1.21, 1.22, and 1.24. Recall that the latter group of simulations was identical to the former, except that the service capacities were doubled. Figure 2.7 displays the average patient waiting times and inventory required per person for these simulations. In simulations 1.11 and 1.21 and in simulations 1.12 and 1.22, increasing staffing capacity significantly decreases the expected per patient waiting times, although there is little effect on the expected inventory use per person. Simulations 1.14 and 1.24 do not display this trend; these simulations experienced the highly oscillating patient demand discussed earlier. We see that the difficulty in responding to the highly unpredictable patient demand pattern outweighed the limitations presented by staffing capacity in this case.

Of course, it is unsurprising that doubling service capacity yields some improvement in performance. In most public health emergency responses, staff are a limited commodity, so in experiment 2 we explored the impacts of smaller increases in service capacity and dynamic staffing decisions. Recall that the simulations in groups 2.11-2.13, 2.21-2.23, and 2.31-2.33 were each identical except for varying service capacities. For simulations 2.22-2.23 and 2.32-2.33, the service capacities varied over time with the expected demand rates. The bottom graph in Figure 2.4 shows the patient delay for these simulations. Average patient delay decreases significantly as staffing capacity increases in simulations 2.11-2.13. Simulations 2.22-2.23 and simulations 2.32-2.33 also demonstrate the value of increasing service capacity. However, we observe that patient waiting times are significantly lower in simulation 2.22 compared with simulation 2.21, even though the total service capacity remains almost constant. The main difference between the two simulations was the allocation of the service capacity

over time; in simulation 2.22 service capacity increases when patient demand is high and decreases when demand is low. We refer to this as a dynamic staffing plan.

However, the same pattern is not present in simulations 2.31 and 2.32. Instead, patient delay increases under the dynamic staffing plan implemented in simulation 2.32 due to inventory shortages. In simulation 2.32, the increased staffing capacity matches the arriving demand in the beginning of the time horizon, but there is insufficient inventory at the PODs at that point, so the extra service capacity goes unused. Later in the time horizon, when inventory arrives, the service capacity has decreased so the inventory cannot be dispensed as efficiently as in simulation 2.31. This example emphasizes the importance of modifying a staffing plan to conserve service capacity when inventory is unavailable.

Finally, we wish to address the value of increasing the number of PODs in a distribution network. In experiment 3, we explored the consequences of increasing the number of PODs in a dispensing network. The numbers of PODs operating in the dispensing networks were 2, 4, 8, and 16 in simulations 3.11, 3.12, 3.13, and 3.14, respectively. Figure 2.8 shows that both average patient delay and average inventory required per person increase significantly when the number of PODs increases. Of course, this does not account for the increased travel time that would be required of people when there are very few PODs in a distribution network, but the value of decreased waiting times is not solely one of individual convenience. Decreased waiting times also decrease the likelihood of crowd control problems and the level of security and support facilities that would be required. If capacities are not a limiting factor in a response network,

as in simulations 3.23 and 3.24, increasing the number of PODs will not present any problems. But in a public health setting, resources are always limited, and simulations like these can help public health officials justify the construction of more compact POD networks.

A good command and control system could allow public health officials to retain some benefits of a compact distribution network even when a large number of PODs are necessary to serve a population. Such a system would allow for rapid sharing of information about staff availability and service rates, patient demand, and inventory levels, so that resources could be shared between different PODs. It could also be used to inform the public about current waiting times at PODs, to encourage people to seek service at lower demand times or less-busy PODs. Dynamic staffing plans could also be implemented: if demand information were collected and shared with a central staffing office, newly arriving staff could be assigned to PODs that have both high patient demands and sufficient inventory to serve them. And, of course, centralized command and control would allow us to use an information-rich inventory allocation policy like the TCA or LR methods to serve patients with a minimum of delay and inventory, allowing us to achieve our primary goal of minimizing mortality and morbidity following.

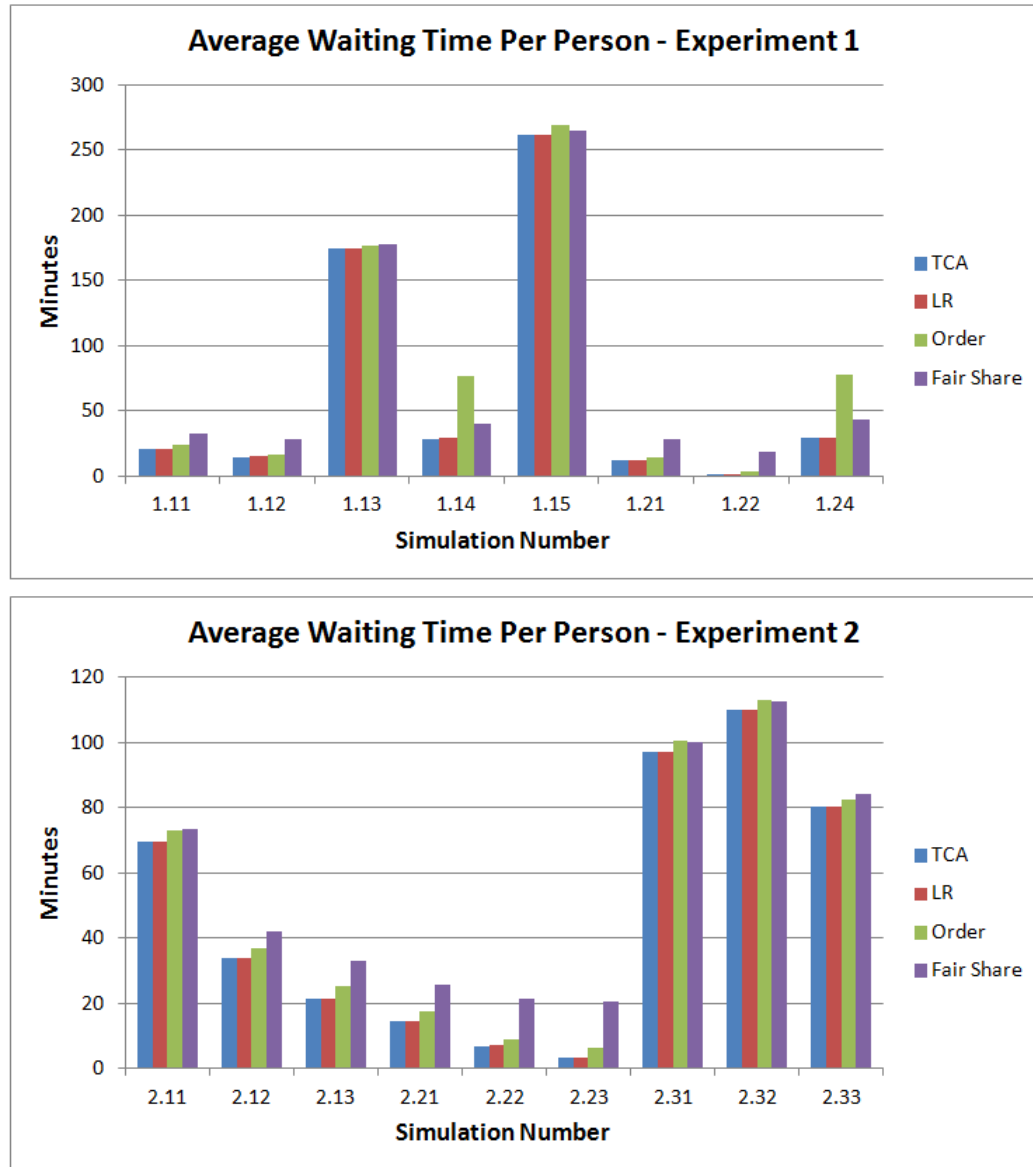


Figure 2.4: Average per patient waiting time for experiments 1 and 2.

2.5 Future Work

We have presented a model of the public health emergency response supply chain that would be used to support a mass-dispensing campaign following

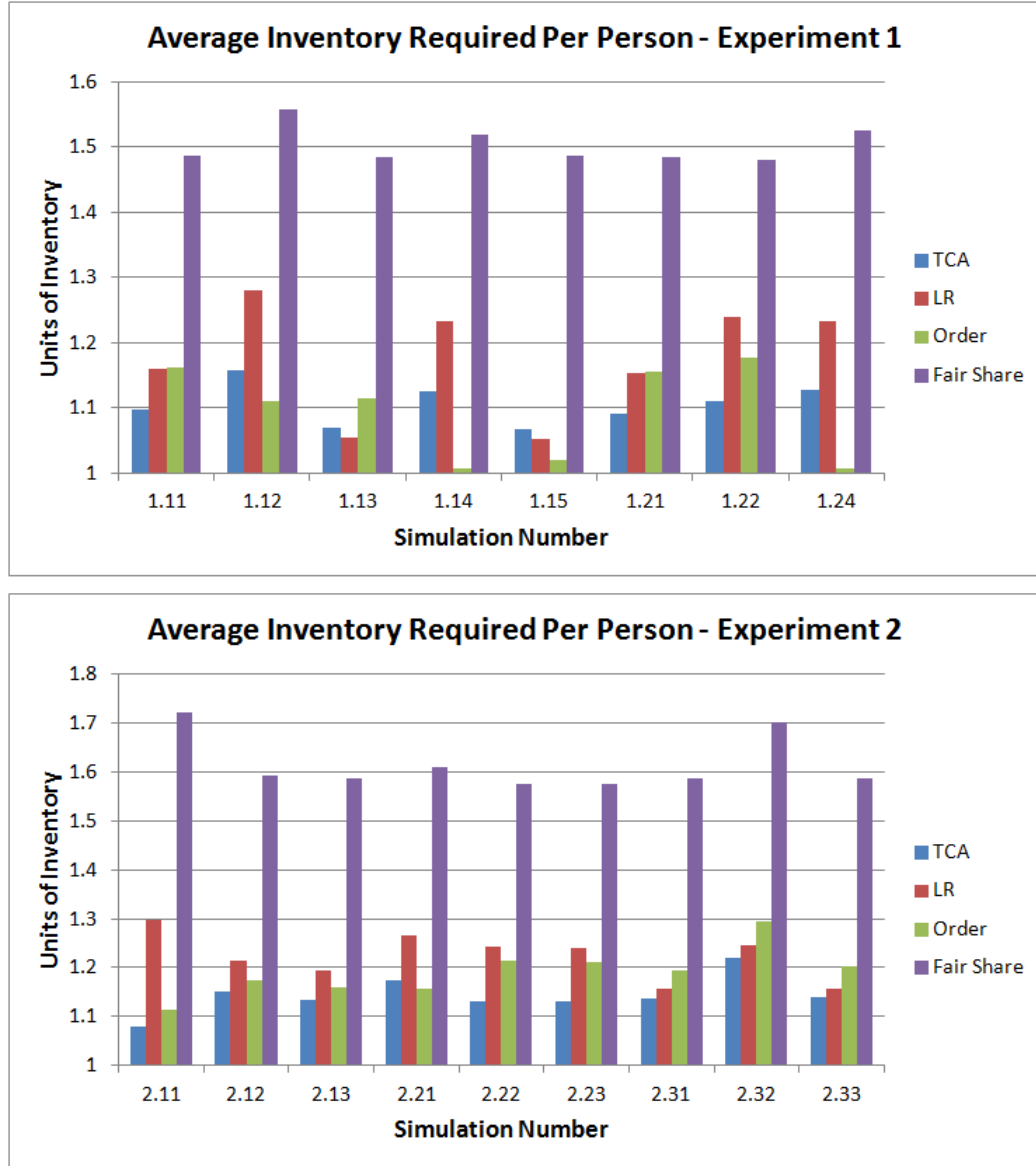


Figure 2.5: Average units of inventory required per person served for experiments 1 and 2.

an emergency such as an inhalational anthrax attack. We constructed two inventory allocation methods: the myopic Truncated Cumulative Approximation method and the Lagrangian Relaxation method. Both may be used for more general service capacity-constrained three echelon distribution networks, but we are primarily interested in their value during an emergency response effort.

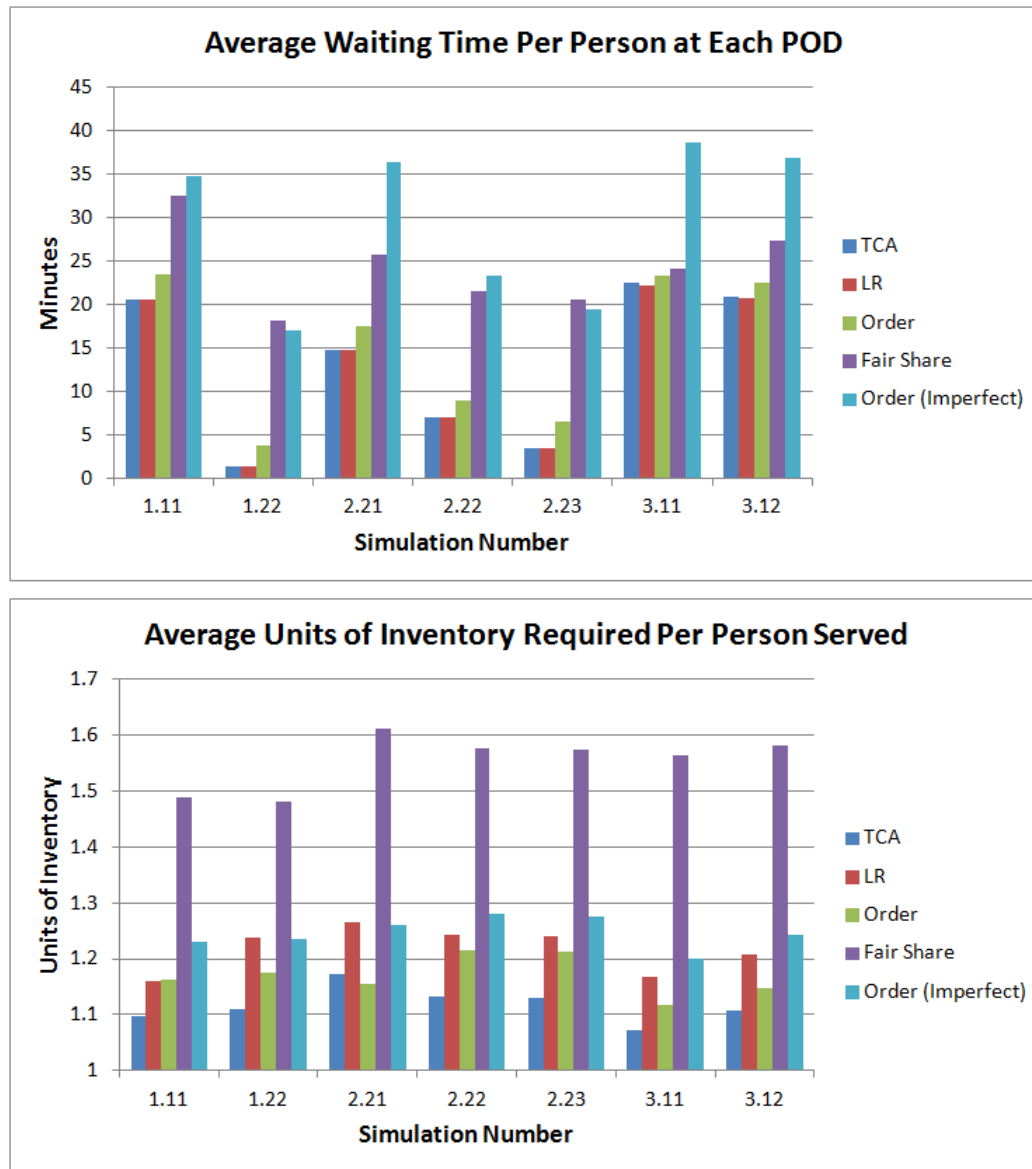


Figure 2.6: Average per patient waiting time and inventory required per person for simulations 1.11, 1.22, 2.21-2.23, and 3.11-3.12.

We presented two lower bounds and used these to show that the TCA and LR allocation methods perform well, particularly in comparison to the allocation methods currently in use by public health authorities. Finally, we demonstrated how our model can help public health authorities answer questions about the value of a command and control system, the potential value of dynamic staffing

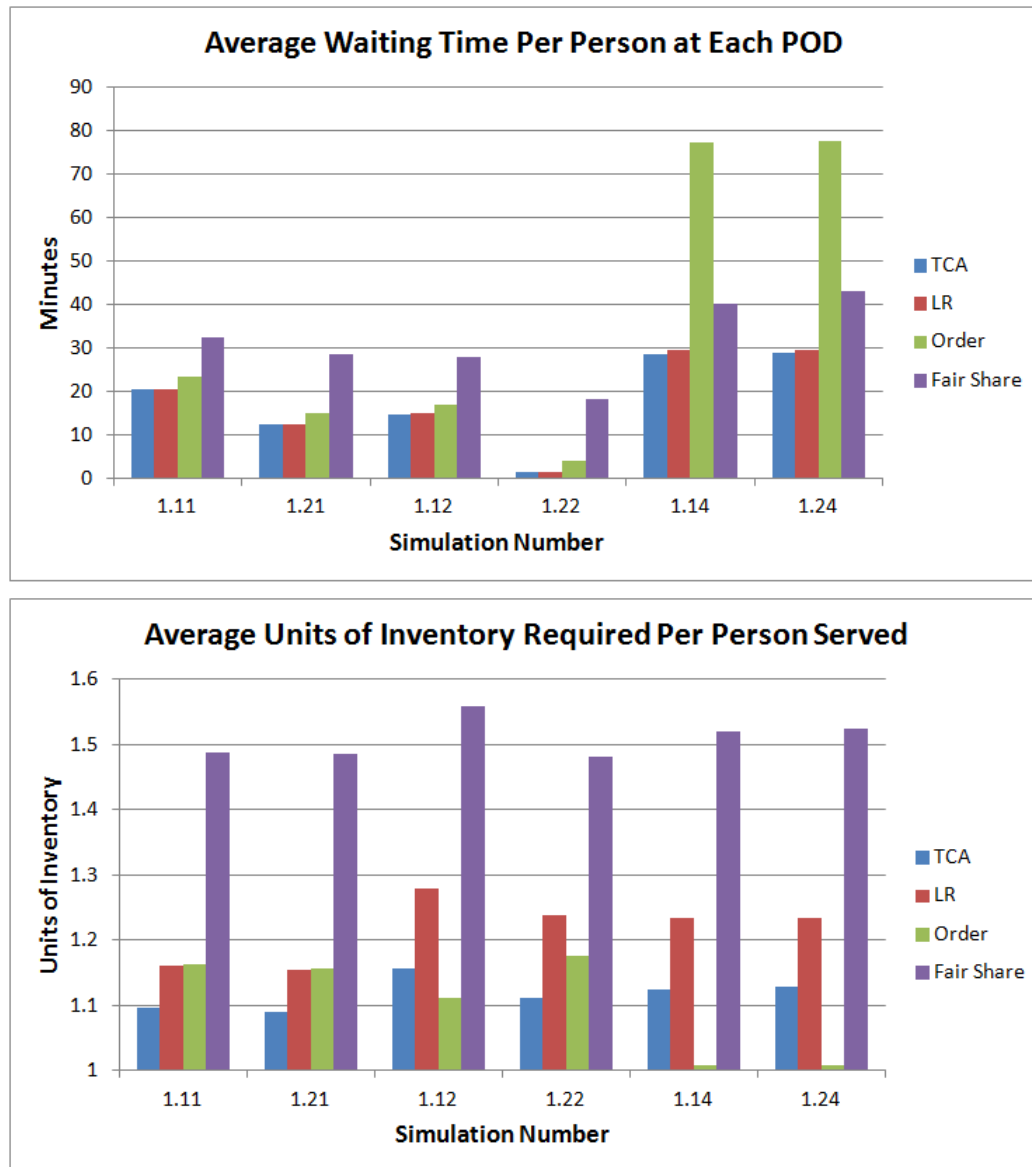


Figure 2.7: Average per patient waiting time and inventory required per person for simulations 1.11, 1.12, 1.14, 1.21, 1.22, and 1.24.

plans, and the benefits of more compact POD networks. Future work will include a more comprehensive simulation study to explore more realistic POD networks and an investigation of the TCA look-ahead period, to determine whether shorter periods may provide similarly strong results. Extensions to the model may include allowing staffing levels to be changed over time in re-

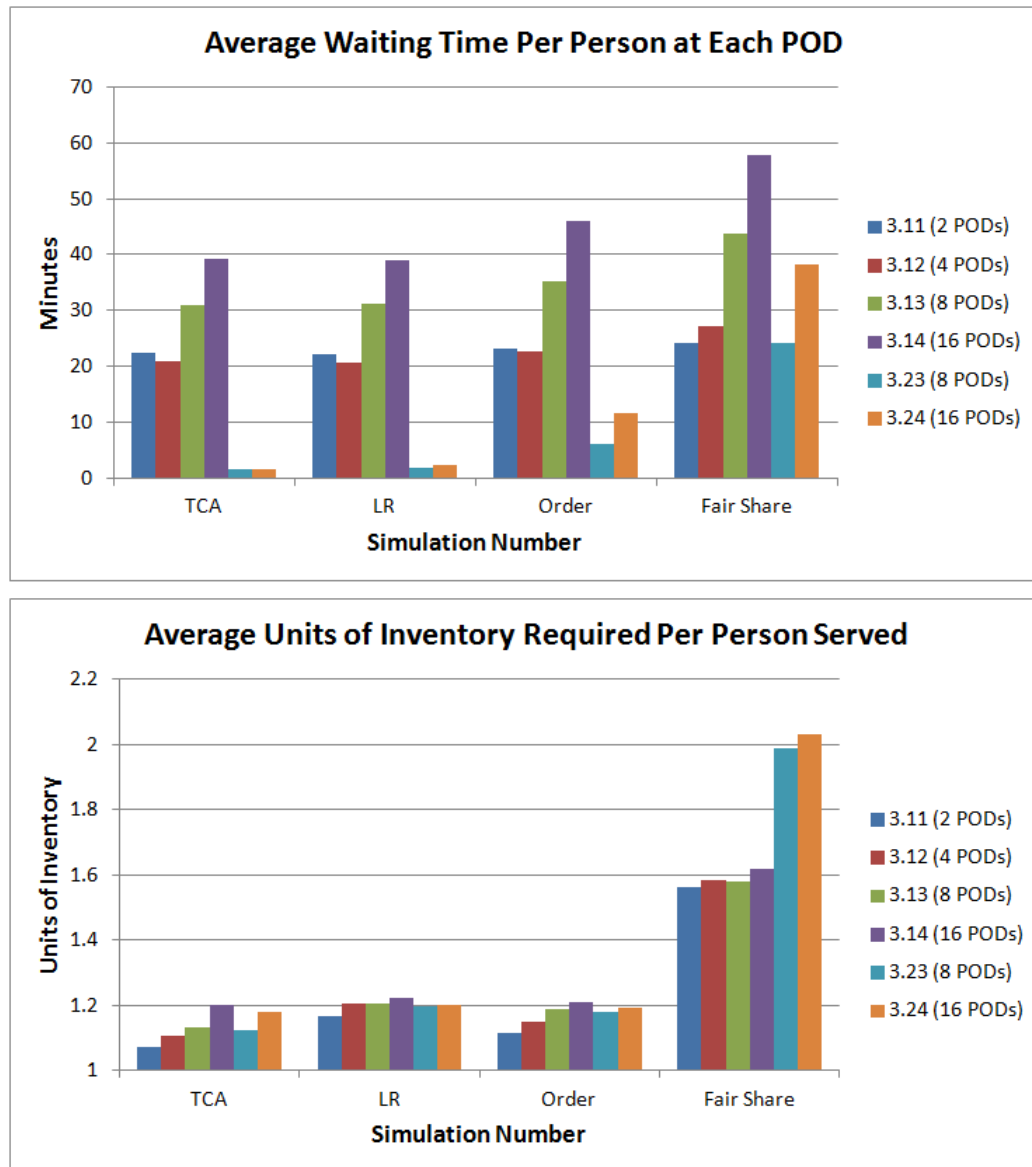


Figure 2.8: Average per patient waiting time and inventory required for experiment 3.

sponse to observed patient demands and the increasing number of staff who may become available once the emergency response is underway.

CHAPTER 3

MASS PROPHYLAXIS SIMULATION MODELS

When making staffing and capacity planning decisions, it may be desirable to consider some portions of the system in greater detail than was included in Chapter 2. In this chapter, we describe two simulation models that focus on particular aspects of the public health emergency response network. The first of these, presented in Section 3.1, is called the Dynamic Point of Dispensing Simulator (D-PODS) and includes a detailed model of staffing and patient flow within a single POD. D-PODS allows users to explore the value of various POD layouts and different staffing strategies under nonstationary patient demand scenarios. The second model, presented in Section 3.2, is called the Emergency Supply Chain Operations Evaluator (ESCOE). It allows users to simulate the operations of a complete emergency response network. One can use ESCOE to explore different network structures and logistical capacities and to evaluate system performance under a variety of conditions.

Both models described below include user-friendly interfaces and are implemented in Excel and Visual Basic for Applications. This choice of implementation makes the models easy to run on any computer with Microsoft Office, making them accessible to the public health officials who are the target user group. User manuals for the models are included in the appendices. In the following sections, we describe the models in detail, give brief overviews of their user interfaces, and show examples that illustrate the types of policy insights that one can draw from them.

3.1 Dynamic Point of Dispensing Simulator

The Dynamic Point of Dispensing Simulator (D-PODS) is a flexible simulation tool that can model and simulate the performance of a wide variety of POD designs. During the last decade, a number of researchers have created software tools to help POD planners better organize, staff, and operate their PODs, as described in Section 1.1 [Hupert *et al.*, 2002, Lee *et al.*, 2006b, Lee *et al.*, 2006a, Aaby *et al.*, 2006]. These tools have all proven useful in planning and operating POD exercises; but all three have assumed either stationary or deterministic patient arrival patterns. D-PODS eliminates this assumption to more accurately reflect the significant uncertainty that is present in any emergency scenario and its impact on system operations.

A POD must perform a set of tasks such as collecting personal information from arriving patients, informing patients about the treatment or prophylaxis that they will receive, basic triage to determine what treatment is appropriate for each patient, and drug dispensing. The exact tasks required and the type of staff who may perform them vary between different states and counties. For example, in some states, antibiotics and vaccines must be dispensed by medically trained personnel, while other states have relaxed this requirement during declared emergencies. The amount of information provided at the POD and the degree of detail in triage or medical evaluation also vary by area. However, basic organization of a POD generally remains the same.

Most PODs consist of a sequence of stations. Staff members at each station perform some task for each patient, such as distributing information or dispensing antibiotics. Patients move through the POD from station to station to receive

complete treatment. For a POD constructed to dispense antibiotic prophylaxis following an anthrax attack, the stations could include a greeting station, where patients receive paperwork from general staff to fill out and some basic information; a triage station, where medically trained staff quickly review each patient's paperwork to ensure that they are eligible to receive antibiotics and then check for basic symptoms of prodromal (early-stage) anthrax; a medical evaluation station, where more experienced medical staff do a more thorough examination of potentially ill patients; and a drug-dispensing station, where patients receive bottles of antibiotics. Different types of patients may require different resources or have different arrival patterns. For instance, more time may be required to serve non-English speakers or mobility-impaired individuals. D-PODS allows planners to describe whatever set of stations has been selected for their area, or to experiment with different types of layouts.

However, even with a fixed POD layout and a set of patient types there is significant uncertainty about the number of patients who will arrive over time. Since every emergency is unique and a large-scale anthrax attack has never occurred, there is no way to know exactly when or how many people will seek care. The rate at which staff can serve patients will vary, as well. The goal of D-PODS is to help users understand the dynamics of patient flow within a POD under a variety of scenarios. We allow POD planners to estimate the consequences of different potential layouts and staffing plans. In the following sections we describe the model and its interface in more detail, and then show how D-PODS can be used to model a particular example scenario and help address policy-related questions.

3.1.1 Model Description

D-PODS is a discrete event simulation model of operations within a single POD. The POD consists of a set of stations; each station performs a particular task (such as triage or drug dispensing) and is assigned staffing levels that may change over time. Patients move from station to station to receive service. We can model many different types of patients, who may require different service levels. Patients of each type arrive to the POD according to independent nonhomogeneous Poisson processes whose means vary by time and by patient type.

The POD has a maximum capacity; if a patient arrives when the POD is full, the patient enters a queue outside the POD. Patients in this queue are admitted to the POD as other patients finish service and leave the POD. Upon entering the POD, all patients enter the queue for the first station in the POD. Each station is modeled as a single queue, multi-server system, in which each staff person assigned to the station is a server. Lee et al. found that triangular probability distributions model service times well in an anthrax dispensing exercise [Lee *et al.*, 2006b]. We also use triangularly distributed service time random variables; the distribution parameters depend on the station and the type of patient being served. We do not model changes in staff efficiency due to learning curves or fatigue.

After completing service at a station, each patient moves to his next station, which is chosen according to a transition probability matrix. Travel time between the stations is negligible; the patient immediately enters the queue for his next station. This process repeats until the patient completes service. Normally this means that the patient will receive antibiotics or, if the patient is ill, will be transported to a hospital. Other exit possibilities, such as being sent to a

special clinic for further evaluation, may also exist in some POD designs.

PODs may remain open continuously or they may close for some portion of each day. When a POD closes, all patients who have entered the POD will be served, including those who are waiting in the queue for the first station. Patients who have not yet entered the POD will be sent away. The POD will remain open for as long as it takes to complete service for the remaining patients inside the POD. For the sake of simplicity, we assume that staffing schedule and patient arrival distributions are identical for each day of the dispensing campaign.

3.1.2 Model Interface

A large number of user inputs are required to describe the POD layout, staffing plan, and patient demand parameters. The interface consists of a sequence of Microsoft Excel worksheets that guide the user through the input process, which is outlined in the D-PODS menu interface and shown in Figure 3.1. The user then runs the simulation, generating a large quantity of output data which is stored in an Access database. The Excel interface allows users to explore these data in both tabular and graphical formats. We will give a brief overview of the tool here, and the User Manual, presented in the appendix, provides more detail. It explains how to get started using D-PODS and how to enter information on the user input sheets, as well as how the simulation outputs can be analyzed and interpreted.

The most time-consuming part of running D-PODS is simply describing the input. The input sheet asks the user to enter the duration of the dispensing

D-PODS Menu

In order to run the program, follow the steps as shown below.
The program may not work if executed in the incorrect order.

Step 1	Construct the model	Model
Step 2	Input arrival rate	Arrivals
Step 3	Input the service time parameters	Service
Step 4	Establish staffing levels	Staffing
Step 5	Enter simulation parameters	
	Number of simulation replications Random Seed	2 150
Step 6	Input Case Name	[Enter Name]
Step 7	Run the simulation	Run
Step 8	View the Results. The results are displayed in both tabular and graph form.	Output Tables
		Output Graphs

Other Options

Select an existing case.	Existing Case
Manage case list.	Manage Cases
Go to Cover Page to Start Over	Start Over

Figure 3.1: The main menu of the D-PODS interface.

campaign and the number of hours that the POD will be open each day and to define the structure of the POD, including the number of stations, their names, and the station transition probability matrix that determines each patient's path from station to station within the POD. The next input sheet allows the user to describe the patient arrival patterns by defining a set of time intervals and

the expected number of patients of each type to arrive in each interval. We assume that the mean arrival rate is constant during each interval. Next, the user defines the triangular service time distributions for each station. One of the patient types is assigned to be the “base type;” other types may be assigned a “service time increase factor,” which adjusts their expected service times accordingly without requiring the user to define new distributions for each type. The last input step involves setting the staffing levels for each station over time. This sheet includes a simple staffing calculator that estimates staffing requirements using queueing approximations from [Buzacott & Shanthikumar, 1993], which would provide optimal staffing levels in a stationary setting. Users may use, modify, or ignore these advised staffing levels.

Finally, the user may choose some number of simulation replications and run the model. Detailed output is saved in a Microsoft Access database, but users may easily view some of the results in tables that display summary statistics describing patient arrival rates and throughput, as well as waiting and service times by station. Graphs of queue length, staff utilization, and patient arrivals over time are generated by station and for the entire POD. In the following section, we will describe a set of sample inputs and show the kinds of insights that one may draw from D-PODS.

3.1.3 Staffing Policy Implications

We now want to explore the impact of nonstationary patient arrival patterns on a POD’s operations and determine the value of instituting nonstationary staffing policies in response to observed demands. We will also investigate the

relationship between the number of PODs in a response network and the total staffing requirements across the network and explore the importance of constructing an effective command and control system.

In the examples discussed below, we represent and evaluate the performance of a large POD dispensing antibiotics in response to an inhalational anthrax attack. We run the model for 48 hours, which is the federal goal for completing antibiotic prophylaxis following an anthrax attack, as discussed in Chapter 1. We assume that the POD closes for two hours each day to allow for restocking and cleaning, but otherwise remains open to serve as many people as possible. The POD is expected to serve up to 11,000 people per day, since a service rate of 500 people per hour is considered reasonable for medium-sized PODs in many cities [Hupert, 2011]. We will vary the rate at which these people arrive over time, but the total expected demand will remain constant in all of the examples below.

The POD modeled will have a standard layout with four stations: Greeting, Triage, Medical Evaluation, and Drug Dispensing. To keep our discussion as simple as possible, we have run the simulations described below with only one patient type. Most patients will not require medical evaluation, but 5% of individuals will show symptoms of anthrax at the Greeting station and be sent for evaluation. At the more thorough Triage station, an additional 5% of patients will be identified as symptomatic and be sent for evaluation. However, 95% of those who were sent to the Medical Evaluation Station will still be sent to the Drug Dispensing station and subsequently released; only 5% of those evaluated will be taken to a health center for treatment. Figure 3.2 shows a diagram of patient flow within the POD. The staffing levels at these stations will vary over

time; we will describe the potential staffing strategies in the examples below.

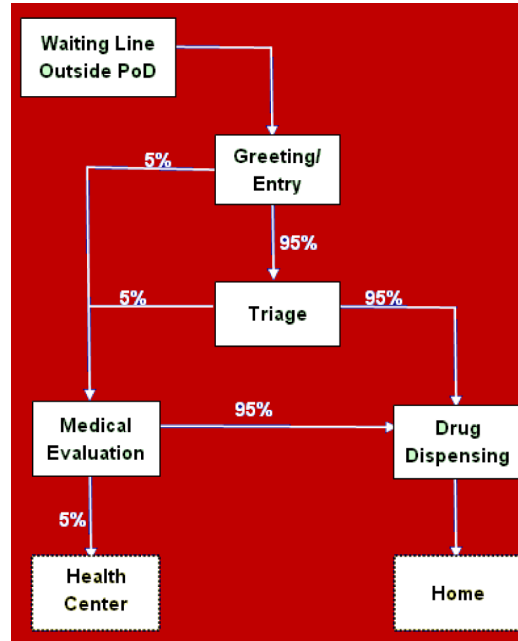


Figure 3.2: POD patient flow diagram.

For ease of data analysis, we assume that the POD is sufficiently large so that its total capacity will never prevent patients from entering the POD. The service time parameters at each station are based on discussions with the RAND Corporation and CDC staff; they are very similar to the numbers used in an anthrax dispensing exercise by Lee et al. [Lee *et al.*, 2006b]. These values are summarized in Table 3.1 below along with the patient service path transition probabilities.

Parameter Name	Value	Units
Transition Probability (Greeting to Medical Evaluation)	0.05	(n/a)
Transition Probability (Triage to Medical Evaluation)	0.05	(n/a)

Transition Probability (Medical Evaluation to Hospital)	0.05	(n/a)
Service time for Greeting/Entry station	Triangular(0.25,0.50,1)	minutes
Service time for Triage station	Triangular(1,2,3)	minutes
Service time for Medical Evaluation station	Triangular(2.5, 5, 10)	minutes
Service time for Drug Dispensing station	Triangular(0.5, 1, 2)	minutes

Table 3.1: Simulation parameter values.

The Impact of nonstationary Patient Arrival Patterns

The goal of our first set of simulation experiments is to quantify the impact of nonstationary patient arrival patterns on POD performance. As mentioned earlier, most POD planning tools assume a constant rate of patient arrivals and constant staffing levels during all operating hours. However, this type of constant, predictable demand seems unlikely in an emergency scenario. Spikes and drops in arrival rates could occur for a variety of reasons. Arrivals might increase over time as more people learn about the POD, or decrease as other PODs open. Arrivals might spike in response to the availability of public transit or drop due to inclement weather. We consider three scenarios, labeled *A*, *B*, and *C*, in which patient demand varies throughout the day, as shown in Figure 3.3, and we will simulate each with two different staffing strategies.

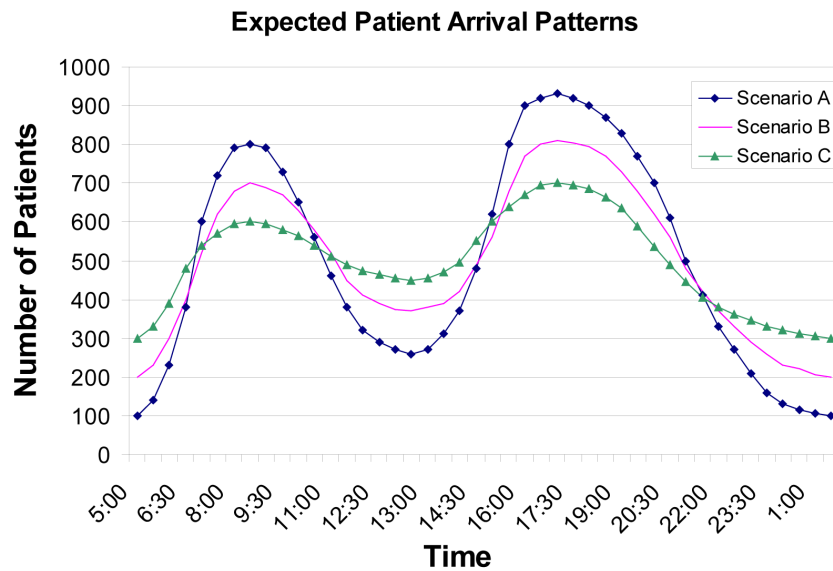


Figure 3.3: Expected patient arrival scenarios.

If we were to staff to accommodate patients arriving at a constant expected rate of 500 per hour, to yield a total expected demand of 11,000 people over the 22 hours during which the POD is open, then we would expect to underutilize the staff during periods when the demand is low and overwhelm the staff when the demand peaks. We would also expect to see long waiting times when the demand peaks in any of the three scenarios shown in Figure 3.3. However, if we could accurately predict the average patient arrival rate over time and staff accordingly, we would expect to provide much better service to patients. We call any staffing plan that changes over time a “dynamic staffing plan.” We calculated a “constant staffing plan” using the simple queueing calculation built into D-PODS to determine approximate the staffing levels for a constant patient arrival rate of 500 people per hour. To define the dynamic staffing plans, we used the same queueing calculation, but allowed the staffing levels to change every two hours, and we staffed to accommodate the maximum patient arrival

rate in every two hour period.

Figure 3.4 shows that our predictions about system performance are correct on both counts. When a dynamic staffing plan is instituted, the average patient time spent in the POD decreases by a factor of 10 to 18, depending on the arrival scenario, while the number of staff-hours required only increases by 10 to 14%.

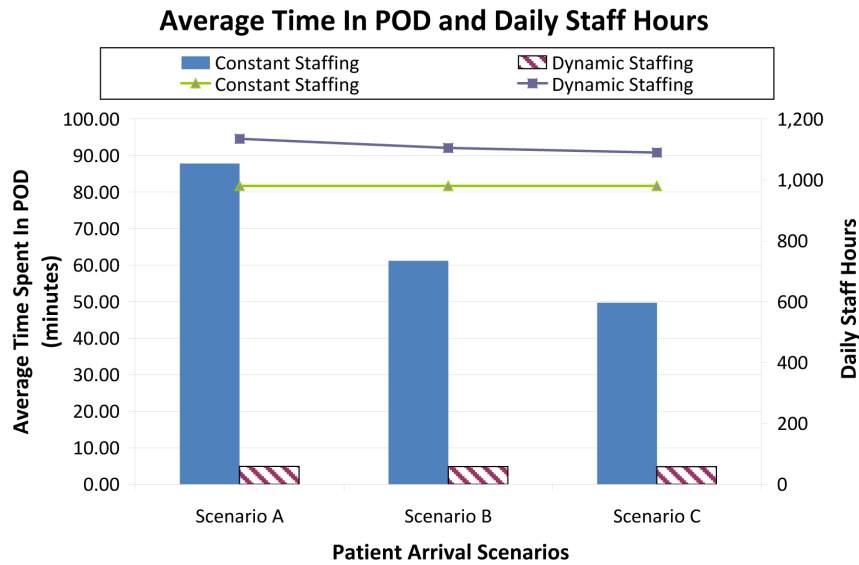


Figure 3.4: Average patient time spent in the POD and staff-hours required each day, given constant or dynamic staffing plans.

However, the averages displayed in Figure 3.4 do not show the whole story; we also need consider how events unfold over time. Figures 3.5 and 3.6 show the patient queues under both constant and dynamic staffing plans for patient arrival scenario A at the greeting and triage stations, respectively. When the constant staffing plan is in effect, the queue at the greeting station becomes very large when patient demands peak, while it remains at a reasonable size when the dynamic staffing plan is in place. The queue length for the triage station, shown in Figure 3.6, is slightly smaller under the constant staffing plan, but

does not grow significantly for either plan. When the constant staffing plan is in place, the Greeting station acts as a bottleneck and buffers the rest of the POD from the unmanageable demand. The excessively long queues at the Greeting station are worrisome not only because they significantly inconvenience individuals, but long waits could also result in balking or crowd control problems.

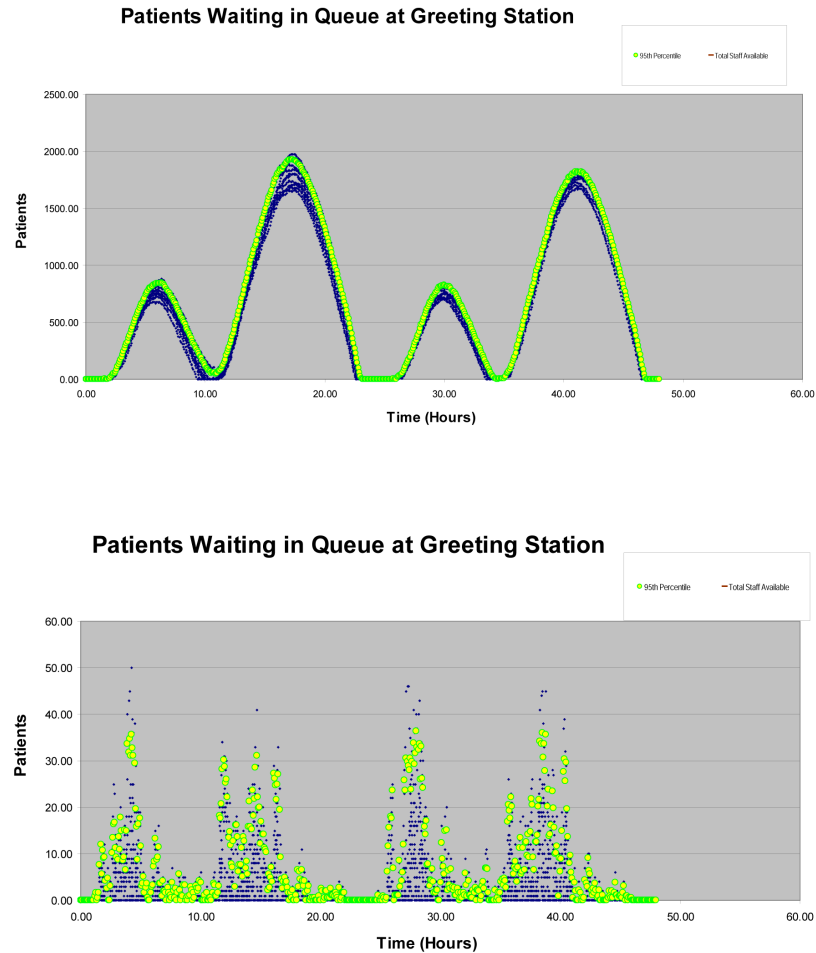


Figure 3.5: The length of the queue at the Greeting Station for Scenario A for Constant Staffing (top) and Dynamic Staffing (bottom). The small blue dots show the queue lengths from 10 simulation replications for every five minute interval; the yellow-green circles show the 95th percentile queue length for each five minute interval.

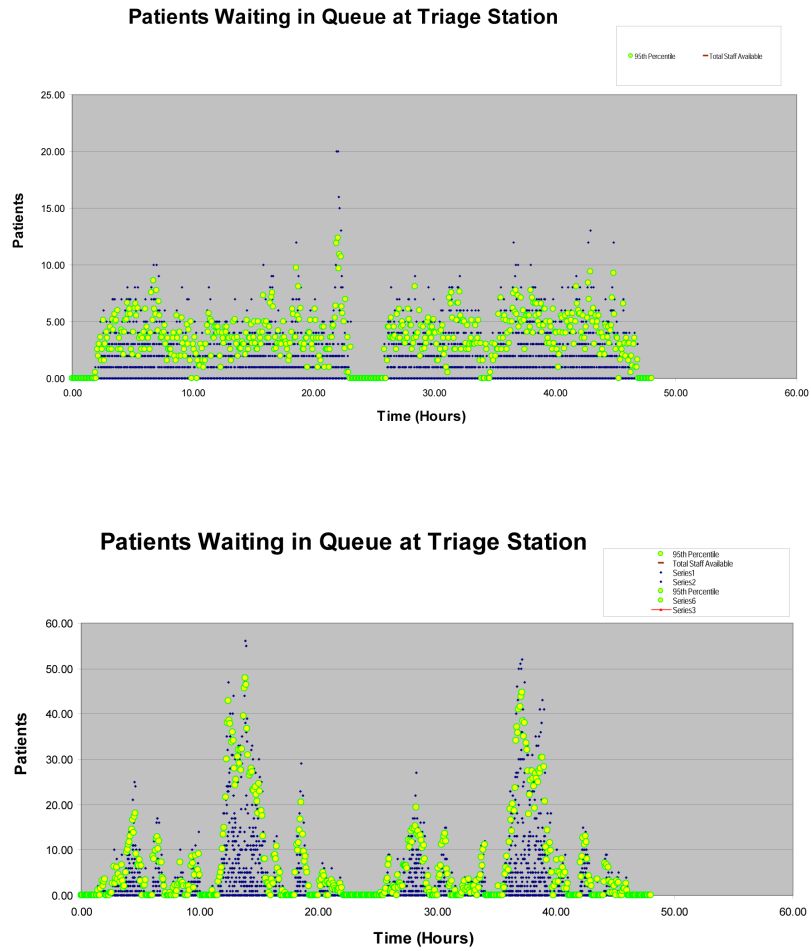


Figure 3.6: The length of the queue at the Triage Station for Scenario A for Constant Staffing (top) and Dynamic Staffing (bottom). The small blue dots show the queue lengths from 10 simulation replications for every five minute interval; the yellow-green circles show the 95th percentile queue length for each five minute interval.

Clearly, the ability to modify staffing levels over time as patient arrivals vary is important to a POD's ability to serve patients efficiently. But our examples have assumed that planners can predict the true expected patient demand rates over the time horizon, which is unlikely. In the following example, we will consider how a system might perform when there is a time lag in adjusting staff

to patient demands.

The Value of Flexible Staffing

Suppose that staffing levels are determined by forecasted demand levels, which are calculated dynamically throughout the day. We will allow staffing levels to change every two hours, but during the two hour periods, staffing levels remain constant. We consider four staffing plans. In Plan 1, we forecast perfectly. That is, we know the exact expected patient arrival rates for the next two hours, and we staff to accommodate the maximum arrival rate; this is equivalent to the dynamic staffing plan discussed in the previous section. In Plan 4, we know the expected total number of patient arrivals for the next two hours, and we staff assuming that they arrive at a constant rate. In Plans 2 and 3, we use a very simple forecasting mechanism; we will assume that the patient demand at the time of forecast will remain constant for the foreseeable future. In Plan 2, we observe demand and determine staffing levels 30 minutes before these new staffing levels will go into effect. In Plan 3, we have a delay of 1 hour. In general, we will expect Plans 2 and 3 to under-staff when the patient arrival rate is increasing and over-staff when it is decreasing.

We ran simulations of the POD operating under each staffing plan with patient arrival Scenario A, defined in Figure 3.3. In Figure 3.7 we see the average patient waiting times for each of the staffing plans. Implementing staffing plan 3 results in the longest patient waiting times, but notice that these are still significantly better than the waiting times observed for the constant staffing plan discussed in the previous section (see Figure 3.4). Dynamically adjusting staff, even using a naive forecasting method like those used in Plans 2 and 3, de-

creases average patient waiting times by more than half.

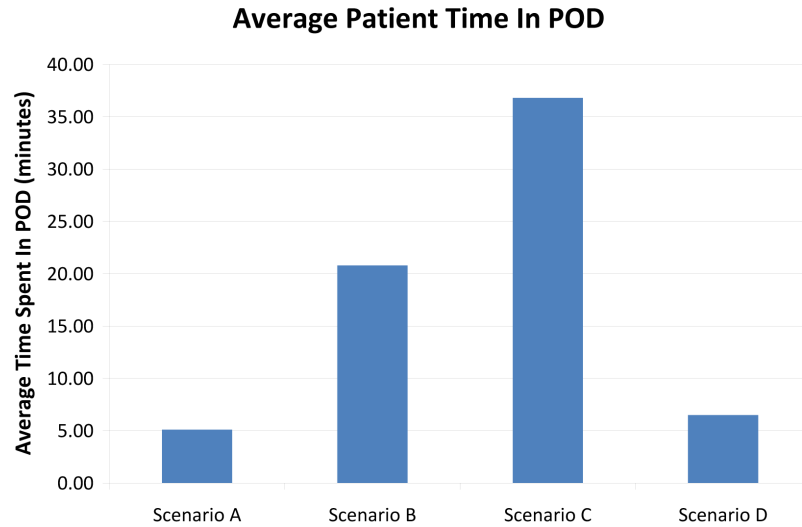


Figure 3.7: Average patient time spent in the POD for patient arrival Scenario A for each of the four staffing plans.

Figure 3.7 seems to imply that Plans 1 and 4 are almost equally effective. This would be impressive, since the forecasting mechanism used to set staffing levels for Plan 4 requires less information about the actual expected patient arrival rates than the perfect forecasting used by Plan 1. However, Figures 3.8 and 3.9 show the dynamics of these two simulations, and we see that the queues at both the Greeting and Triage stations grow significantly longer under Staffing Plan 4. At the Greeting Station, the queue grew longer than 100 people in many of the replications under Plan 4, but under Plan 1 the queue seldom exceeded 30. At the Triage Station we see that the queue sometimes grew longer than 200 under Plan 4, but generally remained stable under Plan 1.

We show these results only for patient arrival Scenario A, which is the most variable of the three patient arrival patterns considered. The performances of all of the staffing plans improve under the less variable patient arrival pat-

terns, Scenarios B and C. All four plans would yield identical outcomes if patient arrival rates were constant. If the patient arrival rates changed exactly when staffing shifts changed and remained constant during each staffing interval, then Plans 1 and 4 would provide identical staffing levels. In general, the performance of the system under Plans 1 and 4 will be similar, but Plan 1 will yield slightly better POD performance and Plan 4 will yield slightly higher staff utilization rates, since Plan 1 staffs for the maximum expected patient arrival rate during each staffing interval, while Plan 4 staffs for the average.

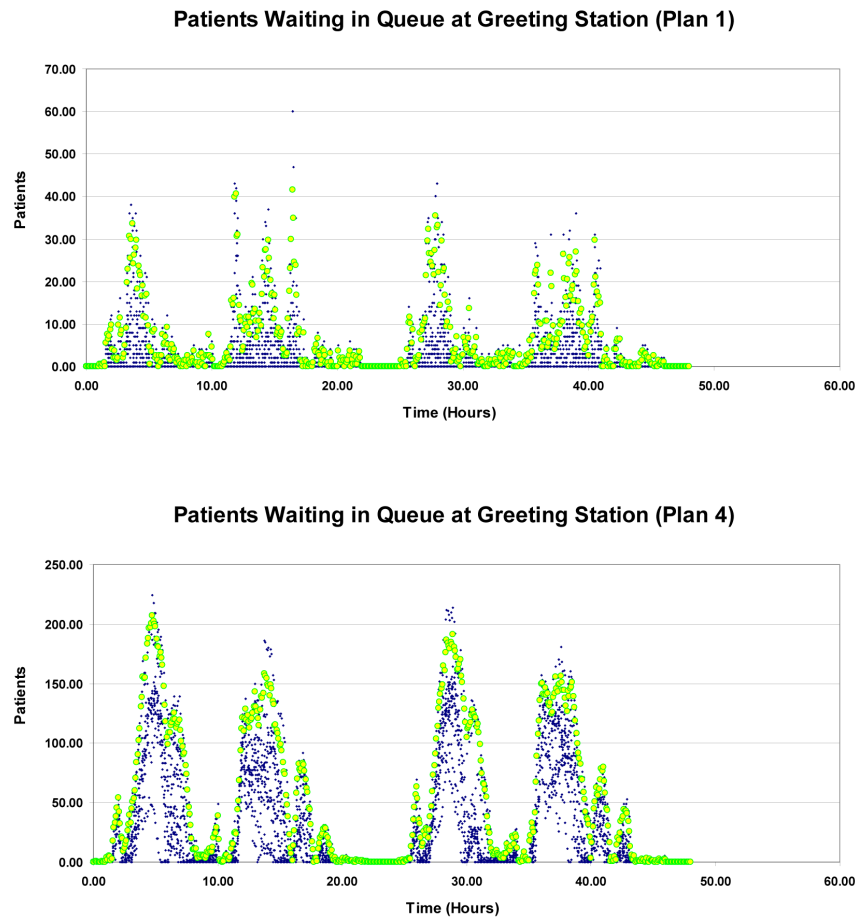


Figure 3.8: Queue lengths at the Greeting station for Plans 1 and 4 with patient arrival Scenario A.

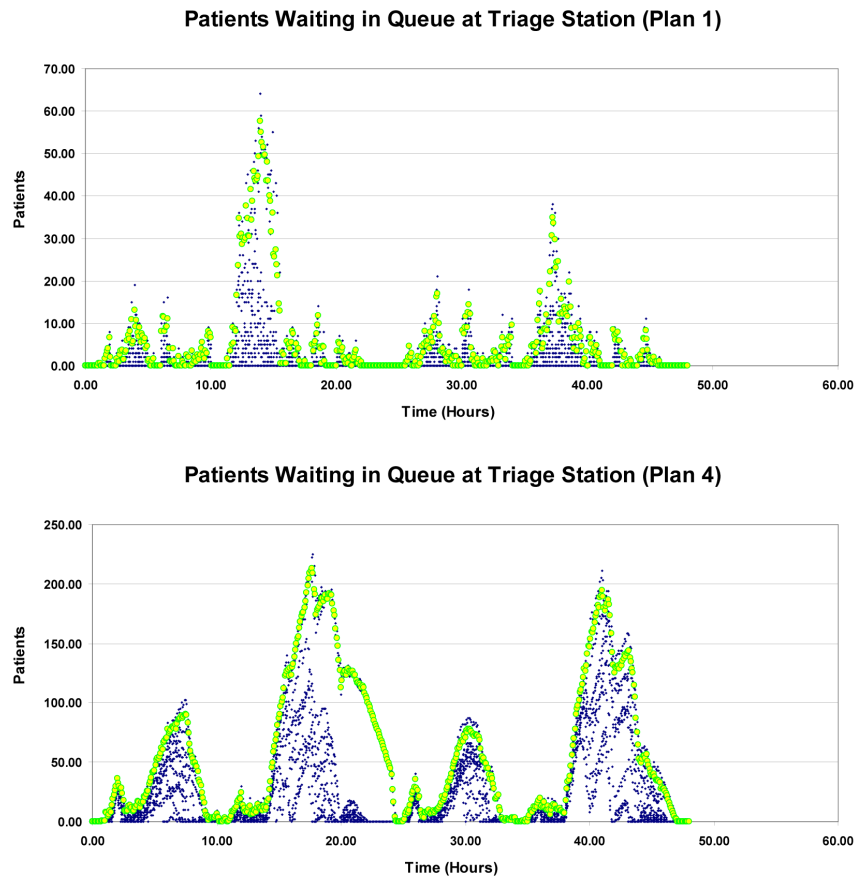


Figure 3.9: Queue lengths at the Triage station for Plans 1 and 4 with patient arrival Scenario A.

It is clear that the quality of patient demand forecasting can have a significant impact on patient waiting times and queue lengths. Less variable, or “flatter,” demand curves are much more forgiving of the forecast quality. Some cities, such as New York City, have constructed online tools to tell patients the current waiting times at various PODs throughout the city [Starr, 2012]. The goal of such tools is to encourage individuals to go to under-utilized PODs or wait to seek treatment until queue sizes have dropped, thereby flattening the demand across the POD network. The results shown here strongly support the development of such information-sharing strategies, which could help provide good POD performance across the dispensing network, especially if combined with simple forecasting methods and a dynamic staffing system. But dynamic staffing plans are only useful if there is sufficient staff available to fulfill these plans. It is also essential to design a response network that can be operated by the limited number of staff who will be available during the first hours after an emergency occurs. In the following section, we show how the overall design of the POD network affects the number of staff that will be necessary to run it.

The Impact of Network Design on Staffing Requirements

There is pressure on some local public health departments to open a very large number of PODs so that individuals in the community will not need to travel very far to receive treatment [Starr, 2012]. We ran two simple experiments to show the consequences of increasing the number of PODs in a dispensing network. Consider a network of 10 small PODs, which each serve an average of 500 patients per hour. We compare this with a network consisting of only one large POD which serves an average arrival rate of 5,000 patients per hour.

Both systems will experience the same total expected daily patient demand. For each system, we used the queueing formulae built into D-PODS to calculate the staffing levels necessary to limit average patient waiting times to five minutes at each station within the PODs.

We find that each of the small PODs requires 984 staff-hours to provide an average patient waiting time of 6.72 minutes; this gives a total of 9,840 staff-hours to operate the POD network for two days. The large POD, on the other hand, requires only 8,664 staff-hours to provide an average patient waiting time of 6.28 minutes. That is, almost 15 percent more staff-hours are required to provide slightly worse service to patients in the 10 POD network, compared to the single POD. Furthermore, additional fixed costs would likely be required to operate a POD network that is more widely-distributed over many locations.

There are many factors that affect capacity-planning decisions for PODs, but in general one can show that increasing the number of PODs also increases the number of staff-hours required to provide a desired level of service to the same number of patients. Clearly, then, if the number of staff-hours is limited, having fewer larger PODs would provide the best service for patients. For large emergencies, opening only a few PODs would be infeasible. However, an effective command and control system could facilitate the creation of a virtual single, large POD, if it allowed for monitoring and adjusting staffing levels and service capacity throughout the system. As we have demonstrated, the number of staff required to meet patient needs in a timely manner will be reduced substantially if such a virtual system could be implemented.

3.1.4 Discussion

We used D-PODS to demonstrate that nonstationary and uncertain patient arrival patterns have a significant impact on POD performance; that dynamic staffing policies can greatly improve patient waiting times and queue lengths; and that POD networks require fewer staff-hours when fewer large PODs are used, compared to many smaller PODs. Patient waiting times and queue lengths are minimized in an environment in which POD staffing levels are appropriately calculated to match expected patient arrival rates. These results strongly suggest that a responsive and robust POD system can exist only if a command and control system is in place that can balance staff and other resources throughout a dispensing network. If such a system is not developed, then either significantly larger staffs must be employed or patients will experience very long waiting times; neither of these outcomes is desirable. We also note that these observations are not limited to the setting of staff levels. Greater amounts of other resources such as medical supplies, equipment, and space will all be needed if scarce resources are allocated ineffectively. If we could predict patient arrival rates with great accuracy and allocate staff and other resources accordingly in advance, then the need for an effective command and control system would be reduced. But there is so much uncertainty regarding where and when spikes in patient demand will arise that developing mechanisms to balance loads throughout the POD network by moving staff and other scarce resources among PODs should be considered essential.

D-PODS can help public health planners better understand the operations of PODs and the importance of developing flexible plans that account for the inherent uncertainty of emergency scenarios. A planner may experiment with

D-PODS to construct potential staffing plans for different types of demand profiles and to estimate POD service rates more accurately over time. In the next section we describe a model that allows planners to consider uncertainty and interdependencies in the full emergency response system.

3.2 Emergency Supply Chain Operations Evaluator

The Emergency Supply Chain Operations Evaluator (ESCOE) is, like D-PODS, a tool that can help public health planners better understand how the emergency response system will work under a variety of uncertain conditions. ESCOE focuses on the flow of inventory from the SNS warehouses to the RSSs to the PODs. The goal of ESCOE is to allow policy makers to study the global consequences of supply chain designs and operating policies.


Setting up and operating the emergency response supply chain is a difficult task, due in large part to the large number of organizations and individuals involved. During an emergency many of these will be called upon to perform tasks far removed from their everyday jobs. For example, office workers may staff warehouses, and public health officials might become managers of dispensing clinics. As mentioned earlier, exercises are difficult to organize, time-consuming, and very expensive to conduct. The exercises that do take place are often incomplete, and they are usually carefully planned and scripted, thereby minimizing the elements of surprise and uncertainty that are part of real emergencies. Furthermore, exercises seldom involve more than one organization, and they never involve all three echelons of the supply chain. The SNS performs its exercises and evaluations, while states run separate events to practice

operating RSSs, and local authorities are often responsible for figuring out how to operate their PODs. To organize an exercise involving all three levels of the distribution network would be overly burdensome and expensive. Instead, simulation models such as ESCOE offer more efficient and affordable opportunities to study the workings of the emergency response network.

The need for a simulation tool like ESCOE is magnified by the fact that the many supply chain design decisions are made by many different individuals. While some efforts have been made to coordinate these decisions, current plans allow the various parts of the system to operate largely independently. Problems could easily arise as a result of insufficient coordination. For example, at PODs, there may be sufficient staff availability to respond to expected patient demands, but if inventory arrives later than expected from the RSS, queues of patients could build up, overwhelming the POD staff. Over the course of the dispensing campaign there may be sufficient staff, transportation capacity, and inventory to accommodate the cumulative patient demand, but the system may still perform poorly if these resources are not carefully coordinated. ESCOE brings the operational details of the system together so that policy makers and planners can explore how the choices they make might affect the overall performance of the supply chain. In the following sections we describe ESCOE and its underlying assumptions in greater detail, and we provide a detailed example of how ESCOE can be used to model an emergency scenario and identify some attributes of effective response system designs.

3.2.1 Model Formulation

Like D-PODS, ESCOE is a simulation tool built in Visual Basic with an Excel-based interface, shown in Figure 3.10. Users enter input parameters through a step-by-step process that helps them describe the emergency response network that they would like to evaluate. The user describes the physical distribution network and the basic capacity constraints that control how inventory flows from one stage to the next.



Menu

Step 1	Total Number of Time Periods:	25
	Length of Each Time Period (hours):	2
	Total Number of Hours in Simulation:	50
Step 2	Construct the Network	Network
Step 3	Describe the Lead Times	Lead Times
Step 4	Describe the Inventory	Inventory
Step 5	Describe the SNS	SNS
Step 6	Describe the FDSs	FDSs
Step 7	Describe the RSSs	RSSs
Step 8	Describe the POD Types	POD Types
Step 9	Describe the Simulation Experiment	Simulation
Step 10	Run the simulation	Run Simulation
Step 11	View the Results. The results are displayed in both tabular and graph form.	Output Tables
		Output Graphs

Figure 3.10: ESCOE interface main menu.

The SNS maintains about ten main warehouses across the country, but information regarding their exact number and locations is not publicly available. Since the SNS is centrally controlled, we model all of the SNS warehouses as a single central location, just as we did in Chapter 2. We continue to abuse definitions slightly and refer to this single central location as “the” SNS or the SNS

warehouse. In addition to the single SNS warehouse, our model allows for the inclusion of Forward Deployed Stockpile (FDS) locations. An FDS is a smaller SNS-maintained warehouse that is strategically located near a major population center so that medical supplies can be delivered to that area faster than the expected 12 hour time horizon. Unlike inventory stored at the main SNS warehouse, which could be allocated to any RSS, the FDSs would only serve one or two nearby RSSs. Unlike the SNS, the FDSs would not be resupplied by vendors during the emergency; their sole purpose would be to provide rapid service to their RSSs in the early stages of an emergency. Currently the SNS is creating some FDSs in response to strong requests from large cities such as New York City and Los Angeles. Thus, we include FDSs in our model to allow planners to explore their potential operational value. Muckstadt and Caggiano provide a model that estimates the monetary costs of operating an FDS over time [Caggiano & Muckstadt, 2010].

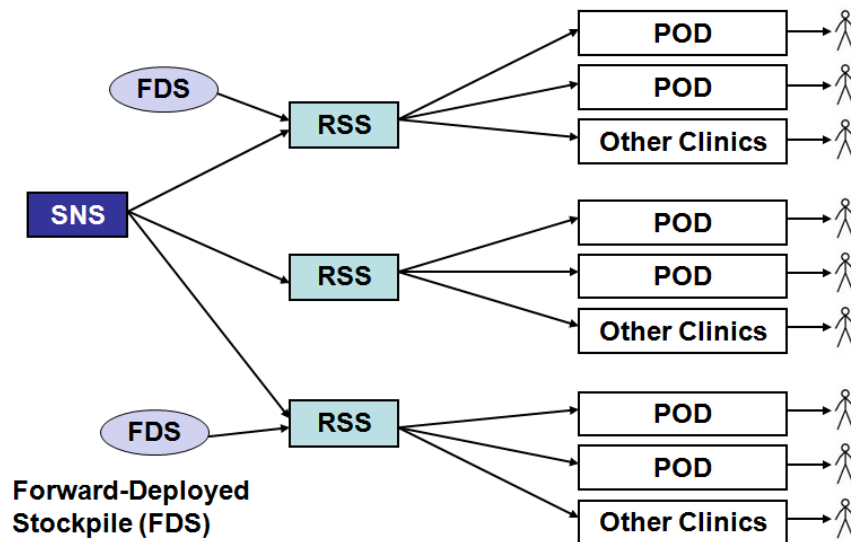



Figure 3.11: Emergency supply chain diagram, including Forward Deployed Stockpiles (FDSs).

Figure 3.11 shows a diagram of the distribution network. The network includes one or more RSSs, as well as at least one POD per RSS. The SNS central warehouse serves all the RSSs. FDSs may serve multiple RSSs, but an RSS can be served by at most one FDS. Each RSS serves a set of PODs, disjoint from the PODs served by other RSSs. Thus, the supply chain network can be represented by an acyclic directed graph; it is not necessarily a tree since RSSs may be served by an FDS as well as the SNS warehouse. Once established, the network topology remains fixed during the simulation. Figure 3.12 shows the ESCOE interface that allows users to input parameters that describe this network topology. After completing the worksheet, the user must click on the buttons at the left side of the screen to move to the next steps of describing lead times, inventory, and the characteristics of the SNS, FDSs, RSSs, and PODs, before finally running the simulation.



Cornell University

Step 1: Construct the Network

Main Menu

Construct Network

Define Lead Times

Describe Inventory

Describe SNS

Describe FDSs

Describe RSSs

Describe POD Types

Describe Simulation

Step A	Input the number of FDSs (maximum of 15)	2
	Input the FDS names	FDS names
Step B	Input the number of RSSs (maximum of 15)	2
	Input the RSS names	RSS names
Step C	Input the number of POD Types (maximum of 10)	3
	Input the POD Type names	POD types
Step D	Input RSS, POD relationships	FDS/RSS
Step E	Input RSS, POD relationships	RSS/POD
Step F	Return to Main Page	Return

FDS	1	2
Name	fds1	fds2

RSS	1	2
Name	rss1	rss2

POD Types	1	2	3
Names	pod1	pod2	pod3
Amount	10	30	20

Input a 1 in the FDS/RSS Relationship Table if the FDS serves the RSS, and input a 0 in the FDS/RSS Relationship Table if the FDS does not serve that RSS.

	rss1	rss2
fds1	1	0
fds2	0	1
SNS	1	1

RSSs

Figure 3.12: The ESCOE interface for describing the distribution network.

Time in the simulation is divided into discrete time periods. In each period

a particular set of events takes place. At each upper echelon location in the distribution network (SNS, FDSs, and RSSs), inventory is allocated and shipped out, and any newly arrived shipments are processed. At the PODs, shipments are received, and patients arrive and either receive service during the period or, if there is insufficient service capacity or inventory, enter a queue to wait until these resources become available. The number of patient arrivals during each time period is randomly drawn from a non-homogeneous Poisson distribution. We assume that these patients arrive at a constant rate throughout a time period, but this rate may vary from period to period. Unlike D-PODS, we do not simulate the internal workings of PODs in detail. Each POD's patient service capacity during a time period is triangularly distributed, with distribution parameters that are functions of the number of staff assigned to the POD.

The ESCOE model also includes some transportation and warehouse logistics. There are limits on the number of pallets that can be loaded and unloaded at each location in each time period. All of these constraints may vary over the course of the simulated time horizon. For the SNS warehouse, FDSs, and RSSs, the number of shipments that can be sent out in each time period is constrained, and the number of pallets that may be sent in each shipment is limited. The transportation lead times between locations in the network are also triangularly distributed random variables, whose parameter values may differ for each pair of locations.

Inventory Allocation Policy

Most state and local emergency response plans call for a standard order-up-to policy to be used in allocating inventory. The par levels, as they are called, are

predetermined for each location in the distribution network, and whenever inventory falls below the par level, a shipment will be sent to raise the stock to par level. In ESCOE, we set par levels to equal the expected demand over the lead time plus safety stock equal to two standard deviations of the estimated demand over the lead time. This is a standard method for setting inventory levels in industry. It is optimistic to assume that the distribution of patient demand over a lead time is known with accuracy, but doing so allows us to ensure that the results given by ESCOE are the best possible outcomes; problems will not arise due to quirks of the inventory ordering policy.

ESCOE also includes provisional allocation rules in case there is insufficient stock in the system to bring all of the locations up to their desired inventory levels. A fair share policy has been implemented, which attempts to distribute the available inventory to locations that fall below their par levels so that the probability of patient demand exceeding stock is approximately equal at each location. Many states do not have contingency plans for the case where RSSs run out of inventory. Hence, it is likely that materials would simply be sent out in a first come, first served fashion until everything is gone, which could leave some PODs unable to operate effectively.

This combination of the order-up-to and fair share policies is used by many commercial and industrial supply chains for planning purposes, and it represents the most sophisticated type of plan currently in use by state and local health departments. However, future work should include adding some of the more complex inventory allocation policies given in Chapter 2. We now illustrate how ESCOE, using the inventory policies discussed here, can be used to model a particular emergency scenario.

3.2.2 Example Scenario: Modeling an Inhalational Anthrax Attack

To demonstrate the functionality of ESCOE we will consider a moderately sized inhalational anthrax attack that affects both southeastern New York State and western Connecticut. As in Section 3.1, we consider a 48 hour time horizon, and we separate it into 24 two-hour time periods. Each state would likely open one RSS warehouse and a number of PODs. We will simulate the system both with and without one FDS per state. In responding to a large attack, Manhattan alone might open as many as 100 PODs, but for this smaller example we will assume that New York opens 50 PODs and Connecticut opens 10. PODs may come in many sizes and each state's plan varies, but for simplicity we assume that there are three basic types of PODs: small, medium, and large, which are staffed to serve up to 2,500, 5,000, and 10,000 patients per day, respectively. We will suppose that New York opens 8 large PODs, 22 medium PODs, and 30 small PODs, while Connecticut opens 2 large PODs and 8 medium PODs.

Western Connecticut is a relatively small region, so we assume that a truck could reach any POD from the RSS within a single two-hour period. New York is a much larger state, but it is not unreasonable to suppose that the RSS serving New York City and southeastern New York State would be located within two hours of the PODs that it serves, so we similarly assume that the lead time from the New York State RSS to its PODs is one time period. For similar reasons, we assume a single period lead time from the FDSs to their respective RSSs, when FDSs are included in the simulation. We allow a twelve hour lead time from the SNS to both RSSs in accordance with the federal goal, although this may be overly pessimistic since in reality the SNS may maintain a stockpile near New

York City. Although the modeling environment permits these lead times to be random, in this example we assume that they are constant. For the purposes of this simulation, we suppose that the lead times include all of the necessary inventory processing time, as well as loading and unloading delays, in addition to the actual travel time from one location to another.

In response to an inhalational anthrax attack the CDC plans to provide all affected individuals with a course of antibiotic prophylaxis, so the PODs will distribute “unit-of-use” bottles of pills. There are two or three types of antibiotics that would be dispensed in practice, but we include only one type in this example due to the current limitations of ESCOE; future work will include extending the model to include multiple inventory types.

We assume that antibiotics are shipped in cases which contain 300 unit-of-use bottles, and pallets hold 32 cases. Inventory initially at the SNS or FDSs must be in pallet-sized quantities. In this example we limit the number of pallets per truck to 26, which is the number of pallets that a standard 53 foot trailer can hold without stacking, since material-handling equipment may be limited at the RSSs and unavailable at PODs. We assume for simplicity that all trucks used in the simulation are of the same size.

The set of parameters described above defines a distribution network topology and some of its basic characteristics. However, we have not yet discussed how we might define the storage limits or the inventory loading and unloading constraints at the SNS, FDSs, or RSSs, nor have we addressed patient demands or service capacities at the PODs. All of these will be defined in the following section, as we illustrate how to set these parameters in order to address a particular policy-oriented question.

Using the Inhalational Anthrax Scenario to Evaluate Policy Decisions

There are many questions that we might ask about the distribution network described above. For example, we could investigate how much inventory would be required to ensure reasonable patient waiting times if demands are wildly unpredictable and the RSS can only ship to a small number of PODs in each time period. Alternatively, we could explore the impact of POD service capacities on system performance, comparing dynamic staffing plans, in which service capacities correspond to patient demand or inventory levels, to constant staffing plans. We could also consider the impact of varying the numbers and sizes of PODs, the patient demand processes, or the POD processing capacities, as well as comparing different inventory policies or the effects of longer lead times. Since our purpose here is to illustrate the use of ESCOE rather than to provide a detailed analysis of the New York-Connecticut inhalational anthrax example, we will focus on just one area of interest, rather than all of these. Since, as mentioned earlier, the CDC is interested in creating a number of FDSs, we will explore whether these might be useful and to what degree.

FDSs have the greatest potential utility if the SNS can only send a small amount of inventory initially, or if PODs and RSSs are ready to begin operating before the initial shipment of inventory arrives from the SNS. For an anthrax scenario, like the one we are currently considering, the former possibility is unlikely since the SNS has run exercises and is confident that it will be able to provide a sufficiently large quantity of inventory to each state in the initial shipment [CDC, 2011]. Instead we will consider the latter option.

Setting up PODs and readying them to serve patients is nontrivial, since this involves securing the POD location, bringing in the necessary equipment, and

calling in volunteers and staff, to name just a few of the required tasks. Thus, PODs may require 12 or more hours to prepare for operations. However, well-organized PODs might be prepared to serve patients in fewer than 12 hours. We consider cases in which PODs open at the fourth hour after the emergency is declared, the eighth hour, and the twelfth hour.

We assume for these examples that patients do not begin arriving to wait for service until two hours before the PODs open. Since the main priority in responding to an inhalational anthrax attack is to distribute antibiotics to as many people as possible, it is reasonable to suppose that the clinics remain open continuously for the 48-hour time horizon. But we would not expect as many patients to seek medical care in the middle of the night as we would expect during the day, so patient arrival rates change throughout the day, rising higher during the main part of the day, decreasing slightly in the evening, and then falling to a lower level overnight.

Consequently, PODs should not be staffed to provide a constant rate of service 24 hours per day. In the model we assume for simplicity that during each eight-hour shift the expected rate of service remains constant, even though in a true emergency productivity would likely change, possibly increasing as staff become more familiar with their job within the POD and decreasing later in a shift due to exhaustion. For this example we supposed that the expected patient demand could be predicted with reasonable accuracy, and expected service capacities were set to be 25% greater than the expected demand in each time period.

The loading, unloading, and storage capacities at each location were assumed to be ample so that we could focus on inventory in this system, rather

than attempting to identify other bottlenecks. We also initialized the network with a very large amount of inventory at the SNS so that it will never require additional supplies during the dispensing campaign. When they were included in the simulation each FDS was provided with one pallet, or 9,600 units, of inventory. We increased the initial SNS stock by the amount of inventory stored in the FDSs when no FDSs were included in the simulation.

Thus, we have described how the remaining model parameters were set. The six different cases we study are network topologies with and without FDSs, and for each topology we consider the possibility of PODs opening at four, eight, and twelve hours after the emergency is declared. The total expected number of patient arrivals and service capacity will remain the same in every scenario. In the following section we describe the results of these six cases and the lessons that one might take away from these examples.

Example Results and Conclusions

We ran each of the simulations for ten replications. Figures 3.13 and 3.14 show the average queue lengths for sample small and large PODs under the four and twelve hour opening time scenarios. Unsurprisingly, when PODs are able to open in hour 4, the cases in which FDSs are present show much shorter patient queue lengths than the cases in which there are no FDSs. This is a somewhat unfair comparison, though, since it is not reasonable to open a POD before there is inventory available to stock it. The early case without FDSs is an example of what could happen if a POD staffed and ready to open with patients waiting before inventory arrives.

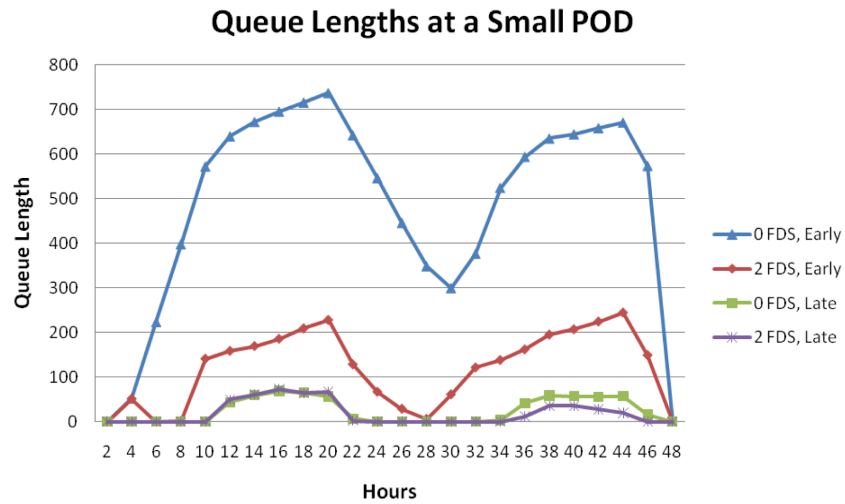


Figure 3.13: Average queue lengths at a small POD for the 0 and 2 FDS cases, with Early (4th hour) and Late (12th hour) POD opening times.

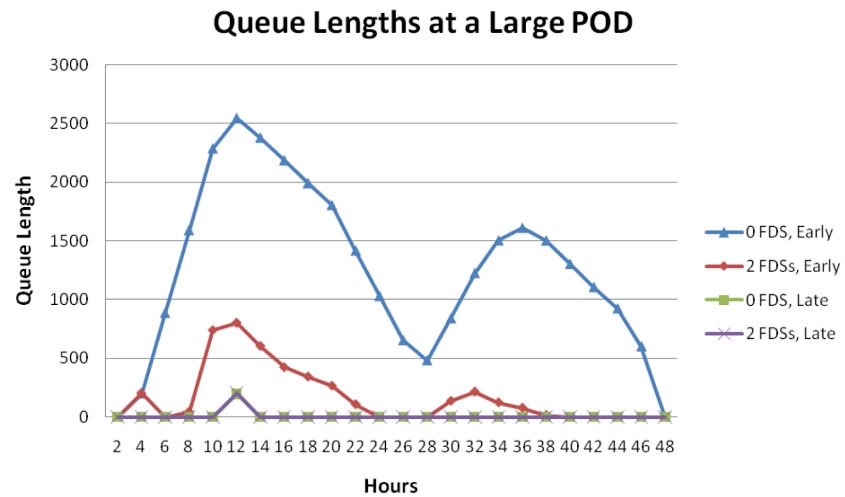


Figure 3.14: Average queue lengths at a large POD for the 0 and 2 FDS cases, with Early (4th hour) and Late (12th hour) POD opening times.

On the other hand, notice that in both Figures 3.13 and 3.14 the presence of FDSs makes little or no difference in the simulation output when PODs open

later. We conclude unsurprisingly that FDSs may be valuable if PODs are prepared to open and begin serving patients significantly before the 12th hour when inventory from the SNS arrives. However, we should not invest in FDSs unless assurance is given from local health departments that they are able to open functioning PODs significantly faster than 12 hours after an emergency is declared. This identifies a possible bottleneck in the distribution network, namely POD setup time, and the CDC might consider an exercise to evaluate realistic POD setup times before constructing FDSs. Policy makers might also wish to weigh the cost of stocking and maintaining FDSs against their potential benefits.

Figures 3.15 - 3.17 illustrate the importance of uncertainty in the system. Figures 3.15 and 3.16 show inventory levels and queue lengths at a single small POD that opens in the twelfth hour without FDS support under two different simulation replications. In the one replication, inventory at the POD drops to zero and a large queue builds up for several time periods near the end of the time horizon, while in the other replication the POD never runs out of inventory and queue lengths remain fairly small. Many public health emergency plans are designed to accommodate the expected patient demands, without acknowledging the uncertainty inherent in patient arrivals and other parts of the distribution network or how this uncertainty may affect system performance. Figures 3.15 and 3.16 provide a clear example of how this uncertainty can affect a system, causing troubles when things do not go as planned, as in simulation replication 7. Figure 3.17 shows ten replications for the same small POD, further emphasizing the potential variance of inventory levels over time due to fluctuations in patient demands and service capacities.

Drawing attention to this variation encourages policy makers to develop flexible response plans that will allow the system to respond to unusually large demands or lower-than-expected service capacities. Planners might consider increasing safety stocks at the RSS to help protect the network from inventory stockouts, or they could maintain a group of staff who move from POD to POD to increase service capacities in response to large queues that may develop. Other ideas could also be considered and modeled with ESCOE, but our goal is to help policy makers identify potential operational problems and to evaluate the impact of alternative operating strategies on response effectiveness.

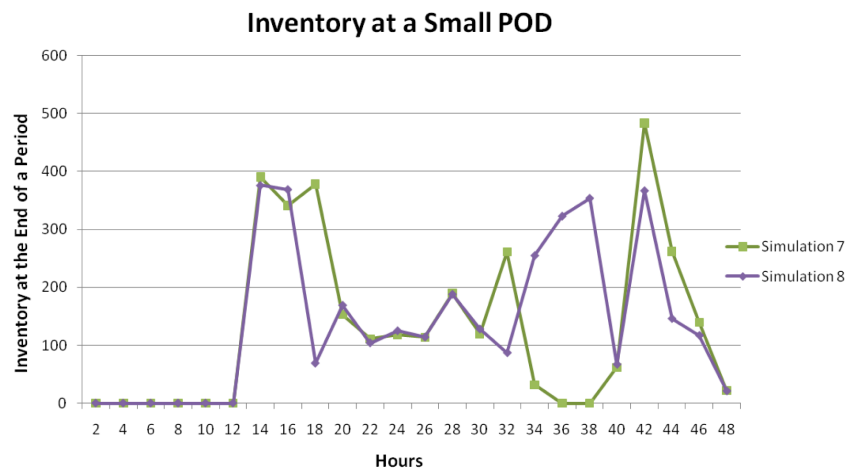


Figure 3.15: Inventory on-hand over time in the seventh and eighth simulation replications for one small POD that opens in the twelfth hour without FDS support.

3.2.3 Discussion

There are many possible mass dispensing network designs, and ESCOE allows for quantitative comparison between these different options. ESCOE also encourages users to consider the impact of uncertainty in the emergency response

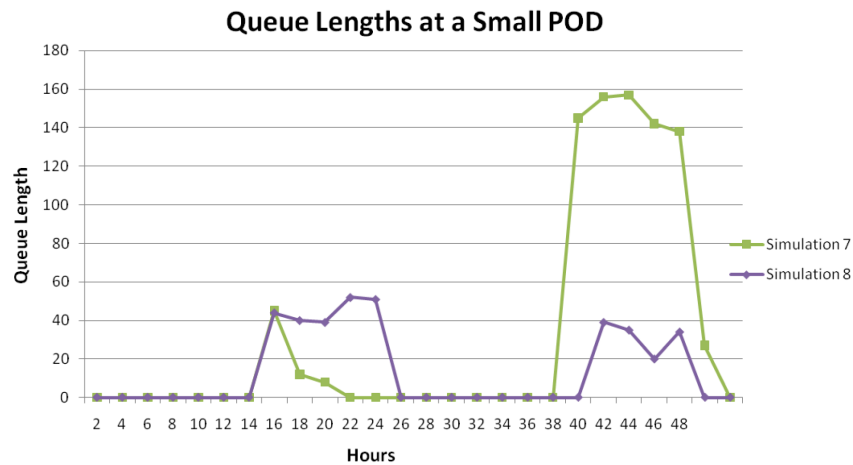


Figure 3.16: Queue lengths over time in the seventh and eighth simulation replications for one small POD that opens in the twelfth hour without FDS support.

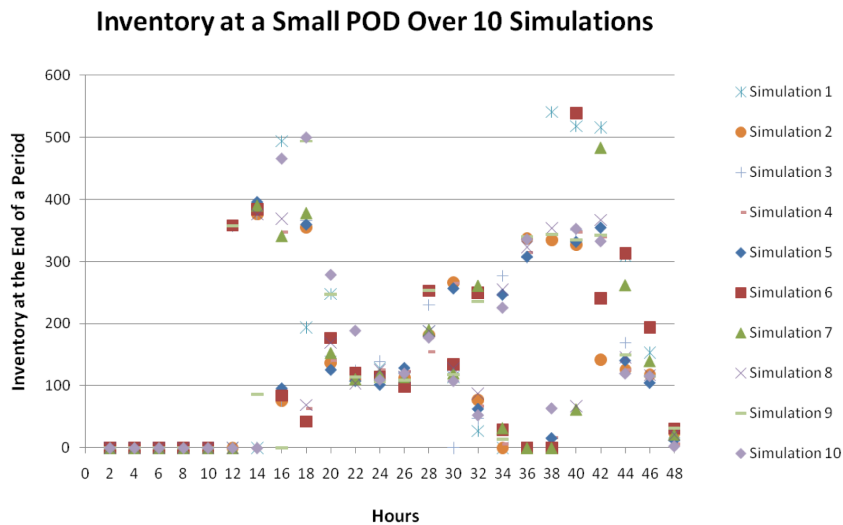


Figure 3.17: Inventory on-hand over time in the seventh and eighth simulation replications for one small POD that opens in the twelfth hour without FDS support.

supply chain. Furthermore, since only the standard Microsoft Office programs Excel and Access are required to run the simulations, ESCOE is easily accessible

to public health officials. Its modular interface could also be extended to allow users to describe particular distribution networks in greater detail. Some future additions to the simulation that are currently under consideration include modeling more complex inventory allocation strategies, multiple inventory types, and more detailed truck routing policies.

Simulations such as ESCOE and D-PODS are essential tools for public health policy-makers responsible for developing effective emergency preparedness plans. Realistic exercises involving many organizations across the supply chain are not feasible, so we must rely on simulation tools like these to understand how decisions made by different groups will affect the network as a whole. Such tools can help planners take a systems-oriented approach to planning emergency response networks that involve large numbers of independent agencies by allowing them to consider the entire scope of the planned response. D-PODS, and models like it, can help planners thoroughly understand particular elements of the system, while ESCOE helps planners consider how these pieces will fit together. These tools let policy makers better understand the consequences of other plans, allowing them to evaluate and improve our true emergency preparedness.

CHAPTER 4

ANTIVIRAL DISPENSING MODELS FOR AN INFLUENZA PANDEMIC

In Chapter 1, we described the need for a controlled-dispensing campaign to distribute antiviral medication in the early stages of an influenza pandemic. Antivirals can reduce the severity and infectiousness of individuals infected with influenza, and the United States' pandemic response plan calls for antivirals to be distributed to sick individuals. Currently, the plan required states to undertake the burden of dispensing antivirals, but compelling arguments have been presented in favor of using the commercial pharmaceutical supply chain to operate this campaign instead [Koonin *et al.*, 2011]. We are working with a team of CDC officials to construct models that will help policy-makers understand how a commercially operated dispensing campaign might work.

There are three large pharmaceutical distributors in the United States: McKesson, Cardinal Health, and Amerisource-Bergen. Each of these operates about 25 warehouses, which supply the 60,000 commercial pharmacies [SK&A, 2011]. The pharmacies generally place orders to and receive shipments from one or more of the distributors every day. In some cases computerized systems are already used to transfer information about inventory needs from many of the pharmacies to their distributors. In general, the distribution network operates very effectively, and the idea of leveraging this system to dispense antivirals is appealing. However, this raises many questions about how the dispensing network will be structured, whether pharmacies can provide sufficient dispensing capacity, and other policy concerns. With regard to the network structure, we must consider which pharmacies will be included in the plan. One possible answer would be to use all of them, but that could lead to undesirable outcomes.

Either massive amounts of inventory could be required to ensure that demand will be met with high probability in each pharmacy, or many pharmacies may run out of stock, causing patients to hunt for antivirals by traveling from pharmacy to pharmacy. Instead, the plan might select a subset of pharmacies, but how will this subset be chosen?

We also must ensure that the chosen pharmacies will have sufficient capacity to meet demand. How will the pharmacies handle a significant increase in demand due to antiviral prescriptions? If only a subset of pharmacies is chosen to participate in the campaign, then these locations may see a much larger than normal number of new patients, who require longer service times. Other questions arise with regard to policy. When will the SNS dispensing campaign begin? Will the free antivirals immediately be accessible to everyone, or will insurance companies pay for their customers' antivirals as long as the initial commercial supply is available? How will the SNS antivirals be processed in the pharmacies' inventory systems?

To identify some of the challenges that would arise, we must consider how influenza would spread across the country. We constructed hypothetical antiviral demand curves based on a variety of historical influenza data to help us better understand the consequences of influenza. The model separates antiviral demand by the HHS regions defined in Table 4.1. We discuss the construction of these curves in the following section, but first we explore the lessons that can be gleaned from them. Figure 4.1 shows the expected antiviral demands by region for a moderate pandemic similar in scale to the 2009 H1N1 pandemic. Notice

that the slopes of the demand curves are extremely steep and the peaks are very large. The demand in region 4 increases from 20,000 units to over 100,000 units in a single week. In most regions there is little or no “ramp up” period that would give pharmacies and distributors time to prepare. Furthermore, the timing of the peak varies significantly from region to region. Hence, the pharmacies in the distribution network must not only be capable of serving very high demands, but they must be prepared to increase their capacity rapidly to keep up with demand. Distributors must increase their on-hand inventories early in the pandemic so that they will be prepared to supply the pharmacies with the large quantities of antivirals that will be required.

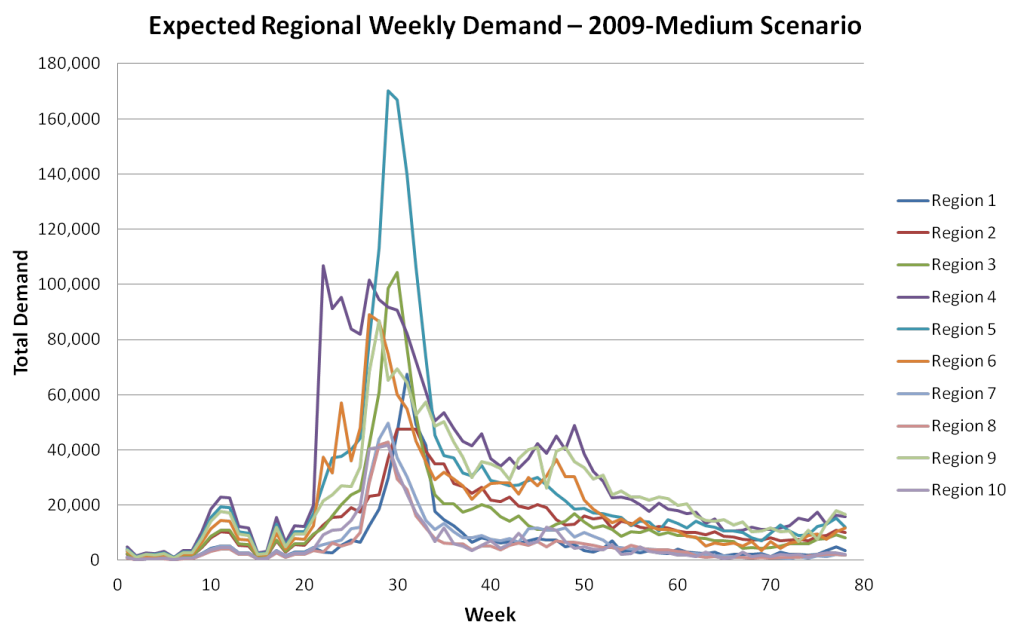


Figure 4.1: Expected patient demands for each of the ten HHS regions during a moderate 2009-like pandemic scenario.

However, because the timing of the pandemic varies by region, the distributors must also hold enough inventory in reserve to supply the regions in which demand peaks later. Since the distributors make daily shipments to the

pharmacies, we suggest that inventory should be sent to pharmacies in small, frequent shipments. This will help ensure that large stocks of antivirals will not remain unused when the demand drops off rapidly at the end of the peak period. If inventory imbalances do arise at the local level, information about the locations of available stock could be provided to the public so that patient demand can self-adjust. These imbalances could also be corrected in part by inventory-sharing between nearby pharmacies, but such a plan would require additional coordination and infrastructure and there is no guarantee that this would be possible.

The length of time during which the pandemic is active also varies by region. Demand for antivirals is elevated for only about ten weeks in regions 8 and 10, but in region 4 demand remains very highly elevated for 15 weeks and stays significantly above the normal level for almost 30 weeks. The unusual curve in region 4 is likely the result of a number of pandemic peaks arising in different parts of the region, which is large and includes eight states in the southeastern United States. However, we still see that the pharmacies must be capable of filling large numbers of antiviral prescriptions over an extended period of time while continuing to serve its regular customer base. Pharmacies must create staffing plans to accommodate these requirements. If a pharmacy's pandemic plan calls for all employees to work overtime to fulfill antiviral prescriptions, the pharmacy may not be capable of maintaining this pace for the complete pandemic.

We also observe from this demand model that inventory shortages are likely to occur in more severe pandemic scenarios. Figures 4.2 and 4.3 show the cumulative national antiviral demand and the SNS inventory levels during two

severe pandemic scenarios. Currently, the SNS has stockpiled 48 million courses of adult antivirals and 12 million courses of pediatric antivirals [Hupert, 2011]. These figures show that during a severe pandemic national shortages would certainly occur, so pharmaceutical distributors must allocate antivirals carefully to ensure that equitable service is provided across the United States.

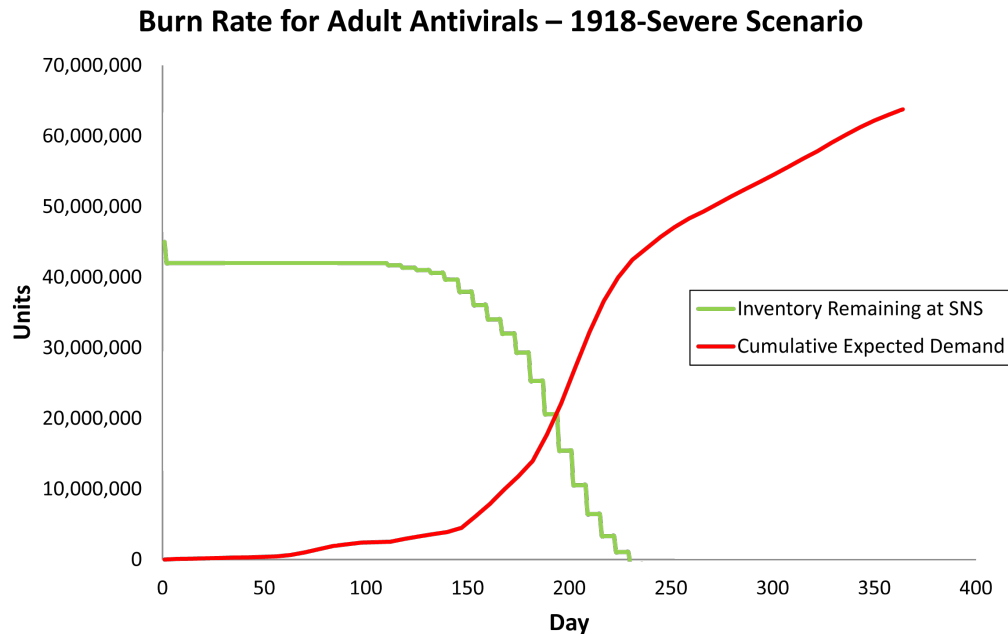


Figure 4.2: Burn rate for adult antivirals in a severe 1918-like scenario.

As we mentioned above, careful staffing would be essential to ensure that the pharmacy can fill the required antiviral prescriptions, but it is unlikely that other capacity constraints would present a challenge to the dispensing effort. One case of antivirals contains 300 bottles of pills, which is more than enough to supply daily demand at each pharmacy during all of but the peak times of the pandemic. Hence, neither storage capacity at the pharmacy nor space on the trucks that transport material daily from the distributor to the pharmacies will be a bottleneck in the distribution network. Each pallet contains 32 cases of

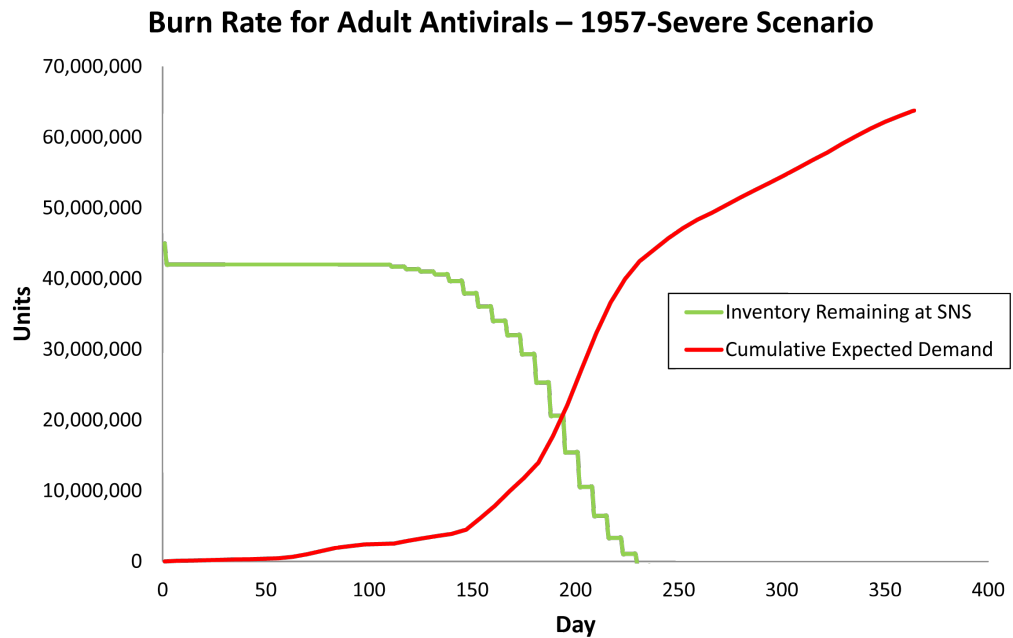


Figure 4.3: Burn rate for adult antivirals in a severe 1957-like scenario.

antivirals, or 9,600 bottles of pills. Even a very large initial shipment from the SNS to the distributors of 10 million courses of antivirals would require sending only about 14 pallets to each of the distribution warehouses, which would fill just over half of a standard 53 foot trailer. This is a trivial amount of inventory for these large distributors, so neither transportation nor storage capacities should be problematic at the SNS-distributor level, either.

We will keep these observations in mind throughout this chapter. In the following section we show how the antiviral demand curves shown in the previous three figures were constructed. In the subsequent section we present a simulation of the pharmaceutical distribution network and use these demands to demonstrate the potential performance of the network under different pandemic scenarios. Finally, we present several models to assist with inventory allocation during a pandemic.

Region 1	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
Region 2	New Jersey, New York
Region 3	Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, West Virginia
Region 4	Mississippi, Alabama, Tennessee, Kentucky, Florida, North Carolina, South Carolina, Georgia
Region 5	Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin
Region 6	Arkansas, Louisiana, New Mexico, Oklahoma, Texas
Region 7	Iowa, Kansas, Missouri, Nebraska
Region 8	Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming
Region 9	Arizona, California, Hawaii, Nevada
Region 10	Alaska, Idaho, Oregon, Washington

Table 4.1: HHS Regions.

4.1 Constructing Antiviral Demand Scenarios

Our goal in this section is to describe a method for estimating demand for antivirals across the United States during hypothetical pandemic scenarios. A number of papers have estimated epidemiological parameters, such as cumula-

tive attack rates and basic reproductive numbers, for the 2009-2010 H1N1 pandemic [Fraser *et al.*, 2009, Yang *et al.*, 2009, White *et al.*, 2009, Tuite *et al.*, 2010, Pre-sanis *et al.*, 2009, Carrat *et al.*, 2010, Nishiura *et al.*, 2010, White & Pagano, 2010]. However, Shrestha *et al.* is the definitive paper whose authors worked with CDC officials to estimate the true influenza attack rate in the United States for the pandemic [Shrestha *et al.*, 2011]. They used surveillance reports from laboratory-confirmed influenza cases and hospitalization data and applied statistical methods to correct for under-reporting of influenza.

Shrestha *et al.* provide data at the national level, but the timing of the peak and its magnitude of the 2009 pandemic varied by region. Since pharmacy capacity also varies significantly by region, more specific regional curves were necessary to determine whether that capacity would be sufficient for antiviral dispensing during a pandemic. The most detailed geographic information available for outpatient influenza cases in the United States was collected by the Influenza-Like Illness (ILI) Surveillance Program [CDC, 2009]. The ILI data include reports from a national network of over 3,000 health care providers who record the number of confirmed influenza and unconfirmed influenza-like illnesses that they observe each week. These data are compiled for each of the ten HHS regions, which are defined in Table 4.1.

We show below how the ILI data and the model by Shrestha *et al.* determine the shape of our regional antiviral demand curves. The magnitude of these curves depends on census reports of population size by age and geographic region and on a set of epidemiological parameters [USCB, 2012]. We considered twelve pandemic scenarios: low, medium, and severe versions of the four pandemics that have occurred in the last hundred years: 1918, 1957, 1968, and 2009.

Our epidemiological parameters were constructed by CDC epidemiologists for each of these pandemic scenarios [Koonin, 2012].

For each scenario, the parameters include AR_a , the clinical attack rate for each age group a ; t_a , the percentage of those ill in age group a that is both included in the treatment protocol and will seek antiviral treatment; and w_a , the percentage of the population in age group a that is not ill, but will seek antivirals. We sometimes refer to the latter group as the “worried well.” The age groups of interest are determined by epidemiological parameters and treatment protocols. Pediatric antivirals are required for individuals aged 0 to 11. Different attack rate curves are expected for people under the age of 18, between the ages of 18 and 64, and ages 65 and above [Koonin, 2012]. We define $\mathcal{A} = \{0-11, 12-17, 18-64, 65+\}$ to be the set of age ranges of interest. The geographic regions of interest are the ten HHS regions for which we have ILI data. Let $\mathcal{R} = \{1, \dots, 10\}$ be the numbers of the HHS regions.

Let POP_{ar} be the population of age group a in geographic region r , as given by the 2010 census data. Then, using the epidemiological parameters defined above, the total number of people in age group a in region r who require antivirals is given by

$$z_{ar} = (AR_a \cdot t_a + w_a)POP_{ar} \quad \text{for } a \in \mathcal{A}, r \in \mathcal{R}. \quad (4.1)$$

Next, we determine when this demand occurs. In some influenza pandemics, a small spring wave of illness occurs in March, April, and May and then mostly disappears before the main outbreak begins in August or September. Unfortunately, the ILI data from 2009 were only collected starting in week

35 (August 24-30) of 2009, which fails to capture this initial curve. To estimate a reasonable demand curve for a spring wave, we turn to the model by Shrestha et al. [Shrestha *et al.*, 2011]. Shrestha's model begins showing influenza cases in week 14 (April 2-8).

Define S_t to be the number of cases occurring in week t according to the Shrestha model [Shrestha *et al.*, 2011]. Define S to be the total number of cases occurring over the full pandemic in the model, so $S = \sum_t S_t$. Define S_F to be the total number of cases that occur in the first wave (weeks 14 through 34) in the model, so $S_F = \sum_{t=14}^{34} S_t$. The percentage of demand that occurs during the first wave is given by $\frac{S_F}{S}$, so for our demand model, the total number of cases for age group a in region r that would occur in the first wave would be $z_{ar} \cdot \frac{S_F}{S}$. However, we wish to allow additional flexibility to explore the impact of spring waves of varying sizes, so we define the user-determined parameter $p \in [0, 1]$ to be the percentage of the total cases that will occur during the spring wave, and the number of cases in the first wave is $z_{ar} \cdot p$. Within the first wave, the fraction of first wave cases that occur during week t is given by $\frac{S_t}{S_F}$. So the expected demand for age group a in region r during week t is given by

$$z_{art} = z_{ar} \cdot p \cdot \frac{S_t}{S_F} \quad \text{for } a \in \mathcal{A}, r \in \mathcal{R}, t = 14, \dots, 34. \quad (4.2)$$

The total number of cases for age group a in region r that occur during the second wave of our model is $z_{ar}(1 - p)$. To determine the fraction of these cases that occurs during week t , we apply the ILI data. These data from 2009 were provided for five age categories: 0-4, 5-24, 25-49, 50-64, and 65+. To determine an antiviral dispensing policy, we need to recalculate the disease data for the age groups in \mathcal{A} . This requires separating the age 5-24 group into three separate

categories: 5-11, 12-17, and 18-24.

Let ILI_{art} be the number of ILI cases recorded for age group a in region r during week t . We define the ILI case data for the age categories $a \in A = \{0 - 11, 12 - 17, 18 - 64, 65+\}$ to be the transformed values ILI'_{art} as follows:

$$\begin{aligned} ILI'_{0-11,r,t} &= ILI_{0-4,r,t} + \frac{POP_{5-11,r}}{POP_{5-24,r}} \cdot ILI_{5-24,r,t}, \\ ILI'_{12-17,r,t} &= \frac{POP_{12-17,r}}{POP_{5-24,r}} \cdot ILI_{5-24,r,t}, \\ ILI'_{18-64,r,t} &= ILI_{25-49,r,t} + ILI_{50-64,r,t} + \frac{POP_{18-24,r}}{POP_{5-24,r}} \cdot ILI_{5-24,r,t}, \text{ and} \\ ILI'_{65+,r,t} &= ILI_{65+,r,t}. \end{aligned}$$

This calculation assumes that, during the 2009-2010 influenza season, the attack rate for H1N1 was uniform across the 5-24 age category. This seems to contradict one of our reasons for further dividing that category, namely that the attack rate will vary by age group. However, lacking more finely recorded data, we cannot justify using any other method of estimating the demands.

In 2009, the total number of ILI cases for group a in region r is $ILI'_{ar} = \sum_t ILI'_{art}$. The fraction of cases for week t is given by $\frac{ILI'_{art}}{ILI'_{ar}}$. Since all of the ILI cases occurred during the second wave of the 2009 H1N1 pandemic, this number is the fraction of the second wave cases that occur in week t . So the expected number of people of age group a in region r who require antivirals in week t is

$$z_{art} = z_{ar} \cdot (1 - p) \cdot \frac{ILI'_{art}}{ILI'_{ar}} \quad \text{for } a \in \mathcal{A}, r \in \mathcal{R}, t = 35, \dots, T, \quad (4.3)$$

where $T = 91$ is the last week of data available from the ILI records. Equations

(4.2) and (4.3) together define the weekly demand by age group and region for a complete pandemic scenario.

4.2 Simulating System Performance

The demand models presented in the previous section provided some insights into the requirements of the distribution network, but they do not capture the fundamental uncertainty present during any influenza pandemic. We constructed a simulation of the system to help policy-makers identify potential pitfalls of using the pharmaceutical supply chain and to understand more completely how this uncertainty could affect the operations of the response network. We developed the simulation in collaboration with a group of CDC officials who provided information about current and future SNS operations to set the parameter values used in our examples below.

We built the simulation in Visual Basic and Microsoft Access, with a Microsoft Excel interface to make it accessible for our audience of policy-makers. There are three areas in which user input is required: the demand model, the distribution network structure, and the rules for dispensing stocks from the SNS inventory. For the former, the interface allows users to select one of the twelve pandemic scenarios described in the previous section to simulate. Users may further modify these scenarios by changing the epidemiological parameters or by directly increasing or decreasing expected demands for particular regions in particular weeks. The user also selects one state, for which supply and demand will be simulated at the pharmacy level; in all other states, only cumulative supply and demand values will be determined.

We allow for several pharmacy types in the model. In the examples described below, we will assume three such types: Large, Medium, and Small. We populate the simulation with the actual numbers of pharmacies in each state; in our example, we assume that all chain pharmacies are type Large, mass merchant and supermarket pharmacies are type Medium, and independent pharmacies are type Small. The user may select the percentages of pharmacies of each type that are included for each state. Pharmacies have service capacities which indicate the maximum number of antiviral prescriptions that they could fill each day. These capacities are assumed to be constant within a region and are assigned by the user. The proportions of demand assigned to pharmacies of each type are also assumed to be constant over time within each region. In the examples, we assume that twice as many patients seek service at medium pharmacies as at small pharmacies, and twice as many seek service at large pharmacies compared to medium ones.

Some information about distribution warehouse locations is publicly available, but there is information describing which distribution warehouses serve particular pharmacies is proprietary. For simplicity, we assume that there is one distribution warehouse in each state which serves all of the pharmacies in that state. We do not model any constraints on service or transportation capacities at the distributor warehouses since, as we mentioned earlier, the quantities of antivirals under consideration would not strain distributor capacities. There are shipping lead times between the SNS and the distributors and between the distributors and the pharmacies. In the example, the distributor-to-pharmacy lead times are one day, since most pharmacies already receive daily shipments from their distributors. The SNS-to-distributor lead times are two days.

Recall that the SNS presently has on-hand 48 million courses of adult antivirals and 12 million courses of pediatric antivirals. The SNS plans to begin its inventory distribution effort with an initial push of inventory to the distributors and pharmacies. The size of the initial push is determined by the user; the default values are 8 million courses of adult antivirals and 3 million courses of pediatric antivirals, for a total 11 million, which was the size of the initial inventory push in the 2009 H1N1 pandemic [HHS, 2009]. After the initial push, the SNS will send new shipments periodically in response to orders from the distributors; the default frequency of shipments is once each week.

Some portion of the SNS inventory would be sent directly to state public health authorities to serve special populations. Since our goal in this simulation is to evaluate the ability of the commercial pharmaceutical supply chain to dispense most of the antivirals, we simply assume that each state would serve some percentage of its population. We allow the user to set these percentages and assign some portion of the SNS stockpile to be reserved for the state public health authorities. In our examples, both of these are set to five percent.

We now describe the simulation environment. We model time in periods; each period is one day long. Each day, inventory shipped from the SNS one lead time ago is received at all of the distribution warehouses. As we mentioned above, we simulated operations at the pharmacy level only for a single, user-selected state. For the other states, we calculate only cumulative performance measures. For the chosen state, inventory shipped from its distribution warehouse to the pharmacies one lead time ago is also received. Then each of these locations places an order on its supplier. Each location forecasts its demand over the lead time and sets its order-up-to level to this quantity plus safety

stock equal to twice the square root of this demand, which is two standard deviations if the demand is Poisson distributed. Each location uses a simple look-back forecasting method; that is, it assumes that daily demand in the future will equal the previous day's demand. This type of simple forecasting is a reasonable reflection of what these systems might actually implement, but extensions to the model could include using a triple exponential smoothing method, which would allow us to capture long-term and local trends in the demands.

After orders have been placed, if the current day is one on which the SNS makes a shipment, the SNS makes an allocation to the distributors. If sufficient inventory is available at the SNS, all distributors will receive their desired order quantities. Otherwise, each distributor will receive an equal fraction of its order quantity. This process is repeated at the distributor in the chosen state as it allocates inventory to the state's pharmacies.

Next, we draw random patient demands. Demand at the regional level is assumed to be Poisson distributed. In the previous section we showed how the expected weekly demands are determined for each region. We determine the initial daily expected demand for each region by linearizing the expected weekly demands over the days of the week so that there are no "jumps" in demand between weeks. To ensure that each randomly drawn set of patient demands results in approximately the same total number of patients served, we add one modification to these expected demands. The expected daily demand for a particular region on a given day is the sum of this initial demand and additional factor. The extra factor is the previous day's expected demand minus that day's actual sampled demand. This ensures that, if fewer people than expected sought service on the previous day, then the current day's demand will

compensate by being larger. The demand for a chosen region on a given day is drawn from a Poisson distribution whose mean is constructed as described here. Each region's demand is randomly assigned to states within the region in proportion to the states' populations. For the chosen state, the state's demand is further allocated to pharmacies in that state, in proportion to the demand ratios mentioned above.

Once the day's demand's are known, we can calculate the number of people served and unserved at each pharmacy in the chosen state, as well as the remaining inventory. Unserved patients do one of two things, depending on the user-selected policy: either they return to the same pharmacy the following day for antivirals, or they stop seeking to fill their prescriptions. The latter choice is not unreasonable, since antivirals are unlikely to be effective in reducing the severity of illness unless treatment begins immediately. Another reasonable patient behavior would be to seek treatment at nearby pharmacies on the same day. The CDC has recently provided us with county-level pharmacy data, so this extension to the model is currently underway.

For the states whose pharmacies are not simulated, we assume that the state dispensing capacity is equal to the sum of all of its pharmacies' capacities. The number of patients served in each of these states is the minimum of the state dispensing capacity, the total patient demand in the state, and the inventory on-hand at the state's distributor. The same rules for unserved patients apply, as described above.

4.2.1 Example Results

In this section we present a simple example and results to explore the simulation output. We considered the four moderate pandemic scenarios (1918, 1957, 1968, 2009), simulating pharmacies for both New Jersey and Georgia, which were chosen because they are typical states from two regions with noticeably different expected demand curves. We included half of the pharmacies in each state in the dispensing network and set antiviral dispensing capacities of 200, 100, and 50 for the Large, Medium, and Small pharmacies, respectively. In each case we simulated the pandemic for one year. We also explored the consequences of patient behavior. We compared system performance when patients continue returning until they receive service (the “Patients Return” policy) and when patients abandon their search for antivirals if their chosen pharmacy cannot provide service (the “Patients Leave” policy).

First, we consider the efficiency of the overall network. Figure 4.4 shows the number of patients served under each pandemic scenario for the two possible patient behavior patterns. We see that the different pandemic scenarios have very different inventory requirements; the 1918 and 1957 scenarios are significantly more severe than the 1968 and 2009 scenarios. Also, many more patients are served under the Patients Return policy. This seems to indicate that such a policy is preferable, since the goal of the dispensing campaign is to provide antivirals to as many people as possible. However, antivirals are most effective in reducing the severity of illness if they are taken soon after an individual is infected with influenza. If individuals spend multiple days seeking service, the window of effectiveness will be lost. If patients do not return for service, far less inventory is required and the patients who are served will be individuals

for whom the antivirals are most useful. While inventory will not be a limiting factor in a moderate 2009 or 1968 pandemic scenario, almost the complete stockpile is consumed in the moderate 1918 scenario. If the population could be served equally effectively with far less inventory, that would be very highly desirable. Of course, during an actual pandemic public health officials cannot completely control patient behavior; they may use mass media to educate the public and encourage them to behave in certain ways (e.g., asking people to refrain from filling antiviral prescriptions that are more than one day old), but such requirements would probably not be enforced. Panicked individuals and the worried well may continue to seek treatment, even if it is unlikely to be effective. The actual performance of the system during a pandemic is likely to fall somewhere between the two extreme policies (Patients Return and Patients Leave) discussed here, so we may think of these examples as upper and lower bounds on inventory requirements and service rates.

A natural question that arises from this discussion is whether service capacity limitations, inventory shortages, or both cause patients to remain unserved under these policies. In Table 4.2 we show the first days on which patient demands for adult antivirals exceed available inventories under the Patients Leave policy. Demand never exceeds capacity in any of these scenarios, but stockouts occur less than halfway through the year. Table 4.3 shows the first day on which pharmacy capacity is insufficient for the total patient demand under the Patients Return rule. The Patients Return scenarios result in much higher effective demands, since each day returning patients must be served in addition to newly arriving patients. However, even under this demand-intensive policy, the effective demand never exceeds service capacity under the 2009 or 1968 pandemic scenarios. These tables indicate that inventory shortages are the main prob-

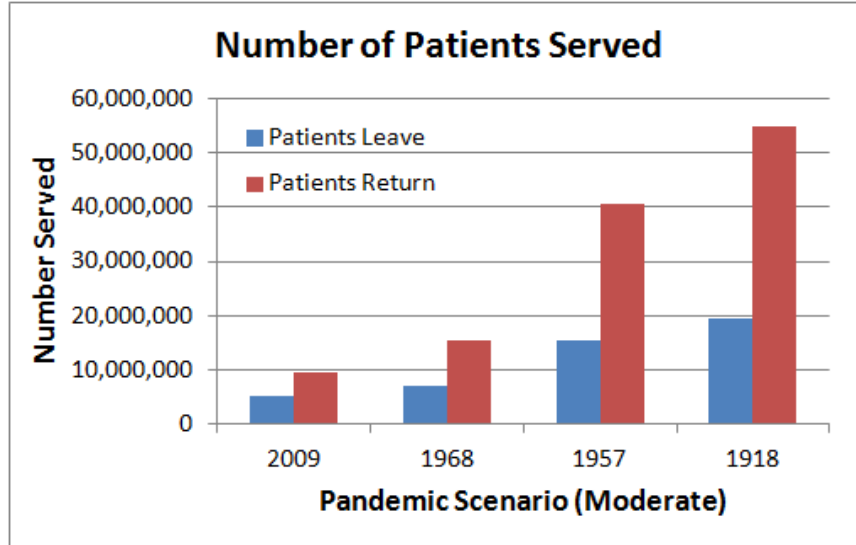


Figure 4.4: Number of patients served in the distribution network under two possible patient behavior strategies: patients leave the system stop seeking antivirals if inventory is unavailable or patients return until they are served.

lem in the system. A better inventory allocation policy or a more sophisticated forecasting method could significantly improve system performance under either patient behavior pattern. Recall that the current forecasting policy predicts future demand based on the previous day's demand. Even with a reasonable quantity of safety stock, this forecasting method is insufficient when demand increases rapidly as the pandemic proceeds. Inventory constantly lags behind demand, so large numbers of backorders accumulate under the Patients Return policy and large numbers of patients remain unserved under the Patients Leave policy.

We can also explore patient demand at the pharmacies more directly. Figure 4.5 shows the average and maximum daily demands for New Jersey and Georgia for each pandemic scenario when patients do not return to keep seeking service. We make two observations. First, the average daily demand is very

Pandemic	Large		Medium		Small	
Scenario	NJ	GA	NJ	GA	NJ	GA
2009	154	148	149	149	168	150
1968	98	99	119	120	115	85
1957	69	69	66	66	66	66
1918	61	65	60	61	59	59

Table 4.2: First day on which a **stockout** of adult antivirals occurs at pharmacies in New Jersey (NJ) and Georgia (GA) in moderate pandemic scenarios under the Patients Leave policy.

Pandemic	Large		Medium		Small	
Scenario	NJ	GA	NJ	GA	NJ	GA
2009	Never	Never	Never	Never	Never	Never
1968	Never	Never	Never	Never	Never	Never
1957	206	156	206	156	205	155
1918	205	155	205	155	204	154

Table 4.3: First day on which demand **exceeds service capacity** at pharmacies in New Jersey (NJ) and Georgia (GA) in moderate pandemic scenarios under the Patients Return policy.

low; for large portions of the pandemic the pharmacies will not experience any strain on their resources, even under the more severe 1918 and 1957 pandemic scenarios. Even the maximum demands never exceed half of the pharmacy capacities defined for this example. Given a better inventory allocation scheme, the pharmacies in New Jersey and Georgia could easily serve all of the demand that arises during any of the four pandemic scenarios shown.

However, in Figure 4.6 we see that the system may be overwhelmed when patients return repeatedly seeking service. The average daily demand is low, but the maximum effective patient demand (which includes returning and newly arrived patients) is more than double the available service capacity at all pharmacy types under the moderate 1957 and 1918 pandemic scenarios. This

highlights the need for improved inventory management; we know from Figure 4.5 that the maximum number of patients arriving on each day is always reasonable. If most patients were always served immediately, large queues such as the ones we observe in the bottom graph of Figure 4.6 would never have the opportunity to develop. Unfortunately, perfect inventory management will be impossible during a pandemic due to highly unpredictable demands, so pharmacies must construct staffing plans that will allow them to respond to higher than expected demands.

Notice in Figure 4.6 that, for the 2009 and 1968 pandemic scenarios, the maximum effective demand remains reasonable even when patients return until they are served. Thus, this example indicates that for mild pandemic scenarios, the commercial pharmaceutical network would be very effective in operating a controlled dispensing campaign, even with extremely simple inventory policies such as the ones currently implemented in this simulation. For more severe pandemic scenarios, the commercial system has sufficient capacity to serve the necessary demands, but inventory and other limited resources must be allocated carefully to ensure that sick people throughout the country will have access to antivirals.

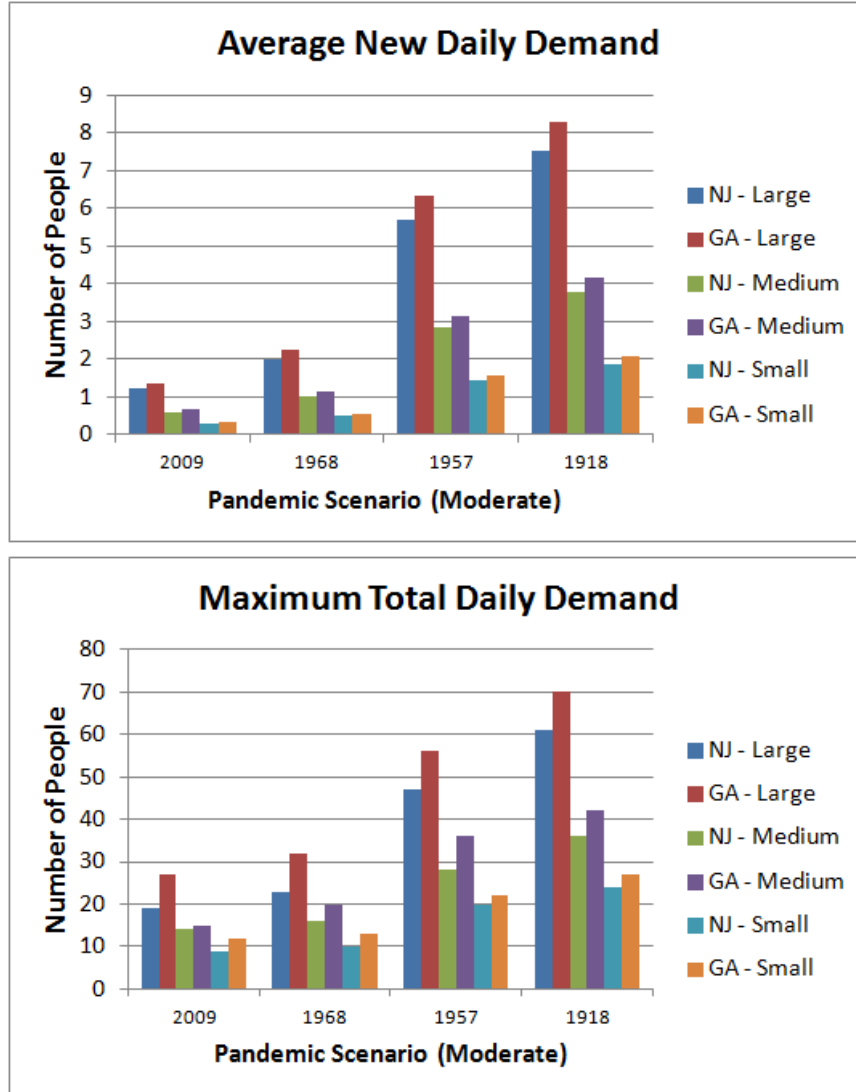


Figure 4.5: The top graph shows the average demand for antivirals each day and the bottom shows the maximum demand on a single day, under the Patients Leave policy.

4.3 Antiviral Allocation Models

In this section, we present some more sophisticated inventory allocation models that could be used to improve the performance of an antiviral dispensing campaign operated by the commercial pharmaceutical supply chain, discussed

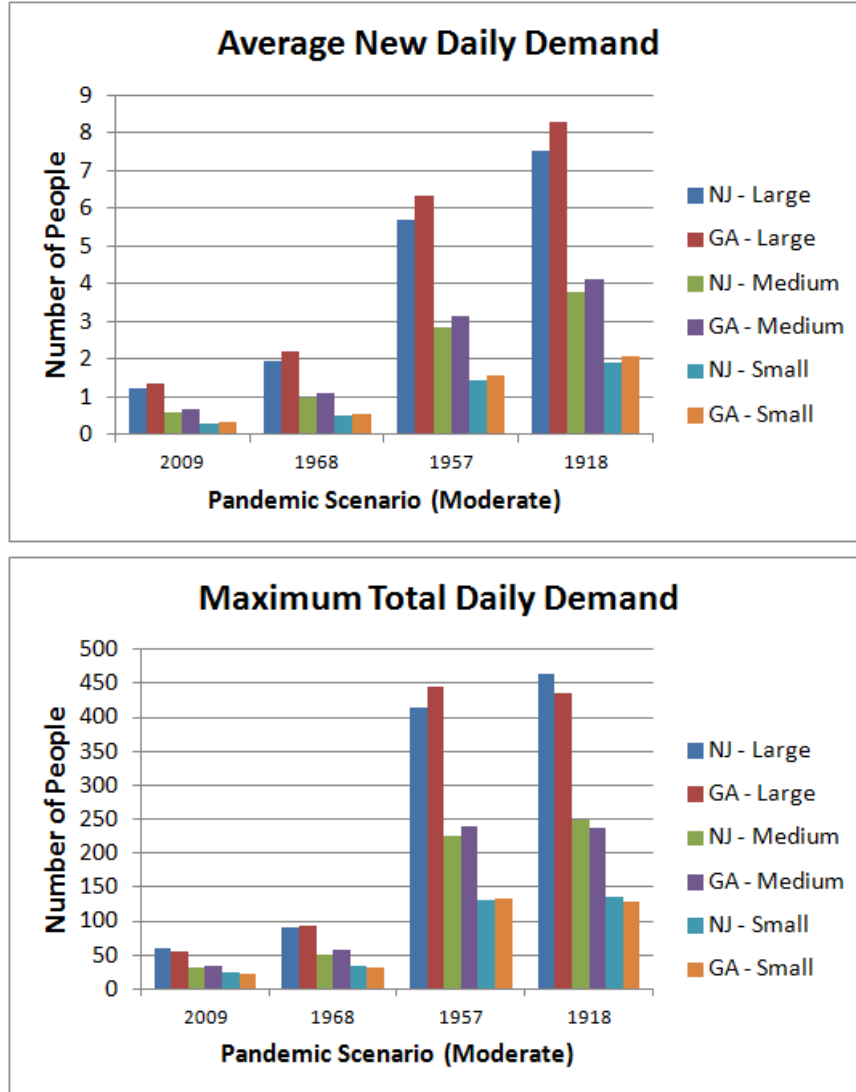


Figure 4.6: The top graph shows the average newly arriving demand for antivirals each day and the bottom shows the maximum actual demand (new arrivals + returning patients) on a single day, under the Patients Return policy.

in the previous section. The inventory allocation methods implemented in the simulation are reasonable representations of how the distribution network may operate in practice, but we showed in Chapter 2 that centralized allocation methods can significantly improve performance. In the next two subsections we present several inventory models for making allocations from the SNS to

the distributors and from the distributors to the pharmacies.

4.3.1 SNS-Distributor Inventory Allocation

In the simulated distribution network we assumed that one distribution warehouse served each state and that cross-shipping was allowed between all of the warehouses. In fact, there are three large pharmaceutical distributors in the United States, each of which maintains its own distribution warehouses. All three would likely be involved in the distribution effort, and while they may cross-ship inventory between their own warehouses, it is unlikely that they would share with one another. The SNS may ship directly to each of the distribution warehouses, or the SNS may ship only to a few select locations, allowing the distributors to determine how this inventory should be shared among their warehouses. Below, we present models for both of these possibilities.

We define \mathcal{B} to be the set of the three large distribution companies. Each distributor must own at least one warehouse from which to ship inventory to the pharmacies; let \mathcal{W} be the set of all of these warehouses. For $b \in \mathcal{B}$, let $\mathcal{W}(b)$ be the set of warehouses that belong to distributor b . As we discussed above, the SNS may ship antivirals to some or all of these warehouses; let \mathcal{W}^* be the set of warehouses to receive direct shipments from the SNS. For each $b \in \mathcal{B}$, there must exist at least one $w \in \mathcal{W}(b)$ such that $w \in \mathcal{W}^*$. Let \mathcal{H} be the set of pharmacies included in the network, and let $\mathcal{H}(w)$ be the set of pharmacies served by warehouse $w \in \mathcal{W}$. We assume that each pharmacy is served by exactly one warehouse; that is, for any two warehouses $w, v \in \mathcal{W}$, $\mathcal{H}(w) \cap \mathcal{H}(v) = \emptyset$. Let \mathcal{R} be the set of geographic regions of interest; these may represent HHS

regions or states or some other areas. Without loss of generality, we will also assume that each warehouse serves only pharmacies in a single region. We can simulate a warehouse that serves multiple regions by colocating several warehouses in our model with 0 lead times for moving inventory between them. Let $R(w)$ be the regional location of warehouse $w \in \mathcal{W}$; we know that $R(w) \in \mathcal{R}$.

The SNS sends out shipments periodically, once every L days, where L is likely to be 7 as in the simulation example discussed in the previous section. We will refer to each period of L days between shipments as a “cycle.” There is a lead time, τ_w , from the SNS to each warehouse $w \in \mathcal{W}^*$ which includes the time for picking, packing, and receiving inventory, as well as the transportation time. We model two possible allocation mechanisms: either the SNS ships directly to all of the warehouses, so $\mathcal{W}^* = \mathcal{W}$, or the SNS ships only to a subset of warehouses, so $\mathcal{W}^* \subset \mathcal{W}$. In the second case, the warehouses that receive shipments from the SNS subsequently cross-ship inventory to other warehouses owned by the same distributor. Let τ_{vw} be the time required to send inventory from warehouse v to warehouse w .

We model two inventory types, pediatric (type 1) and adult (type 2). Let $\mathcal{I} = \{1, 2\}$ be the set of inventory types. Let $D_{w,t,t+k}^i$ be the cumulative demand for inventory of type $i \in \mathcal{I}$ at warehouse w for days $t, \dots, t+k$. We assume that the distribution of $D_{w,t,t+k}^i$ is known only for the near future. Let \bar{x}_{wt}^i be the echelon inventory position for inventory type i at warehouse w at the beginning of period t . Let \bar{y}_{wt}^i be the echelon inventory position after inventory allocation decisions have been made and shipments sent in period t .

B	the set of all distributors
\mathcal{W}	the set of all distributor warehouses

$\mathcal{W}(b)$	the set of all distributor warehouses that belong to distributor $b \in \mathcal{B}$
\mathcal{W}^*	the set of all distributor warehouses to receive shipments from the SNS
\mathcal{H}	the set of all pharmacies in the network
$\mathcal{H}(w)$	the set of all pharmacies to receive shipments from warehouse $w \in \mathcal{W}$
$R(w)$	the location of warehouse $w \in \mathcal{W}$, where $R(w) \in \mathcal{R}$
L	the length of a cycle (i.e., the number of days between shipments from the SNS)
τ_w	the lead time for a shipment from the SNS to warehouse $w \in \mathcal{W}$
τ_{vw}	the lead time for a shipment from warehouse v to warehouse w , where $w, v \in \mathcal{W}(b)$ for some $b \in \mathcal{B}$

Table 4.4: Influenza model notation.

Let us now consider the case in which the SNS ships directly to every warehouse. There are several constraints that may affect inventory allocation from the SNS. First, the SNS may choose to hold back some inventory in each period t . Let z_{0t}^i represent the minimum quantity that the SNS is required to retain on day t . In 2009, the SNS initially shipped out 8 million courses of adult antivirals, but no further shipments were sent for some months. Since the initial quantity of adult antivirals is currently 48 million, we would have $z_{0t}^2 = 40,000,000$ for

the months following the initial shipments if the same decision was made in this model. So, on day t , we have the allocation constraint

$$\sum_{w \in \mathcal{W}} \bar{y}_{wt}^i \leq \bar{x}_{0t}^i - z_{0t}^i \quad \text{for all } i \in \mathcal{I}. \quad (4.4)$$

Furthermore, the SNS may choose to limit the amount of inventory available to each region, to ensure that it does not receive more than its “fair share” of antivirals. This type of “fairness” constraint has long been a centerpiece of SNS decision-making. Let p_{r1}^i be the initial inventory of type i available to warehouses in region r , and p_{rt}^i to be the remaining inventory of type i available to be allocated to region r on day $t > 1$. Thus, on day t we have the constraint

$$\sum_{w: R(w)=r} (\bar{y}_{wt}^i - \bar{x}_{wt}^i) \leq p_{rt}^i \quad \text{for all } r \in \mathcal{R} \text{ and } i \in \mathcal{I}. \quad (4.5)$$

In Chapter 2 we did not assume that imbalance was unlikely since there were a large number of PODs serving very unpredictable demand patterns over a short time horizon. However, we now model a relatively small number of warehouses serving a large number of pharmacies, so demand will be “smoothed” at the warehouse level. While it is difficult to estimate exact demand levels with accuracy far into the future, we may be confident that significant imbalance will not arise because each the SNS will gradually ship out inventory over time, so each warehouse will never have excessive stock of inventory. Hence, we make the balance assumption, which means that we do not need to add the constraint $\bar{y}_{nt}^i \geq \bar{x}_{nt}^i$.

As we mentioned above, we assume that the demand distributions are only known for the near future, so we make decisions to minimize cost only over the

short term. Since our goal is to limit the inventory remaining unused at the end of a period as well as the patients unserved, we charge per-unit holding costs of h_{wt}^i and backorder costs of b_{wt}^i for inventory of type i at warehouse w in period t . Thus, our cost function in period $t + k$ is

$$C_{w,t+k}^i(y) = E\left[h_{w,t+k}^i(y - D_{w,t,t+k}^i)^+ + b_{w,t+k}^i(D_{w,t,t+k}^i - y)^+\right]. \quad (4.6)$$

In period t , our goal is to minimize the costs that will be incurred as a result of this decision. A shipment sent in period t arrives at warehouse w in period $t + \tau_w$. The next shipment is sent in period $t + L$ and arrives in period $t + \tau_w + L$. Hence, we wish to minimize the inventory that remains unused and the patients who are unserved in period $t + \tau_w + L - 1$. Thus, our model can be written

$$\begin{aligned} G_t(\bar{\mathbf{x}}_t, \mathbf{p}_t) &= \min \sum_{w \in \mathcal{W}, i \in \mathcal{I}} C_{w,t+\tau_w+L-1}^i(\bar{y}_{wt}^i) \\ \text{such that} \quad &\sum_{w \in \mathcal{W}} \bar{y}_{wt}^i \leq \bar{x}_{0t}^i - z_{0t}^i \quad \text{for all } i \in \mathcal{I} \\ &\sum_{w: R(w)=r} (\bar{y}_{wt}^i - \bar{x}_{wt}^i) \leq p_{rt}^i \quad \text{for all } r \in \mathcal{R} \text{ and } i \in \mathcal{I}. \end{aligned} \quad (4.7)$$

To find an optimal solution to model (4.7), we first calculate

$$\bar{y}_{wt}^{i*} = F_{w,t+\tau_w+L}^{-1} \left(\frac{b_{w,t+\tau_w+L}}{b_{w,t+\tau_w+L} + h_{w,t+\tau_w+L}} \right),$$

for each warehouse $w \in \mathcal{W}$. If $\sum_{w \in \mathcal{W}} \bar{y}_{wt}^{i*} \leq \bar{x}_{0t}^i - z_{0t}^i$ and $\sum_{w: R(w)=r} (\bar{y}_{wt}^{i*} - \bar{x}_{wt}^i) \leq p_{rt}^i$ for all $r \in \mathcal{R}$ and $i \in \mathcal{I}$, then the optimal solution is $\bar{y}_{wt}^i = \bar{y}_{wt}^{i*}$. Otherwise, we must use marginal analysis for each region to find values of \bar{y}_{wt}^i that satisfy

the regional constraint. If $\sum_{r \in \mathcal{R}} p_{rt}^i \leq \bar{x}_{0t}^i - z_{0t}^i$, then this solution is optimal. If $\sum_{r \in \mathcal{R}} p_{rt}^i \geq \bar{x}_{0t}^i - z_{0t}^i$, then we must perform an additional marginal analysis step to reduce some of the inventories further to satisfy this constraint.

We now transfer our focus to the case where the SNS ships only to a subset of the warehouses, $w \in \mathcal{W}^*$. In this setting, there is a set of allocation decisions made at the SNS and then cross-shipment decisions are made at the warehouse after one lead time. We assume now that the lead time from the SNS to any warehouse in \mathcal{W}^* is given by τ , where $\tau < L$. Let the echelon inventory positions of the warehouses after a shipment is sent from the SNS in period t be \hat{x}_{wt}^i ; these shipments are constrained by inventory availability and shipping constraints from the SNS. The shipments arrive in period $t + \tau$ to all warehouses $w \in \mathcal{W}^*$. These warehouses immediately send cross-shipments to other warehouses that belong to the same distributor. The demand from periods $t, \dots, t + \tau - 1$ has already been observed when these decisions are made. Our goal, as before, is to minimize the inventory unused at the end of the cycle and the number of patients unserved up until the next shipment arrives from the SNS in period $t + \tau + L$, so we minimize costs incurred in time $t + \tau + L - 1$. We write the model as

$$\begin{aligned}
G_t^2(\bar{\mathbf{x}}_t, \mathbf{p}_t) &= \min \sum_{b \in \mathcal{B}} E_{t,t+\tau} \left[H_{b,t+\tau} \left(\sum_{w \in \mathcal{W}(b)} [\hat{x}_{wt}^i - \mathbf{D}_{\mathbf{w},t,t+\tau-1}] \right) \right] \\
\text{such that} \quad &\sum_{w \in \mathcal{W}} \hat{x}_{wt}^i \leq \bar{x}_{0t}^i - z_{0t}^i \quad \text{for all } i \in \mathcal{I} \\
&\sum_{w: R(w)=r} (\hat{x}_{wt}^i - \bar{x}_{wt}^i) \leq p_{rt}^i \quad \text{for all } r \in \mathcal{R} \text{ and } i \in \mathcal{I}. \\
&\hat{x}_{wt}^i = \bar{x}_{wt}^i \quad \text{for all } w \notin \mathcal{W}^* \text{ and } i \in \mathcal{I}.
\end{aligned} \tag{4.8}$$

where

$$\begin{aligned}
H_{b,t+\tau}(\hat{\mathbf{x}}_{bt}) &= \min \sum_{w \in \mathcal{W}(b), i \in \mathcal{I}} C_{w,t+\tau+L-1}(\bar{y}_{w,t+\tau}^i) \\
\text{such that} \quad &\sum_{w \in \mathcal{W}(b)} \bar{y}_{w,t+\tau} \leq \hat{x}_{bt}^i \quad \text{for all } i \in \mathcal{I}.
\end{aligned} \tag{4.9}$$

Problem (4.9) can be solved for a large number of values of $\hat{\mathbf{x}}_{bt}$ using marginal analysis to define the function $H_{b,t+\tau}(\hat{\mathbf{x}}_{bt})$. Once this function is known, we can solve problem (4.9) using the same two stage marginal analysis process that we described for model (4.7). In the following section, we construct an inventory allocation model distribution to the pharmacies.

4.3.2 Distributer-Pharmacy Inventory Allocation

At the pharmacy level, cross-shipments of inventory are unlikely, but customers may be sent to nearby pharmacies that have excess antivirals on-hand if shortages arise at some pharmacies. While forcing sick customers to travel from store to store in search of inventory would not be ideal, it is preferable to making them wait for a future shipment to arrive, since the effectiveness of antivirals decreases significantly if treatment is delayed. In this model, our first goal is to minimize unused inventory at pharmacies and the likelihood of patients arriving to a pharmacy where no inventory is present. Our second and more important goal is to minimize the likelihood that all of the pharmacies in a particular area experience a stockout. We assume that all of the pharmacies in a local area are served by the same distribution warehouse. Let $\mathcal{H}(w, k)$ be the k^{th} subset of pharmacies served by warehouse w such that all of the pharmacies may share patient demand.

Pharmacies receive daily shipments from the distribution warehouses. This inventory can be used to serve patients the same day, so we have zero lead times between each warehouse and its pharmacies. The costs incurred at each pharmacy n on day t for inventory type i include per-unit holding cost h_{nt}^i and per-patient relocation costs c_{nt}^{Ri} which are charged if the patient cannot receive inventory at the first pharmacy he visits. The cost incurred for inventory of type i at pharmacy n on day t is thus given by

$$C_{nt}^i(\bar{y}_{nt}^i) = E\left[h_{nt}^i(\bar{y}_{nt}^i - D_{nt}^i)^+ + c_{nt}^{Ri}(D_{nt}^i - \bar{y}_{nt}^i)^+\right], \quad (4.10)$$

where, as before, \bar{x}_{nt}^i and \bar{y}_{nt}^i are the echelon inventory positions before and after allocation decisions are made, respectively. There is also a cost charged if the entire group of pharmacies $\mathcal{H}(w, k)$ runs out of inventory. Let c_{nt}^{Ei} be the cost incurred for patient who cannot be served with the available inventory. We assume that $h_{nt}^i < c_{nt}^{Ri} < c_{nt}^{Ei}$.

Due to the unpredictability of demand and the fact that patients will be sent to seek inventory at different pharmacies in the event of stockouts, we can reasonably assume that inventory imbalance will not be a problem. Therefore, the only constraints on inventory allocation are

$$\sum_{n \in \mathcal{H}(w)} \bar{y}_{nt}^i \leq \bar{x}_{wt}^i \quad \text{for all } i \in \mathcal{I}.$$

Thus, the complete model for the pharmacies served by warehouse w is

$$g_{wt}(\bar{\mathbf{x}}_t) = \min \sum_{n \in \mathcal{H}(w)} C_{nt}^i(\bar{y}_{nt}^i) + \sum_k c_t^{Ei} E \left[\left(\sum_{n \in \mathcal{H}(w,k)} (D_{nt}^i - \bar{y}_{nt}^i) \right)^+ \right] \quad (4.11)$$

such that $\sum_{n \in \mathcal{H}(w)} \bar{y}_{nt}^i \leq \bar{x}_{wt}^i$ for all $i \in \mathcal{I}$.

Because of the emergency resupply cost term, the optimal inventory level for each pharmacy cannot be determined as easily as in the previous section. We must first calculate lower bounds on the optimal values by neglecting the cost of the emergency resupply; the actual desired inventory level would likely be higher to decrease the likelihood of paying the emergency resupply cost. Let \bar{y}_{nt}^{Li} be the lower bound on the optimal level for inventory of type i at pharmacy n . Then

$$\bar{y}_{nt}^{Li} = \left\lceil (F_{nt}^i)^{-1} \left(\frac{c_{nt}^{Ri}}{c_{nt}^{Ri} + h_{nt}^i} \right) \right\rceil.$$

If $\bar{x}_{wt}^i \geq \sum_{n \in \mathcal{H}(w)} \bar{y}_{nt}^{Li}$, then the inventory assignments will have $\bar{y}_{nt}^i \geq \bar{y}_{nt}^{Li}$ for all pharmacies n , because this will minimize the cost at each location; some pharmacies may receive additional inventory to minimize the expected cost of emergency resupply. These additional units may be assigned using a marginal analysis-type algorithm. If $\bar{x}_{wt}^i < \sum_{n \in \mathcal{H}(w)} \bar{y}_{nt}^{Li}$, we may also use a marginal analysis algorithm to determine the optimal solution $\bar{\mathbf{y}}_t$ and the optimal cost $G_t(\bar{\mathbf{x}}_t)$.

All three of the models (4.7) - (4.11) include a number of approximations and assumptions, but they provide solvable methods for making inventory allocation decisions that take into account the state of all of the locations in the distribution network and information about demand.

4.4 Ongoing and Future Work

In this chapter, we have discussed the possibility of using the commercial pharmaceutical supply chain to operate a controlled antiviral-dispensing campaign during an influenza pandemic. We constructed a set of hypothetical regional antiviral demand curves based on historical epidemiological data. We also built a simulation of the pharmaceutical dispensing network and showed that it can perform well under a variety of scenarios, but that improved inventory allocation methods are necessary. We proposed several such methods in the previous section.

We are continuing to collaborate with CDC officials to make the simulation a useful public health decision-making tool. Some improvements that are currently underway include allowing patients to seek service at multiple pharmacies on the same day and displaying histograms of patient demand to better represent pharmacy throughput requirements. We also plan to implement more sophisticated forecasting methods as well as the inventory allocation strategies described above to show the high level of performance that the system may be capable of under a strong command and control system. This work will be ongoing as we continue modifying and improving the models to support the CDC in constructing a highly effective pandemic influenza response plan.

APPENDIX A

D-PODS USER MANUAL

A.1 Introduction

The Dynamic Point of Dispensing Simulator (D-PODS) is a tool designed to help local public health officials evaluate the operations of a single Point of Dispensing (POD) during a mass prophylaxis campaign. D-PODS is a Monte Carlo simulation model that accepts user inputs describing a POD's layout, patient arrival patterns, and staffing plan, and then outputs a probabilistic assessment of system performance over time. This user manual explains the features and limitations of D-PODS and offers some advice on using D-PODS to successfully model POD systems.

D-PODS is implemented in Microsoft Excel, Access, and Visual Basic. A large number of user inputs are required to run the simulation. These are entered by the user in a series of four worksheets. Once the simulation has run, a number of output statistics and plots are automatically generated by D-PODS. This guide will explain how to get started with D-PODS and describe the information required by the user input sheets. The guide will further identify potential pitfalls and assumptions implicit in the model that may cause surprising results. Finally, the guide will discuss how the simulation outputs can be analyzed and interpreted and how D-PODS can be used to explore policy decisions.

A.2 Glossary of Important Terms

1. Arrival Rate - the expected hourly rate at which patients arrive to the POD (this can vary throughout each day).
2. Arrival Types - different groups of patients who may require different service times based on their special needs. These might include single adults, mobility-limited individuals, non-English speakers, etc.
3. Maximum Average Waiting Time - the longest acceptable patient waiting time in minutes. This input is used by the D-PODS staffing calculator to estimate reasonable staffing levels.
4. Queue Length - the total number of people waiting to be served at a given location.
5. Routing Probability - the probability that a patient will be routed from one station to the next station in the layout of the model. The routing probabilities are sometimes called “transition rates.”
6. Service Time Distribution - the probability distribution that governs the length of time that will be required to serve a single patient at a work station. These distributions are assumed to be triangular and hence are defined by minimum, most likely (mode), and maximum values.
7. Service Time Increase Factor - the percent increase in service time required by different arrival type groups. The service times for different arrival types are calculated by drawing a random service time from the appropriate service time distribution and then multiplying this number by one plus the service time increase factor. For example, if the service time increase factor is 100%, the service time doubles for that arrival type.

8. Simulation Replication - one repetition of the simulated prophylaxis campaign.
9. Time Period - the duration of time over which the expected patient arrival rate to the POD is constant, also referred to as an “arrival interval.”
10. Worker Interval - the duration of time for which the number of staff is constant at all stations.

A.3 Working with D-PODS

To use D-PODS to model a particular POD, a larger number of user inputs will be required. These data are entered in a series of four worksheets.

A.3.1 Getting Started

You will need Microsoft Excel and Access, versions 2003 or later, as well as the files “PoD_DB.mdb” and “PoD_v27.xls.” Make sure that you have both the files “PoD_DB.mdb” and “PoD_v27.xls” stored in the same directory; otherwise D-PODS will not run. No other files should be placed in the D-PODS folder because they may cause errors to arise while running the program. If you ever try to run the simulation and get an error that stops the simulation before it completes, then open the directory where these files are stored and delete all files except these two.

To get started, open the Excel file “PoD_v27.xls.” A “Security Warning” message box may prompt you to set appropriate security settings; if this happens,

select “Enable Macros.” Once the file is open, you will need to ensure that your version of Excel includes all of the required libraries. To do this, go to Tools → Macros → Visual Basic Editor. From the Visual Basic Editor window, go to Tools → References. Make sure that the following references are selected:

1. Visual Basic For Applications
2. Microsoft Excel 11.0 Object Library
3. Microsoft DAO 3.6 Object Library
4. Microsoft Forms 2.0 Object Library
5. OLE Automation
6. Microsoft Office 11.0 Object Library
7. Microsoft ActiveX Data Objects 2.8 Library
8. Microsoft OLE DB Error Library.

Then click “Okay” and return to the Excel-PoD_v27 window. You should only need to complete this step the first time you run D-PODS.

A.3.2 Creating or Selecting a Case

The “PoD_v27.xls” file opens to a cover page. Click the “Begin Model” button in the lower right-hand corner of the screen, and the “Case Selection” page, shown in Figure A.1, will appear. A “case” is a complete set of user inputs and simulation outputs for a particular POD. You may choose to create a new case from scratch by choosing the “Create New” option or you may open an existing case by choosing the “Existing Case” option. If you select the former, you will

be asked to confirm the choice and then taken to the “D-PODS Menu” page. If you select the latter, a user form, shown in Figure A.2, will pop up, showing a list of all cases currently in the database.



Figure A.1: Case selection screen shot.

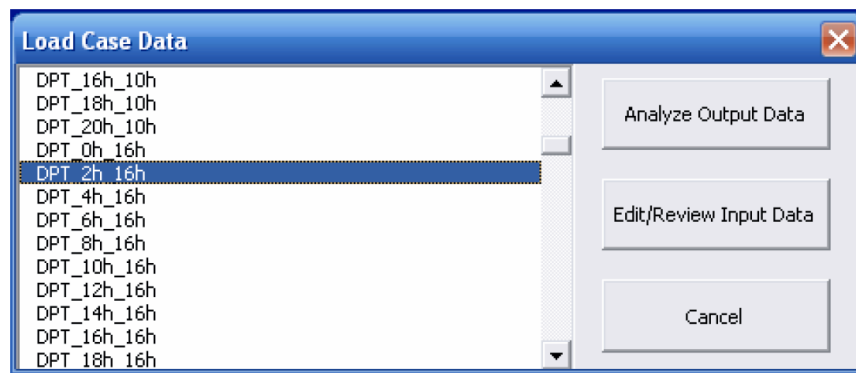


Figure A.2: Load case data screen shot.

After selecting the case of interest, you can choose to “Analyze Output Data,” which will allow you to immediately begin reviewing the case’s output, or “Edit/Review Input Data,” which will allow you to examine and make changes to the case’s inputs. You can also click “Cancel,” which returns the screen to the “Case Selection” page. Clicking “Edit/Review Input Data” will bring you to the “D-PODS Menu” page, while clicking “Analyze Output Data” button will bring you directly to the “Output Tables” page.

A.3.3 D-PODS Menu

The “Main Menu,” shown in Figure A.3 shows the different choices you have when you are working with D-PODS and suggests the best order in which to perform each task. Steps 1 through 4 guide you in filling in the information required by each input worksheet. Steps 5 through 7 are concerned with the details of saving inputs and running the simulation. Step 8 takes you to various sheets that display the simulation output.

The buttons at the bottom of the menu allow you to work with the database file or to restart the program. If you click the “Existing Case” button, you will be prompted with the same “Case Selection” form shown in Figure A.2. You can then select a different case to study. The “Manage Cases” button brings up a similar form, but this one, shown in A.4, allows you to delete cases. Deleting old cases allows you to manage the size of the database and to clean out old data that are no longer of interest. Before a case is deleted from the database, a warning box will appear to confirm you want to delete the case. The “Start Over” button will bring you back to the cover page and allows you to start from scratch.

A.3.4 Step 1: Constructing the Model

When you click the “Model” button in Step 1 of the D-PODS menu, the “Construct the Model” worksheet will appear, which is shown in Figure A.5. To complete this input step, complete Steps A, B, and C in order. In Step A, you first specify the duration of the simulation in days. You may want to begin by considering only one or two days, rather than the entire duration of the planned

D-PODS Menu

**In order to run the program, follow the steps as shown below.
The program may not work if executed in the incorrect order.**

Step 1	Construct the model	Model
Step 2	Input arrival rate	Arrivals
Step 3	Input the service time parameters	Service
Step 4	Establish staffing levels	Staffing
Step 5	Enter simulation parameters	
	Number of simulation replications	2
	Random Seed	150
Step 6	Input Case Name	[Enter Name]
Step 7	Run the simulation	Run
Step 8	View the Results. The results are displayed in both tabular and graph form.	Output Tables
		Output Graphs

Other Options

Select an existing case.	Existing Case
Manage case list.	Manage Cases
Go to Cover Page to Start Over	Start Over

Figure A.3: D-PODS menu screen shot.

prophylaxis campaign, because the simulation run time is proportional to the number days simulated; doubling the number of days approximately doubles the time required to run the simulation. Also, it is important to note that D-PODS assumes that all days are identical. That is, the POD is required to have the same operating plan and patient arrival pattern for each day. Beyond ran-

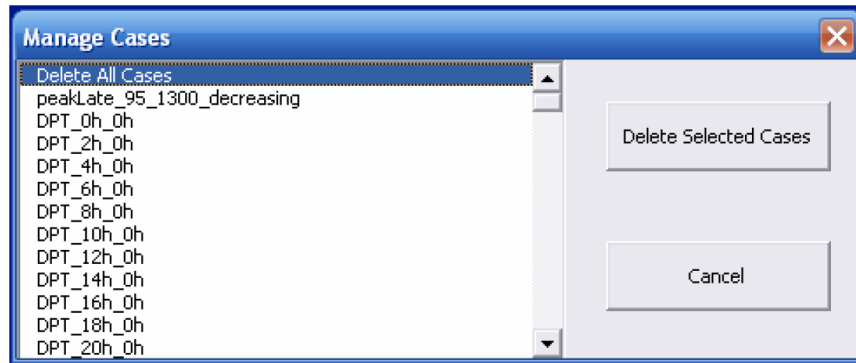


Figure A.4: Manage cases screen shot.

dom variation, the only reason why two days might have significantly different performances is if patient arrivals from one day cannot be served by the end of the day and must be helped during the following work day. We discuss this further below.

Step 1: Construct the Model

Main Menu

Construct Model

Generate Arrivals

Set Service Parameters

Establish Staffing Levels

Run Simulation

Step A		
Input duration of the campaign in days		2
Input hours of operation per day		22
Step B		
Number of Stations (maximum is 15)		4
Input station names and exits		Station Names
Input station transition probabilities		Transition Rates
Step C		
Input PoD capacity		50000
Step D		
Return to Main Page		Return

Transition Rates Table

	Greeting	Triage	Medical Eval	Drug Dispensing	Exit
Greeting		0.95	0.05	0.00	0.00
Triage			0.05	0.95	0.00
Medical Eval				0.95	0.05
Drug Dispens					1.00

The first station is the immediate station after entering the PoD. All subsequent stations should be listed in a downward flowing manner. Meaning, if station j can be reached from station i, then i should be listed before station j. Likewise, one must exit from the last station.

Station Names Table

Station	1	2	3	4
Name	Greeting	Triage	Medical Eval	Drug Dispensing
	Exit			

The transition rate matrix below indicates the probability of flowing from state i to state j within the PoD. The Exit column in the transition rate matrix below the probability of flowing out of the PoD from any of the stations.

Figure A.5: Step 1 screen shot.

The “hours of operation per day” is the length of time during which patients are allowed to arrive and enter the POD. In Figure A.5, this value is set to 22.

That means that any patients who arrive to the POD between the 22nd and 24th hour of the day will be turned away without service. However, any patients who have already entered the POD will be served as long as staff is available, even if this means that the POD remains open for longer than 22 hours. It is possible that not all of these patients will be served by the time patients begin arriving again for the next day; in this case, the newly arrived patients simply queue behind the waiting patients. On the last day of the simulation, however, the simulation will stop at the correct time, even if some patients have not been served.

Step B asks that you describe the layout of the POD by declaring how many workstations will be in use and their names. D-PODS assumes that each station has one queue where patients wait for service. There can be 0 or more staff working at each station, and the person at the front of the queue will be served by the next available staff person at that station. When the patient has finished receiving service, he immediately enters service at the next station or the queue for that station, if no server is currently available. Travel time between stations is assumed to be insignificant.

To complete Step B, first enter a whole number for the “number of stations” and then click the “Station Names” button. This generates the “Station Names Table.” Be careful, however, because clicking this button also clears current entries from the table, and the “Undo” function in Excel will not let you get the information back. If you want to save the information, copy it to another location before clicking the button. It is important to enter the station names in the correct order, because it is not possible for patients to travel from higher-numbered stations to lower-numbered stations. So, if you name the first station “Greeting”

and the second station “Triage,” then it will be possible (but not required) for patients to move from Greeting to Triage, but it is not possible to go from Triage to Greeting. Note that this makes it impossible for patients to visit any station in the POD more than once. So, the first station listed in the “Station Names Table” should correspond to the immediate station that patients encounter upon entering the POD. The last station in the POD is a “dummy” station called Exit, which indicates that patients leave the POD after receiving treatment.

After completing the “Station Names Table,” you can click on the “Transition Rates” button; this will generate the “Transition Rates Table,” which allows you to input patient routing probabilities through the POD. So, for example, the entry in row 3, column 4 of the table gives the probability that a patient at station 3 will move to station 4 after she has finished receiving service at station 3. Since all patients must move to a new station after service, the probabilities in each row must sum to one. If the values do not add up to one, a warning box will appear when you try to move to a new worksheet. Notice that some cells in the “Transition Rates Table” are greyed out; this enforces the condition described above, making it impossible to move from a higher-numbered station to a lower-numbered one. It also makes it impossible for a patient to remain at the same station.

In Step C, you can enter the POD capacity, which sets the maximum number of patients allowed in the POD at any point in time. The capacity may depend on the size of the building or it may be set low to limit the risk of transmitting disease. When the number of patients exceeds the POD capacity, patients form a line outside of the POD and wait until another patient leaves before entering the queue for the first station inside the POD. Since these patients may be

turned away without being counted at the end of the day and they are not included in statistics describing the average time in the POD, some of the output statistics may be misrepresentative if the queue outside the POD grows large. In the experiments described in Chapter 3, the POD capacity was set to be almost five times larger than the expected number of daily patient arrivals, so that all patients who arrived to the POD immediately entered the queue for the first station.

Once Steps A - C are complete, you can proceed to “Step 2: Input Arrival Information.” There are two ways to do this. One option is to return to the “D-PODS Menu” page by clicking the “Return” button in Step D and then clicking the “Arrivals” button on the menu page. The other option is to simply click the “Generate Arrivals” button on the left side of the screen.

A.3.5 Step 2: Input Arrival Rate Information

Step 2 allows you to enter information describing patient arrivals throughout each day. Figure A.6 shows the “Input Arrival Rate Information” worksheet. In Step A, you must decide how many patient arrival intervals you want to use to describe the daily patient arrival pattern. You will be able to set a new expected patient arrival rate for each arrival interval, but during each interval the expected patient arrival rate remains constant. The number of arrival rate intervals per day, can be anywhere between 1 and 96. Choosing 96 as the value means that each arrival interval will be 15 minutes long, because there are 24 hours in each day.

In Step B, you can set the simulation start time. If you want to simulate a

Step 2: Input Arrival Rate Information		
Step A	Input number of arrival intervals	12
Step B	Input start time (Must be input in time format)	7:00 AM
Step C	Input number of arrival types. Arrival types include single person, families, elderly, non english speakers, etc.	2
Step D	Generate chart based on information provided in steps A and B	Chart
Step E	Fill in the highlighted cells of the chart on the right based on the headers	
Step F	Return to Main Page	Return

Interval Number	Hours per Interval	Interval Start Time	Interval End Time	Arrivals Per Hour: Type 1	Arrivals Per Hour: Type 2
1	2.00	7:00 AM	9:00 AM	300	150
2	2.00	9:00 AM	11:00 AM	200	100
3	2.00	11:00 AM	1:00 PM	500	250
4	2.00	1:00 PM	3:00 PM	500	250
5	2.00	3:00 PM	5:00 PM	700	350
6	2.00	5:00 PM	7:00 PM	400	200
7	2.00	7:00 PM	9:00 PM	200	100
8	2.00	9:00 PM	11:00 PM	100	50
9	2.00	11:00 PM	1:00 AM	100	50
10	2.00	1:00 AM	3:00 AM	50	25
11	2.00	3:00 AM	5:00 AM	0	0
12	2.00	5:00 AM	7:00 AM	200	100

Figure A.6: Step 2 screen shot.

POD scenario in which patients begin arriving before the POD opens, you can set the staffing levels at the beginning of the day to 0 at each station in Step 4: Establish Staffing Levels. If you want to simulate a POD that opens before patients begin coming, enter “0” for the patient arrival rates at the beginning of the day in Step E.

Step C allows you to enter the number of patient arrival types. Your basic arrival type may be single, mobile, English-speaking adults, but other useful arrival types may include disabled individuals, school children, non-English speakers, or any other groups of special interest who require service times different from the average. In Step 3, it will be possible to define different service rates for the different patient arrival types.

Step D requires that you click the “Chart” button to generate the patient arrival rates table. Then, in Step E you can define the arrival intervals and the patient arrival rates during each time interval. In the first input column you will enter the duration of each arrival interval in hours; these durations do not all need to be the same, but each one must be rounded to the nearest quarter hour

and the values must sum to 24 hours. The next two columns show the start and end time of each interval. The last two columns allow you to enter the expected patient arrival rates during each interval. Any blank spots in the chart will be treated as 0's. Note that the values entered here are the expected rather than the actual arrival rates. Patients arrive to the POD according to a nonhomogeneous Poisson process, which means that the interarrival times between subsequent patients are distributed according to a Poisson distribution. The values in this chart provide the means for the Poisson process at each point in time.

Note that the arrival rates can only be entered for a single day, because we assume that the arrival pattern on the subsequent days is identical. As mentioned earlier, if the expected arrival pattern varies by day, then separate simulations must be run for each individual day.

After you have completed the table, you may proceed to Step 3: Service Time Parameters. To do this, either click "Return" to go back to the "D-PODS Menu" page and then click the "Service" button, or simply use the "Set Service Parameters" button on the left side panel.

A.3.6 Step 3: Service Time Parameters

Step 3 allows you to input parameters that govern how long it takes to serve each patient type at each station. Figure A.7 below shows the Step 3 worksheet. The service time at a given station for each patient is drawn from that station's service time probability distribution, which is assumed to be triangular. D-PODS further assumes that each patient is processed individually and that the service time parameters for each station and arrival type remain constant

over time; there is no accounting for changes in speed due to worker fatigue or learning curve.

Step 3: Service Time Parameters			
Step A	Generate chart to input percentage increases in service time and chart to input parameters of service time for a single person for each station.	Chart	
Step B	Fill in the cells of the chart on the right with the percent increase. The first arrival type (single person) is the baseline and has zero percent increase in service time.		
Step C	Fill in the cells of the chart on the far right with the parameters of service time for a single person (in minutes).		
Step D	Return to Main Page	Return	

Arrival Type	Service Time Increase Factor
1	0%
2	20%

Station	Minimum	Most Likely	Maximum
Greeting	0.25	0.50	1.00
Triage	1.00	2.00	3.00
Medical Eval	2.50	5.00	10.00
Drug Dispensing	0.50	1.00	2.00

Figure A.7: Step 3 screen shot.

In Step A, you must click the “Chart” button to generate the two charts on the right side of the screen. The charts include rows for the stations you defined in Step 1 and the number of arrival types you entered in Step 2. Default charts may be displayed before you click this button, but it is important that you click the “Chart” button if you have changed number or names of stations or the number of arrival types.

Step B requires that you fill in the Service Time Increase Factor table. Choose one of your arrival types to be the base arrival type; enter “0” or “0%” for this arrival type’s service time increase factor. The service time increase factors for the other arrival types describe how the processing time of the other arrival types compare to your base arrival type. For example, if Arrival Type 1 is your base arrival type, with a service time increase factor of 0 and the service time increase factor for Arrival Type 2 is 20%, then the service times for Type 2 patients will be 20% longer than the service times for Type 1 patients, on average. If Type 2 patients have the same service time distributions as Type 1 patients, then you should set the Type 2 service time increase factor to 0%. If Type 2 patients are processed 20% faster than Type 1 patients, then for the Type 2 service

time increase factor you should enter “-20%.” All numbers in this table must be between -100 and 100. Note that the same service time increase factor is applied to all of the stations.

To complete Step C, you must fill out the far right table. Enter the parameters that describe the triangular service time distributions at each station for Arrival Type 1. The minimum and maximum times are strict limits on the service times, and the most likely time is the peak of the triangular distribution, which means it is the mode of the service times. All three parameters have units of minutes. The minimum time must be less than or equal to the most likely time, which must be less than or equal to the maximum time. If all three values are the same, then the service time will be constant at that station.

After completing both tables, you may proceed to Step 3: Service Time Parameters. To do this, either click “Return” to go back to the “D-PODS Menu” page and then click the “Staffing” button, or simply use the “Establish Staffing Levels” button on the left side panel.

A.3.7 Step 4: Staffing Requirements

The fourth input step is setting the staffing plan at each station throughout each day. Figure A.8 shows a screen shot of the Staffing Requirements sheet. Step A allows you to set the maximum number of times during the day that you want to change the number of staff at each station. If you plan to keep the same number of staff at each station all day long, you may enter a 1 here. The period of time during which staffing levels at each station are constant is called a “worker interval.” D-PODS does not allow worker intervals shorter than two

hours, so the maximum number of worker intervals each day is 12. That is, you may not change the number of staff at each station more than 12 times each day. The second line reminds you the start time for the POD operations, which you entered in Step 2: Arrival Rate Information.

Step 4: Staffing Requirements													
Step A	Input number of worker intervals per day		8										
	Start time from arrivals page		7:00 AM										
Step B	Generate staff interval chart		Chart										
Step C	Input number of hours per worker interval												
Step D	Calculate the maximum arrival rate per hour for each worker interval		Arrival										
Step E	Fill in the desired maximum average queue waiting time in minutes in the chart below												
Step F	Calculate advised number of workers per interval based on the max in Step D		Advice										
Step G	Fill in the highlighted columns for the desired number of staff at each location for each worker interval. If you'd like to use the advice data, click the button to the right.		Use advice?										
Step H	Return to Main Page		Return										

Worker Interval	Hours per Interval	Interval Start Time	Interval End Time	Maximum Arrival Rate	Max Avg Waiting Time	Greeting Advice	Greeting Input	Triage Advice	Triage Input	Medical Eval Advice	Medical Eval Input	Drug Dispensing Advice	Drug Dispensing Input
1	3	7:00 AM	10:00 AM	450	5	5	5	16	16	9	9	10	10
2	3	10:00 AM	1:00 PM	750	5	8	8	26	26	13	13	16	16
3	3	1:00 PM	4:00 PM	1050	5	11	11	36	36	17	17	22	22
4	3	4:00 PM	7:00 PM	1050	5	11	11	36	36	17	17	22	22
5	3	7:00 PM	10:00 PM	300	5	4	4	11	11	7	7	7	7
6	3	10:00 PM	1:00 AM	150	5	2	2	6	6	5	5	4	4
7	3	1:00 AM	4:00 AM	75	5	1	1	3	3	4	4	2	2
8	3	4:00 AM	7:00 AM	300	5	4	4	11	11	7	7	7	7

Figure A.8: Step 4 screen shot.

Step B involves clicking the “Chart” button to generate a new chart at the bottom of the page. Note that clicking this button will clear the information currently in the table, and this cannot be undone. Be sure to save any important information about your previous work before clicking “Chart.” If the number of worker intervals, number of stations, and station names are the same as those currently shown in the chart, you do not need to click the “Chart” button.

Step C requires that you enter the duration of each worker interval in the

“Hours per Interval” column of the table. This column must total 24 hours and each interval must be at least two hours long, but the intervals do not have to be of equal lengths. If the intervals do not sum to 24 hours, then a message will appear at the bottom of the table. The next two columns of the table display the start and end times of each interval.

In Step D you must click the “Arrival” button to display the maximum expected hourly patient arrival rate during each interval in the “Maximum Arrival” column. This information is taken from the patient arrivals table in Step 2: Input Arrival Information. Even if you do not plan to use this information in developing your staffing plan, you must click the “Arrival” button, or a warning will appear that prevents you from proceeding to the next step.

In Step E, you must fill in the maximum desired patient waiting time (or “Max Avg Waiting Time”) column of the table. The numbers in these columns indicate the maximum expected waiting time for each station that you would like to have during each segment of the day. It is not possible to set different maximum expected waiting times for different stations. Once you have entered these values, which must be positive real numbers, you must click the “Advice” button in Step F. The D-PODS staffing calculator then uses a calculation from [Buzacott & Shanthikumar, 1993] to estimate the number of staff required at each station to limit the maximum average patient waiting times. These estimates are shown in “Advice” columns of the table. Note that D-PODS requires that you complete Steps E and F, even if you do not plan to use this information in your staffing plan.

Step G requires you to fill in the “Input” columns of the table with the number of staff at each station during each worker interval. One way that you may

choose to do this is by using the “Use Advice?” button, which copies the values from the “Advice” columns into the “Input” columns. If you click the “Use Advice?” button, a warning will pop up to ask if you are sure that you want to use the advice. If you click “Yes” to indicate that you really do want to use the advice from the D-PODS staffing calculator, any numbers currently in the “Input” columns will be replaced by the advice numbers.

Once all of the “Input” columns are complete, you may return to the “D-PODS Menu” page by clicking “Return” or by clicking the “Menu” page on the left side panel. It is also possible to immediately begin running the simulation by clicking the “Run” button on the left side panel, but if you do this you will not have a chance to save your inputs as a new case or modify the simulation parameters.

A.3.8 Step 5: Simulation Parameters

To complete Step 5, you must input the simulation parameters on the “D-PODS Menu” page. The number of simulation replications is the number of complete prophylaxis campaigns you would like to simulate. If, for example, you set your campaign to be 3 days long and you declare that you want to run 100 replications, D-PODS will simulate the 3 day campaign 100 times. Note that the run time is proportional to the number of replications, so doubling the number of simulation replications will double the time required to run the simulation. If you are not sure how fast D-PODS will run on your computer, start off with only a few replications to avoid crashing Excel. If your computer is fairly new and fast, you should have no trouble running 20 or more replications of D-PODS at

one time. The number of simulation replications should be a positive integer; if you enter a non-integer number, it will automatically be rounded.

You may also enter a new “Random Seed” in Step 5. This number should be a positive integer. It is used in the Visual Basic random number generator to determine the random numbers used in the simulation. If you use the same random seed with exactly the same set of inputs, you will get the same output. However, if you use the same set of inputs but change the random seed, the simulation output will be different (although the average values will likely be quite similar) because a different set of random numbers will be used for the patient arrival and service times. If you want to compare two different sets of inputs and keep the comparison as “fair” as possible, then it is a good idea to use the same random seed for both simulations. This helps eliminate random error as a confounding factor in your comparison. Step 5 can be completed at any point during the input process.

A.3.9 Step 6: Input Case Name

In Step 6 you have an opportunity to save the inputs that you have entered and the simulation output that will be generated when you run the simulation. In the light blue box that initially reads “[Enter Name]” click and type the name you would like to use. If you do this, you will be able to open this case in the future using the Select Existing Case dialog. If you forget to enter a case name before running the simulation and you want to save your work, you have two options. One is to return to the D-PODS menu, enter a case name in the Step 6 box, and then re-run your simulation. Another option is to save the entire Excel

workbook under a new name, but if you do this you may not be able to save all of the simulation output.


A.3.10 Step 7: Run the Simulation

To execute the simulation, you must click the “Run” button in Step 7. The simulation run time depends primarily on the number of simulations runs you have chosen; the number of days in your prophylaxis campaign, and the number of patient arrivals. Large simulations may easily take up to 5 or 10 minutes to complete. If you think the run time is too long, you can stop the simulation by hitting the “Esc” key. Then click “End” in the error dialog box that pops up. If you do this, you must go to the directory where you have saved D-PODS and delete any new files that have been generated; there will likely be one called “Simulation.Output.txt.” If you do not delete this file, D-PODS will not run in the future. You may then modify your simulation by decreasing one or more of the parameters mentioned above to speed up the running time. You will know that the simulation is complete when the mouse pointer changes from a spinning circle or hourglass back into an arrow.

A.3.11 Step 8: View the Results

D-PODS produces several of output tables and graphs that you can use to analyze POD performance. From the D-PODS menu, click on either the “Output Tables” button or “Output Graphs” button to begin viewing the results of your simulation. You can easily move between the two using the buttons on the left

side panel or by returning to the D-PODS menu at any time. Clicking the “Output Tables” button will bring you to the “Output Tables” worksheet, as shown below in Figure A.9.



Cornell University

Output Tables

Number of Arrivals	6,028	6,028	0.00
Total Throughput	6,028	6,028	0.00
Average Throughput (pt./min.)	2.09	2.09	0.00
Average Time in POD (min./pt.)	5.37	5.37	0.00
Average Number in POD (pt./min.)	11.24	11.24	0.00
Average Time in System (min./pt.)	5.37	5.37	0.00
Average Number in System (pt./min.)	11.24	11.24	0.00

Performance Measures at Each Station by Worker Interval:

Select Worker Interval:

Figure A.9: Output tables initial screen shot.

The tables at the top of the sheet show summary statistics that describe the entire prophylaxis campaign. The number of patient arrivals is shown, as is the total patient throughput. Patient throughput is the number of patients who were processed during the campaign. If all patients who arrived were processed before the end of the campaign, these two numbers will be the same. The means and standard deviations are taken across all of the simulation replications.

“Average Throughput” is calculated for each simulation replication by dividing the total throughput by the operating hours. The mean and standard deviation of these numbers are then reported. “Average Time in POD” is calculated for each simulation replication by taking the mean of the patient times in the POD. The numbers reported are calculated from these averages. “Aver-

Performance Measures at Each Station by Worker Interval:			
4			
	Estimate*	Mean	Std Dev
Queue Outside of POD:			
Average Queue Length (patients)	0.00	0.00	0.00
Maximum Queue Length (patients)	0.00	0.00	0.00
Average Wait in Queue (minutes)	0.00	0.00	0.00
Station 1			
Average Queue Length (patients)	0.08	0.14	0.07
Maximum Queue Length (patients)	1.00	1.50	0.71
Average Wait in Queue (minutes)	0.07	0.07	0.00
Maximum Staff Used	2.00	2.00	0.00
Staff Utilization	0.45	0.51	0.08
Station 2			
Average Queue Length (patients)	0.63	0.85	0.32
Maximum Queue Length (patients)	5.00	6.50	2.12
Average Wait in Queue (minutes)	0.36	0.47	0.15
Maximum Staff Used	4.00	4.00	0.00
Staff Utilization	0.78	0.78	0.00
Station 3			
Average Queue Length (patients)	0.00	0.00	0.00
Maximum Queue Length (patients)	0.00	0.00	0.00
Average Wait in Queue (minutes)	0.00	0.00	0.00
Maximum Staff Used	3.00	3.50	0.71
Staff Utilization	0.18	0.22	0.05
Station 4			
Average Queue Length (patients)	0.19	0.20	0.01
Maximum Queue Length (patients)	2.00	2.00	0.00
Average Wait in Queue (minutes)	0.07	0.08	0.01
Maximum Staff Used	3.00	3.00	0.00
Staff Utilization	0.54	0.59	0.07

Figure A.10: Output tables station performance measures screen shot.

age Number in POD” is the average number of patients in the POD at a given point in time. The values reported in the output table are the sample mean and standard deviation of these numbers.

“Average Time in System” and “Average Number in System” are similar to the previous two statistics, except that they include the patients waiting outside the POD, so these means will always be at least as large as the “... in POD”

statistics. If there are never patients waiting outside the POD, these values will be equal to “Average Time in POD” and “Average Number in POD,” respectively.

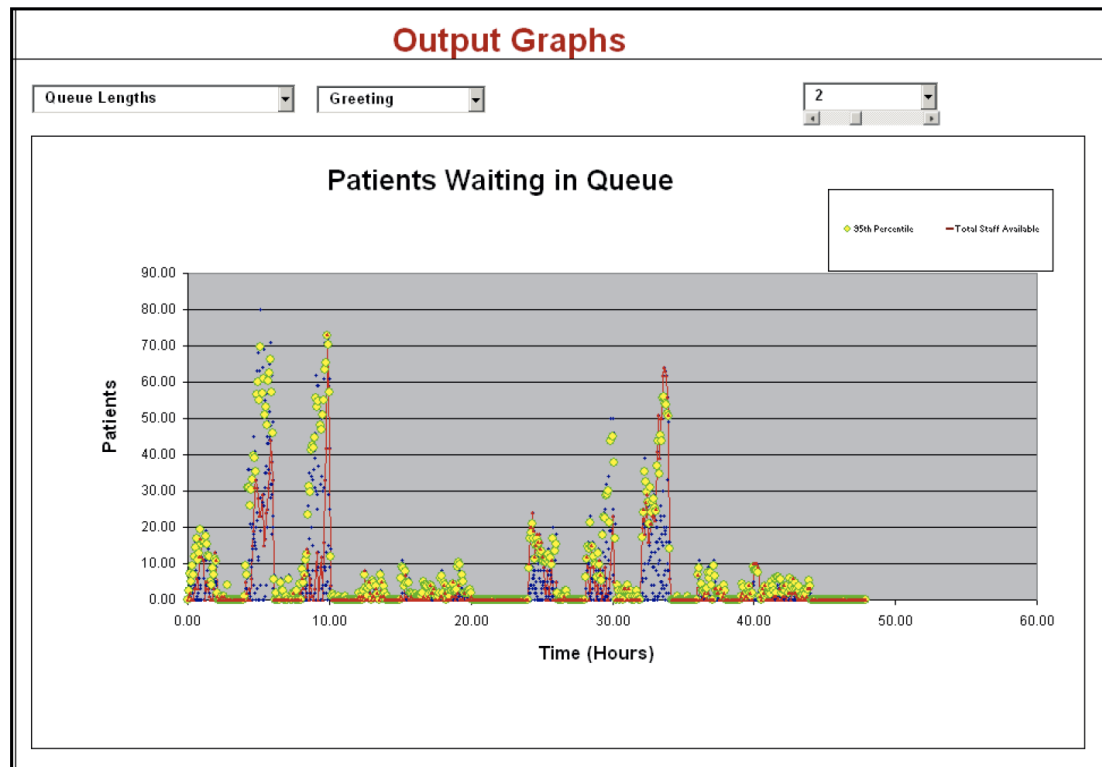


Figure A.11: Screen shot of output graphs after location selection and replication selection.

The statistics are further broken down by workstation and by worker intervals. You can use the drop down box beneath the main output table to select either the entire campaign or a particular worker interval. For the time period chosen, you can see statistics that describe the queue lengths, patient queue times, and staff utilization at each station.

Clicking the “Output Graphs” button on either the left panel from the “Output Tables” page or on the “D-PODS Menu” page brings you to the “Output

Graphs” page. The drop down box allows you to choose the type of data you would display: “Queue Lengths,” “Number of Staff Utilized,” or “Arrivals Per Interval.” Once you have made a selection, two new boxes will appear. The first allows you to select the location for which you would like to see information. This location may be any of the work stations in the POD or “Outside of POD.” However, note that if your POD is large enough to accommodate all arriving patients, then there will not be any interesting data for the “Outside of POD” option. Once you have entered valid options in each of the first two boxes, a plot of the data you have selected will be shown, as in Figure A.11.

Next, you may select a specific simulation replication from the drop down box on the right, or you may choose to view all replications. The scroll bar below the drop down box also controls the simulation replication to be selected. After selecting a specific simulation replication, the data for that replication are highlighted with a red line; in Figure A.11, replication 2 is shown.

The data shown in these plots are drawn from the Access database. In order to plot the data in different ways or view the raw numbers, you must explore the database directly. This is discussed in the next section.

A.4 Navigating the Access Database

When you click the “Run” button to start your simulation, D-PODS immediately saves all input data to the database. The input data are stored in the tables titled “CaseList,” “SingleInputs,” “ArrivalRates,” “ServiceIncreaseFactor,” and “Staffing.”

The “CaseList” table stores the case name specified on the “D-PODS Menu” page. This table also assigns the unique CaseListID to each CaseName, which creates a relationship throughout the tables. All data are stored with the CaseListID included as one of the Access fields. This makes creating queries relating to a specific case name much easier.

The “SingleInputs” table contains all input values that are not entered through a chart within D-PODS, such as the number of simulation replications or the number of days in the prophylaxis campaign. It also contains the station names and station transition probabilities. The values are stored based on the CaseListID, the variable name (varName), and the value the variable holds in the program (varValue).

The “ArrivalRates” table contains the chart shown on the right on the “Step 2” page in D-PODS. The data are organized in the Access table just as it is in D-PODS. The “ServiceIncreaseFactor” table stores the information from the chart on the “Step 3” page in D-PODS that details the service time increase factor for each of the arrival types. The “Staffing” table details the information held in the chart at the bottom of the “Step 4” page in D-PODS.

While the simulation is running, the output data are stored in a text file called “Simulation.Output.txt,” which is transferred to the database at the end of the simulation. These data are moved into the Access tables titled “SingleOutputs” and “Outputs.” The “SingleOutputs” table is organized exactly like the “SingleInputs” table and contains information that is displayed in the tables on the “Output Tables” page in D-PODS, such as the number of arrivals for the duration of the planned campaign. The “Outputs” table contains the output data that are stored at the end of every simulated five-minute interval. These data

include the replication number, the shift number, the event time, the size of the queues at each location, the number of arrivals to each location within the interval, the average wait time for each location, and the number of servers that are busy for each location. These data are organized by CaseListID and can easily be queried to further analyze the output data.

APPENDIX B

ESCOE USER MANUAL

B.1 Introduction

This manual describes a tool called the Emergency Supply Chain Operations Evaluator (ESCOE) that has been created to help policy makers and public health officials evaluate the impact of policy on the operations of the entire supply chain. ESCOE is a Monte Carlo simulation model that accepts user inputs describing a supply chain and then outputs a probabilistic assessment of system performance over time. The goal of this document is to explain the features and limitations of ESCOE as well as to offer advice on using ESCOE to successfully model systems.

ESCOE is implemented in Microsoft Excel, Access, and Visual Basic. A large number of user inputs are required to run the simulation. These are entered by the user in a sequence of eight worksheets in Excel. Once the simulation has run, output statistics and plots are automatically generated by ESCOE. This guide explains how to get started with ESCOE and how to enter information on the user input sheets. The guide will further identify potential pitfalls and assumptions implicit in the model that may cause surprising results. Finally, the guide will discuss how the simulation outputs can be analyzed and interpreted and gives an brief introduction to the structure of the Access database.

B.2 Model Assumptions

ESCOE is a discrete time model; that is, time is divided into periods of equal length. A period may represent any length of time, but the total length of the simulation must be some integer number of periods. For example, a period may represent two hours, which means that the total length of the simulation must be an even number of hours. All parameters of the network, including patient demand rates and service rates, are identical within a single time period, but may change from period to period. At most one shipment may be sent to each location in the distribution network in each time period.

There are four types of locations included in the simulation: the Strategic National Stockpile (SNS); Forward Deployed Stockpiles (FDSs); Receiving, Storing, and Staging Warehouses (RSSs); and Points of Dispensing (PODs). First, we consider the SNS. For this model, we will treat the SNS as if it is one single location with a large stockpile of inventory that is resupplied over time from suppliers with unlimited inventories. Difficulties could arise from this assumption if there are inventory imbalances at the various SNS locations, but this is unlikely given the high degree of organization and cooperation at the SNS level.

The second location type modeled by ESCOE is the Forward Deployed Stockpile (FDS), two of which are shown in the network diagram Figure B.1. An FDS is a federally-controlled stockpile of inventory that can deploy inventory to a small region of the country faster than the SNS. Although there are not currently any FDSs in the SNS distribution network, several are currently being developed. ESCOE users define their distribution networks using a process that we will describe below, and users may choose to include zero or more FDSs in

their networks. The third location type is the Receiving, Storing, and Staging Warehouse (RSS). RSSs receive inventory from the SNS and, in some cases, an FDS, and ship inventory to PODs. FDSs may serve any number of RSSs, but each RSS can be served by at most one FDS.

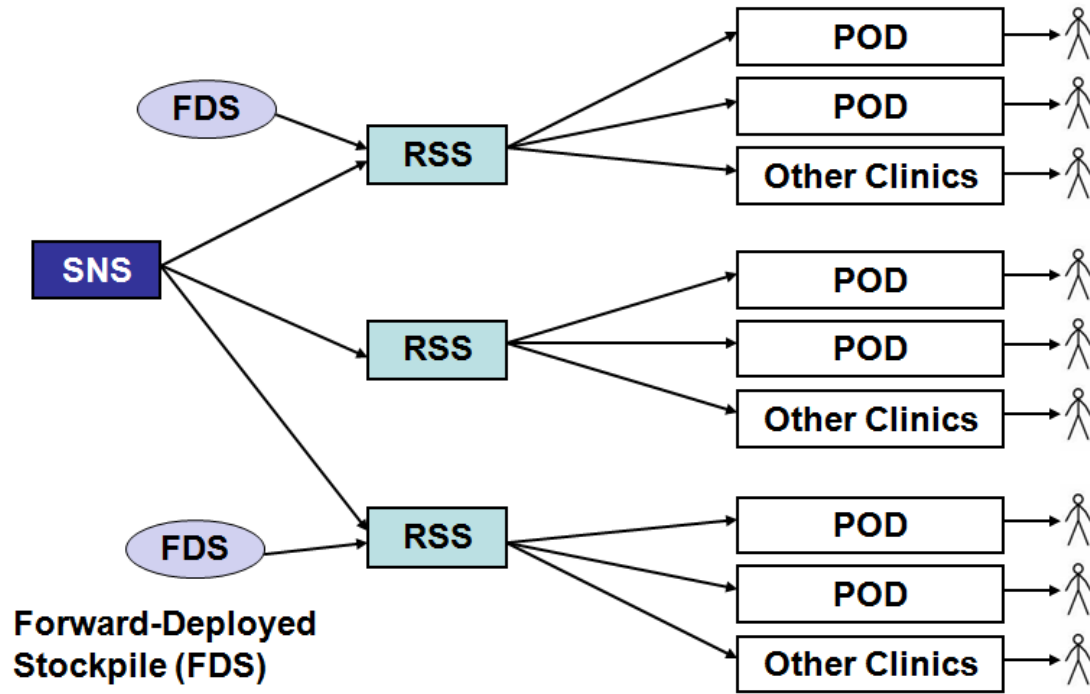


Figure B.1: Emergency supply chain diagram.

The final location type is the Point of Dispensing (POD). A POD is any location where people receive federally-supplied medical countermeasures during an emergency. A POD may be a nursing home or doctor's office or an ad-hoc dispensing clinic set up in a community center, school, or other public space. There may be tens or hundreds of PODs in a distribution network, and describing each individually would be burdensome for an ESCOE user. To reduce the number of input steps required, we allow users to define "POD types." Each POD Type has its own nonstationary arrival patient arrival patterns, opening and closing times, service capacities, resupply lead times, and other key charac-

teristics. There may be many PODs of each type.

Further details regarding each location's operations are described below as we explain how to use ESCOE.

B.2.1 Glossary of Important Terms

1. Arrival Rate - the expected hourly rate at which patients arrive to a POD (this can vary throughout each day).
2. Forward Deployed Stockpile (FDS) - regional federally controlled stockpiles that rapidly provides supplies to RSSs in its region when an emergency is declared.
3. Lead Time - the number of time periods required for medical supplies to be made available at a receiving location following the decision to send them. Lead times must be expressed in an integer number of time periods.
4. Patient Demand - the patients who arrive to a POD in a particular time period plus the patients who arrived during previous time periods but have not yet been served.
5. Point of Dispensing (POD) - location where patients go to receive appropriate medical countermeasures.
6. POD Type - a description of a POD that includes its lead times, unloading rates, arrival rates and service rates over time. The simulation may include many PODs of each POD Type.
7. POD Type Amount - the number of PODs of a particular POD Type within the network. This is an input on the "Construct the Network" page.

8. Queue Length - the total number of people waiting to be served at a given POD at the end of a time period.
9. Receiving, Storing, and Staging warehouse (RSS) - a warehouse that serves a particular state or city by receiving medical supplies from federal stockpiles during an emergency response operation and distributing these supplies to PODs.
10. Simulation Replication - one repetition of the simulated prophylaxis campaign.
11. Strategic National Stockpile (SNS) - federally owned stockpile of medical countermeasures for responding to a variety of emergencies. When an emergency is declared, the SNS rapidly distributes inventory to the affected regions' RSSs.

B.3 Working with ESCOE

To use ESCOE to model a particular supply chain system, a large number of user inputs are required. These data are entered in a series of worksheets. At the beginning of each section below we provide a checklist that identifies what you will need to complete the relevant steps of the model.

B.3.1 Getting Started

You will need:

1. *Microsoft Excel and Access, versions 2007 or later*

2. The files “SNS-CapacitiesModel.mdb” and “ESCOE.xlsm”

Copy both files, “SNS-CapacitiesModel.mdb” and “ESCOE.xlsm,” from the CD and store them in the same directory folder; otherwise ESCOE will not run. No other files should be placed in the ESCOE folder because they may cause errors to arise while running the program. If you try to run the simulation and get an error that stops the simulation from completing, you should go to the directory where these folders are stored and delete all files except these two.

To get started, open the Excel file “ESCOE.xlsm.” A “Security Warning” message box may prompt you to set appropriate security settings; if this happens, select the button to “Enable Macros.” Once the file is open, go to Developer → Visual Basic. From the Visual Basic Editor window, go to Tools → References. Find the following references and mark the corresponding checkboxes.

1. Visual Basic For Applications
2. Microsoft Excel 12.0 Object Library
3. OLE Automation
4. Microsoft Office 12.0 Object Library
5. Microsoft ActiveX Data Objects 6.0 Library
6. Microsoft Forms 2.0 Object Library

Then click “Okay” and return to the Excel - ESCOE window. You should only have to do this the first time you run ESCOE; the References should remain selected after that. However, if the ESCOE ever fails and you cannot understand why, make sure that these references are still selected.

Now you are ready to begin entering information in ESCOE. In the Excel - ESCOE window click on the ESCOE tab at the bottom of the page if it is not already selected.

B.3.2 Selecting an EXCOE Case

The file opens to a cover page. If you cannot see the “Begin Model” button in the lower right-hand corner of the screen, zoom out until the button becomes visible. Click the Begin Model button, and then the Case Selection page will appear, as shown in Figure B.2. A “case” is a complete set of ESCOE inputs. This page lets you choose to load the inputs from a previously defined case or to create a new case.


The image shows a software window titled "Case Selection" with a dark red background. The title is in large white font. Below the title, there is white text that reads: "Do you want to examine an existing case or create a new case? By selecting 'Create New', you will be routed to the menu page. By selecting 'Existing Case', you will be prompted to select the case by name." At the bottom of the window, there are two buttons. The left button is labeled "Create New" and has an unselected radio button next to it. The right button is labeled "Existing Case" and has a selected radio button next to it.

Figure B.2: Case Selection screen shot.

If you select the Create New option, a message box will pop up to ask if you would like to continue. If you answer Yes, you are taken to the ESCOE Menu page. If you answer No, then the program remains on the Case Selection page and awaits your next decision.

If you select the Existing Case button, a user form will pop up to show a list

of all cases currently in the database; this is shown in Figure B.3. Adding new cases to this list is easy; we will explain how to do this later.

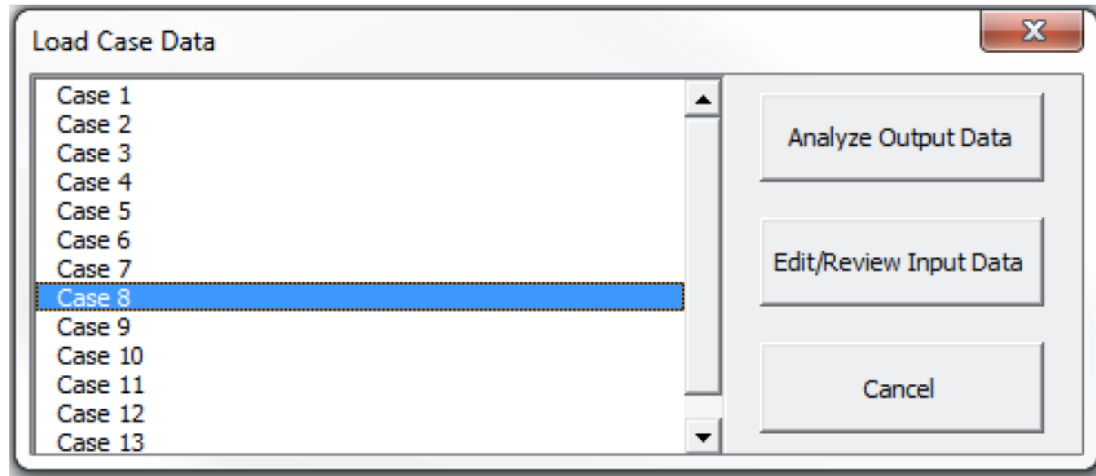


Figure B.3: Load Case Data screen shot.

After selecting your desired case, you can choose to Analyze Output Data or Edit/Review Input Data. Both buttons will load the inputs of the case that you chose. Clicking Edit/Review Input Data will bring you to the ESCOE Menu page where you can go to different sheets and edit the input values. Clicking the Analyze Output Data button automatically runs the simulation after loading the inputs. Be aware that some cases may take a long time to run, so if you are not familiar with a particular case, it is a good idea to choose the Edit/Review Input Data option so that you can check the case inputs before running the simulation. However, no matter which of these two options you select, you will be able to modify the inputs and re-run the simulation.

B.3.3 ESCOE Menu and Step 1

You will need to define:

- 1. The start time of the period that you wish to simulate;*
- 2. The time period length that you want to use in your model; and*
- 3. The total length of time that you want to simulate.*

The ESCOE Menu, displayed in Figure B.4, shows all of steps that are necessary to describe the model inputs, run the simulation, and explore the output. To avoid errors, it is important to work through steps 1 - 9 in order, since the information entered in many steps depends on inputs from previous steps.

Step 1 is completed on the Menu page. First, you may enter the time at which you want to start your simulation. This time should be the earliest time at which any location in your distribution network will begin operating, but not all of the locations in the network will be required to begin operating at that time; we discuss each location type in greater detail later. Next, you give the length of each time period and the total number of time periods that you want to simulate. See the Model Assumptions section for a more detailed discussion of time periods. The end time of your simulation is displayed below the dark grey area in which you enter these values,

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

To complete steps 2 through 9, you must click on their respective buttons and fill in the information on the worksheets that appear. You can return to earlier

ESCOE Menu

In order to run the program, follow the steps as shown below. The program may not work if executed in the incorrect order.

Step 1	Simulation Start Time:	8:00 AM
	Period Length (Hours):	2
	Number of Periods:	24
	Simulation End Time: Day: Time:	3 8:00 AM
Step 2	Construct the Network	Network
Step 3	Describe the Lead Times	Lead Times
Step 4	Describe the Inventory	Inventory
Step 5	Describe the SNS	SNS
Step 6	Describe the FDSs	FDSs
Step 7	Describe the RSSs	RSSs
Step 8	Describe the POD Types	POD Types
Step 9	Describe the Simulation Experiment	Simulation
Step 10	Run the simulation	Run Simulation
Step 11	View the Results. The results are displayed in both tabular and graph form.	Output Tables
		Output Graphs

Other Options

Select an existing case.	Existing Case
Manage case list.	Case List
Go to Cover Page to Start Over	Start Over

Figure B.4: ESCOE Menu screen shot.

steps to check your entries at any time, but if you modify any inputs, you need to work through all subsequent steps in order.

Step 10 runs the simulation. Step 11 takes you to various sheets that display the simulation output. After either of these steps, you can return to the earlier steps to check or change your inputs, but as before, if you make any changes, you must complete all of the higher numbered steps in order.

In the Other Options section below, you can work with the database file or to restart the program. If you click the Existing Case button, you will be prompted with the same user form as on the Case Selection page. You can then select a different case to study.

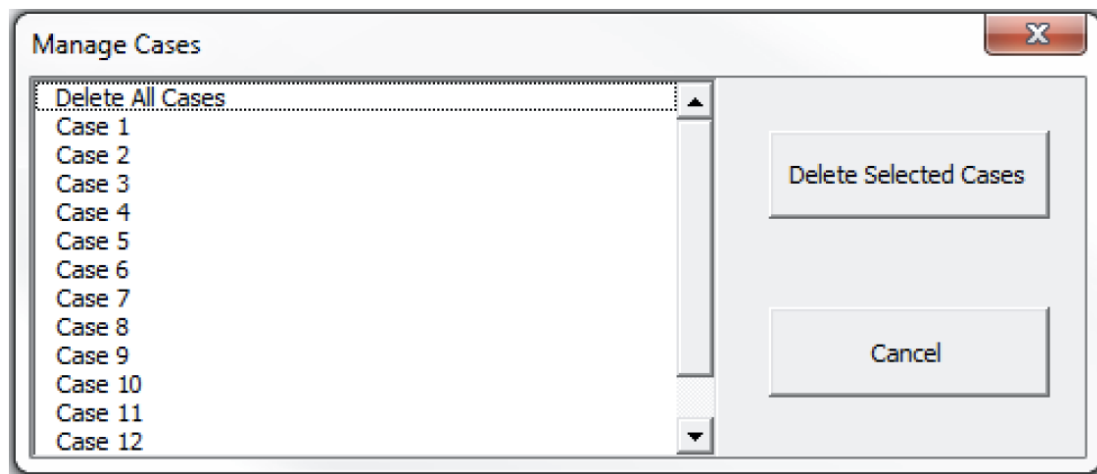


Figure B.5: Manage cases screen shot.

If you click the Case List button, you will be prompted by a user form, shown in Figure B.5, that is very similar to the one shown when Existing Case is clicked. The difference is that this button allows you to delete cases. This allows you to manage the size of the database and to clean out old data that are no longer of interest. The Start Over button will bring you back to the cover page and will

allow you to start the program from scratch.

B.3.4 Step 2: Constructing the Network


You will need to define:

- 1. The number of FDSs in your network;*
- 2. The number of RSSs in your network;*
- 3. The number of POD Types in your network;*
- 4. The number of PODs of each POD Type that are in your network;*
- 5. Which FDSs serve which RSSs; and*
- 6. The number of PODs of each POD Type that are served by each RSS.*

When you click the Network button on the ESCOE Menu, the Construct the Network worksheet, shown in Figure B.6, will appear. On this sheet you must complete Steps A, B, C and D in order. In Step A, specify the number of FDSs that you want to include in your network. Then click on the FDS names button. The FDS Names Table will appear on the right side of the screen. Fill out the table by entering a name for each FDS.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

In Step B, specify the number of RSSs in your network. Then click on the RSS names button. The RSS Names Table will appear on the right side of the screen. Fill out the table, entering a name for each RSS.



Cornell University

Step 2: Construct the Network

Main Menu

Construct Network

Define Lead Times

Describe Inventory

Describe SNS

Describe FDSs

Describe RSSs

Describe POD Types

Step A

Enter the number of FDSs (maximum of 15).

2

FDS names

Step B

Enter the number of RSSs (maximum of 15).

2

RSS names

Step C

Enter the number of POD Types (maximum of 10).

3

POD Types

Step D

Input the Network Relationships in the tables below.

Tables

Step E

Return to Main Page.

Return

FDS Names Table

FDS	1	2
Name	A	B

RSS Names Table

RSS	1	2
Name	C	D

POD Type Table

POD Types	1	2	3
Name	E	F	G
Amount	10	10	10

Input a 1 in the FDS/RSS Relationships Table if the FDS serves the RSS, and input a 0 in the FDS/RSS Relationships Table if the FDS does not serve that RSS.

FDS/RSS Relationships Table

		RSS	
		C	D
FDS	A	1	0
	B	0	1
	SNS	1	1

Figure B.6: Step 2 screen shot.

In Step C, specify the number of POD Types that you want to model. Recall that all pods of the same “type” must share the same parameters, including expected patient demand patterns and service rates. Then click on the POD Types button. The POD Type Names Table will appear on the right side of the screen. Input a name for each POD Type and the number of PODs of that type in your network.

In Step D, click on the Tables button. If there is at least one FDS in your network, the FDS/RSS Relationships Table and the RSS/POD Relationships Table will both appear at the bottom of the screen; if there are no FDSs in your network, only the latter table will appear. Fill in the FDS/RSS Relationships Table first. For each FDS, enter 1’s for in the columns for RSSs served by that FDS and 0’s for the RSSs not served by that FDS. An FDS may serve any number of RSSs, but each RSS may be served by at most one FDS. Also, remember that all

RSSs are served by the SNS, which is indicated by the row of grey cells at the bottom of the table. Continue down to the RSS/POD Relationship Table, which is displayed below in Figure B.7.

Describe Simulation

Run Simulation

View Output Tables

View Output Graphs

Input the number of each POD type served by each RSS in the RSS/POD Type Relationships Table. If the RSS does not serve that POD type input a 0.

RSS/POD Type Relationships Table

	POD Types		
	E	F	G
RSS C	8	5	0
D	2	5	10

Figure B.7: RSS/POD Type Relationship Table Shot.

In each cell of the RSS/POD Relationship Table, indicate how many PODs of each POD Type are served by each RSS. For example, if RSS A serves 5 PODs of POD Type B, you would enter a 5 in the cell corresponding to RSS A, POD Type B. You must make sure that the total numbers of each POD type correspond to the numbers entered in the POD Type Names Table. That is, if you declared that the total number of PODs of type E would be 10, make sure that the numbers in the POD type E column of the RSS/POD Relationship Table add up to 10.

Once Steps A - D are complete, proceed to Define the Lead Times. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step E and then clicking the highlighted Lead Times button on the menu page. The other option is to click the Define Lead Times button on the left side of the screen.

B.3.5 Step 3: Define Lead Times


You will need to define:

- 1. The Lead Time from the SNS to each RSS.*
- 2. The Lead Time from each FDS to each RSS that it serves.*
- 3. The Lead Time from each RSS to each POD Type that it serves.*

In the third input step, you will describe lead times to and from each location. Figure B.8 displays the Define Lead Times worksheet. In Step A, input the lead time (in time periods) from the external supplier to the SNS. In Step B, input the lead time (in time periods) from the SNS to each RSS in the first table. In Step C, enter the lead time (in periods) from each FDS to each RSS. Finally, in Step D, input the lead time (in periods) from each RSS to each POD Type. If an FDS does not serve a particular RSS, the cell corresponding to that FDS-RSS combination may be left blank, and similarly cells for RSSs that do not serve particular POD types may remain blank.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

Once Steps A - D are complete, proceed to Describe the Inventory. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step D and then clicking the highlighted Inventory button on the menu page. The other option is to click the Describe Inventory button on the left side of the screen.



Cornell University

Step 3: Define Lead Times

Main Menu

Construct Network

Define Lead Times

Describe Inventory

Describe SNS

Describe FDSs

Describe RSSs

Describe POD Types

Step A

Enter the lead times (in time periods) from the SNS to the RSSs in the table below.

Step B

Enter the lead times (in time periods) from the FDSs to the RSSs in the table below.

Step C

Enter the lead times (in time periods) from the RSSs to the PODs in the table below.

Step D

Return to Main Page

Return

SNS to RSS Lead Time Table:

	RSS	
	C	D
SNS	1	1

FDS to RSS Lead Time Table:

		RSS	
		C	D
FDS	A	1	1
	B	1	1

RSS to POD Types Lead Time Table:

		POD Types		
		E	F	G
RSS	C	1	1	1
	D	1	1	1


Figure B.8: Step 3 screen shot.

B.3.6 Step 4: Describe the Inventory

You will need to know:

1. The number of inventory types.(Only for future versions of ESCOE.)
2. Characteristics of each inventory type
3. The initial inventory at each location in the network for each inventory type.

In the fourth input step, you will describe the inventory that will be moved throughout your network. Figure B.9 displays a screen shot of the Step 4 worksheet. Future versions of ESCOE will allow you to enter the number of inventory types you wish to consider in Step A. However, the current version of the simulator only permits a single inventory type, so you should enter 1 for this step.



Cornell University

Step 4: Describe the Inventory

Main Menu

Construct Network

Define Lead Times

Describe Inventory

Describe SNS

Describe FDSs

Describe RSSs

Step A	Enter the number of Inventory Types (maximum of 5)	1
Step B	Enter the characteristics of each Inventory Type in the table below.	Inventory
Step C	Return to Main Page	Return

Inventory	1
Name	inv1
Units/Patient	1
Units/Case	1
Cases/Pallet	1
Initial Inventory (Pallets) For:	
SNS	100
FDS A	100
FDS B	100
RSS C	100
RSS D	100

Figure B.9: Step 4 screen shot.

If the inventory table showing below does not display the correct number of inventory types as well as the current FDS and RSS names, you must click the Inventory button in Step B to draw a new inventory type table. Clicking this button may clear the current table, so be sure to copy these values to another

location if you wish to save them. You can now fill out the inventory table. Enter a name for the single inventory type and the number units required to serve each patient. Next, you must enter the shipping characteristics for the inventory: the number of units in each case and the number of cases in each pallet.

The bottom part of the table allows you to define the initial inventory levels in terms of pallets at the SNS, FDSs, and RSSs. The PODs have no inventory at the beginning of the simulation.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

Once Steps A - B are complete, proceed to Describe the SNS. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step C and then clicking the highlighted SNS button on the menu page. The other option is to click the Describe SNS button on the left side of the screen.

B.3.7 Step 5: Describe the SNS

You will need to define:


- 1. The amount of inventory storage space at the SNS;*
- 2. The capacities of the trucks that carry inventory from the SNS to the RSSs;*
- 3. The start and end time for the three time intervals;*

4. *The rate at which trucks departing for RSSs can be loaded; and*
5. *The number of trucks available to send to RSSs in each period.*

In the fifth input step, you will describe the workings of the SNS. Figure B.10 displays a screen shot of the SNS sheet. Step A allows you to limit the amount of space available for inventory storage at the SNS; this is measured in terms of pallets of inventory. In Step B you can specify the capacity of the trucks that carry inventory from the SNS to the RSSs; the model assumes that all trucks used for SNS-to-RSS travel are the same size throughout the simulation. However, ESCOE allows the rate at which trucks can be loaded and the numbers of trucks available to change over the course of the simulation. The worksheet allows three different response phases: Initial, Intermediate, and Final. The truck loading rates and numbers of trucks available can vary between these phases, but must remain constant within a single phase.

For each phase, you must enter the time periods in which the phase starts and ends. ESCOE will show the time corresponding to these periods in the light grey boxes next to each start and end time. Each phase must be at least one time period long, and it starts at the beginning of some period, and finishes at the end of the same or some subsequent time period. For example, in Figure B.10, the Initial time interval lasts from the beginning of period 1 until the end of period 2. The Intermediate interval must start in period 3 or later; if the Intermediate interval does not start at the beginning of period 3, the SNS is assumed to be closed during the gap between the end of the Initial interval and the start of the Intermediate interval.

The outbound truck loading rate is an upper bound on the number of pallets of inventory that may be loaded onto trucks leaving the SNS in each hour.



Cornell University

Step 5: Describe the SNS

Main Menu

Step A	Enter the amount of space available for storage (in pallets).	9
Step B	Enter the capacity of the trucks that travel from the SNS to the RSSs (in pallets).	10
Step C	Complete the Initial Time Interval Table below.	
Step D	Complete the Intermediate Time Interval Table below.	
Step E	Complete the Final Time Interval Table below.	
Step F	Return to Menu Page	Return

Construct Network

C. Initial Time Interval Table		
Initial Interval START time (in periods)	1	Day: 1 Time: 8:00 AM
Initial Interval END time (in periods)	2	Day: 1 Time: 12:00 PM
Loading rate of trucks leaving for RSSs (pallets/hour)	50000	
Number of trucks available (in each period)	100	

Define Lead Times

D. Intermediate Time Interval Table		
Initial Interval START time (in periods)	3	Day: 1 Time: 12:00 PM
Initial Interval END time (in periods)	10	Day: 1 Time: 4:00 AM
Loading rate of trucks leaving for RSSs (pallets/hour)	50000	
Number of trucks available (in each period)	100	

Describe Inventory

E. Final Time Interval Table		
Initial Interval START time (in periods)	11	Day: 1 Time: 4:00 AM
Initial Interval END time (in periods)	12	Day: 2 Time: 8:00 AM

Describe SNS

F. Final Time Interval Table		
Initial Interval START time (in periods)	13	Day: 2 Time: 8:00 AM
Initial Interval END time (in periods)	14	Day: 2 Time: 12:00 PM

Describe FDSs

G. Final Time Interval Table		
Initial Interval START time (in periods)	15	Day: 2 Time: 12:00 PM
Initial Interval END time (in periods)	16	Day: 2 Time: 4:00 AM

Describe RSSs

H. Final Time Interval Table		
Initial Interval START time (in periods)	17	Day: 2 Time: 4:00 AM
Initial Interval END time (in periods)	18	Day: 2 Time: 8:00 AM

Describe POD Types

I. Final Time Interval Table		
Initial Interval START time (in periods)	19	Day: 2 Time: 8:00 AM
Initial Interval END time (in periods)	20	Day: 2 Time: 12:00 PM

Describe Simulation

J. Final Time Interval Table		
Initial Interval START time (in periods)	21	Day: 2 Time: 12:00 PM
Initial Interval END time (in periods)	22	Day: 2 Time: 4:00 AM

Run Simulation

K. Final Time Interval Table		
Initial Interval START time (in periods)	23	Day: 2 Time: 4:00 AM
Initial Interval END time (in periods)	24	Day: 2 Time: 8:00 AM

Figure B.10: Step 5 screen shot.

When modeling a real system, this value would be determined by the number of SNS workers and the availability of truck loading facilities and equipment. Trucks are not modeled individually by ESCOE, but it assumed that a limited number of vehicles are available to deliver inventory to RSSs in each period. The total amount of inventory that may be sent out from the SNS to the RSSs is constrained by number and capacity of the outbound trucks, as well as the loading rates.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

Once Steps A - E are complete, proceed to Describe the FDSs. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step F and then clicking the highlighted FDSs button on the menu page. The other option is to click the Describe FDSs button on the left side of the screen.


B.3.8 Step 6: Describe the FDSs

You will need to define:

- 1. The amount of inventory storage space at each FDS;*
- 2. The capacities of the trucks that carry inventory from the FDSs to the RSSs;*
- 3. The rate at which trucks departing for RSSs can be loaded; and*
- 4. The number of trucks available to send to RSSs in each period.*

In the sixth input step, you will describe the workings of the FDSs. Figure B.11 displays a screen shot of the FDS sheet. You must describe each FDS individually, starting with the first one. Step A requires you to set the amount of space available for inventory storage (in pallets) at the FDS. Step B requires you to specify the capacity of the trucks that move the inventory from the FDSs to the RSSs. In Step C, you can enter the number of trucks that are available to be sent out each hour, and in Step D you can enter the maximum rate at which

pallets of inventory can be loaded onto these trucks. Notice that, unlike at the SNS, ESCOE assumes that all of these values remain constant for the entire simulation; this assumption is made because FDSs are intended to operate only for a short time at the beginning of the emergency response before initial shipments arrive from the SNS.



Cornell University

Step 6: Describe the FDSs

Choose the FDS: A Next FDS

Step A	Enter the amount of space available for storage (in pallets).	50000
Step B	Enter the capacity of the trucks that travel from this FDS to the RSSs (in pallets).	20000
Step C	Enter the rate at which trucks can be loaded (pallets/hour).	100
Step D	Enter the number of trucks available in each period.	50000
Step E	Return to Menu Page Return	

Main Menu

Construct Network

Define Lead Times

Describe Inventory

Describe SNS

Figure B.11: Step 6 screen shot.

Once you have completed Steps A through D for the first FDS, you can click on the Next FDS button. Continue this process until you have entered the information for each FDS in your network.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

Once Steps A - D are complete, proceed to Describe the RSSs. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking

the Return button in Step E and then clicking the highlighted RSSs button on the menu page. The other option is to click the Describe RSSs button on the left side of the screen.


B.3.9 Step 7: Describe the RSSs

You will need to define:

- 1. The amount of inventory storage space at each RSS;*
- 2. The capacities of the trucks that carry inventory from the RSSs to the PODs;*
- 3. The start and end time for the three time intervals;*
- 4. The rate at which trucks arriving from the SNS or FDS can be unloaded;*
- 5. The rate at which trucks departing for PODs can be loaded; and*
- 6. The number of trucks available to send to PODs in each period.*

In the seventh input step, you will describe the workings of the RSSs. Figure B.12 displays a screen shot of the RSS sheet.

You will describe each RSS individually, starting with the first one. Step A requires you to set the amount of space available for inventory storage at the RSS. Step B requires you to specify the capacity of the trucks that carry inventory to the PODs. Similar to the SNS, the simulation time at the RSSs is divided into Initial, Intermediate, and Final time intervals. For each time interval, indicate periods in which the interval starts and ends. The times corresponding to these periods are shown in the light grey boxes next to each start and end time entry.



Cornell University

Step 7: Describe the RSSs

Choose the RSS: C Next RSS

Step A	Enter the amount of space available for storage (in pallets).	1000000
Step B	Enter the capacity of the trucks that travel from this RSS to the PODs (in pallets).	20000
Step C	Complete the Initial Time Interval Table below.	
Step D	Complete the Intermediate Time Interval Table below.	
Step E	Complete the Final Time Interval Table below.	
Step F	Return to Menu Page	Return

C. Initial Time Interval Table

Initial Interval START time (in periods)	1	Day: 1 Time: 8:00 AM
Initial Interval END time (in periods)	2	Day: 1 Time: 12:00 PM
Unloading rate of trucks arriving from the SNS and FDSs (cases/hour)	50000	
Loading rate of trucks leaving for PODs (cases/hour)	50000	
Number of trucks available (in each period)	100	

D. Intermediate Interval Period Table

Initial Interval START time (in periods)	3	Day: 1 Time: 12:00 PM
Initial Interval END time (in periods)	18	Day: 2 Time: 8:00 PM
Unloading rate of trucks arriving from the SNS and FDSs (cases/hour)	50000	
Loading rate of trucks leaving for PODs (cases/hour)	50000	
Number of trucks available (in each period)	100	

Figure B.12: Step 7 screen shot.

You also must enter the rates at which trucks arriving from the SNS or FDSs can be unloaded, the rates at which trucks going to the PODs can be loaded, and the number of trucks available to send out to PODs in each time period. Once you have completed this information for all three time intervals, you can move on to the next RSS by clicking on the Next RSS button. Continue this process until you have entered the information for each RSS.

241

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

Once Steps A - E are complete, proceed to Describe the POD Types. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step F and then clicking the highlighted POD Types button on the menu page. The other option is to click the Describe POD Types button on the left side of the screen.


B.3.10 Step 8: Describe the POD Types

For each POD type, you will need to define:

- 1. The number of relevant time intervals;*
- 2. The start and end period for each time interval;*
- 3. The rates at which inventory can be unloaded from trucks;*
- 4. The average patient arrival rates over time; and*
- 5. The average patient service rates over time.*

In the eighth input step, you will describe the different POD Types. Figure B.13 displays a screen shot of the RSS sheet.

You must describe each POD type individually, starting with the first one. In Step A, you must declare the number of time intervals that you want to consider for this POD type. The intervals need not be the same length, but all of the POD type's parameters must remain constant during each interval. So, if you want to



Cornell University

Step 8: Describe the POD Types

Main Menu

Construct Network

Define Lead Times

Describe Inventory

Describe SNS

Describe FDSs

Describe RSSs

Describe POD Types

Describe Simulation

Run Simulation

View Output Tables

View Output Graphs

Choose the POD type: E Next POD Type

Step A

Enter the number of time intervals (maximum of 12) to be included in the table below.

12

Step B

Complete the table below by entering the Unloading Rate, Arrival Rate, Service Rate, Start Time, and End Time for each time interval.

Rates

Step C

Return to Menu Page

Return

Interval Number	Unloading Rate (units/hour)	Average Patient Arrival Rate (patients/hour)	Average Patient Service Rate (patients/hour)	Interval Start Time (in periods)		Interval End Time (in periods)	
1	0	0	0	1	Day 1 8:00 AM	1	Day 1 10:00 AM
2	5	200	200	2	Day 1 10:00 AM	4	Day 1 4:00 PM
3	10	400	500	5	Day 1 4:00 PM	7	Day 1 10:00 PM
4	10	800	500	8	Day 1 10:00 PM	8	Day 1 12:00 AM
5	10	600	500	9	Day 1 12:00 AM	9	Day 1 2:00 AM
6	10	100	500	10	Day 1 2:00 AM	12	Day 2 8:00 AM
7	0	0	0	13	Day 2 8:00 AM	13	Day 2 10:00 AM
8	10	400	700	14	Day 2 10:00 AM	16	Day 2 4:00 PM
9	10	600	700	17	Day 2 4:00 PM	19	Day 2 10:00 PM
10	10	800	700	20	Day 2 10:00 PM	20	Day 2 12:00 AM
11	10	600	700	21	Day 2 12:00 AM	21	Day 2 2:00 AM
12	10	600	700	22	Day 2 2:00 AM	24	Day 3 8:00 AM

Figure B.13: Step 8 screen shot.

simulate a POD type with few changes in inventory unloading rates, expected patient arrival rates, and expected patient service rates, you can use a small number of time intervals. If you want to model more dynamic rates, you will need to include more time intervals.

Once you have decided on the number of time intervals, click on the Rates button to generate a new table for the POD type if the currently displayed table does not include the correct number of time intervals. All current entries in the table will be deleted when you click this button, so be sure to copy these

numbers to another location if you do not want to lose them. Step B requires that you fill out this table for the current POD type. The right-most columns of the table allow you to enter the start and end periods for each time interval in the dark grey columns. The light grey columns show the actual start and end times of each interval. As with the time intervals for the SNS and RSSs, each interval starts at the beginning of a period and finishes at the end of the same or some later period. If there is a gap between two time intervals, all PODs of this POD type are assumed to be “closed” during this interim; no patients will arrive or be served during this time, but any unserved patients will continue waiting for service until the POD type re-opens.

Inventory is unloaded from trucks that arrive from the RSSs at a rate limited by the numbers given in each table. These unloading rates would be determined both by the number of POD staff and the availability of essential equipment and facilities. The numbers of patients who arrive at PODs of this POD type each hour are drawn from Poisson distributions with means equal to the values that you enter in the Average Patient Arrival Rate column. Similarly, the POD capacities for serving patients are drawn from Poisson distributions with means given by the values in the Average Patient Service Rate column. The actual number of patients served at a POD in each period is the minimum of the total patient demand, the service capacity, and the inventory available. Once you have completed this information for each time interval, you can move on to the next POD Type by clicking on the Next POD Type button. Continue this process until you have entered the information for each POD Type.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able

to click the buttons required to move on to the next steps.

Once Steps A - B are complete, proceed to Describe the Simulation Experiment. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step C and then clicking the highlighted Simulation button on the menu page. The other option is to click the Describe Simulation button on the left side of the screen.

B.3.11 Step 9: Describe the Simulation Experiment

You will need to define:

- 1. The number of simulation replications you wish to run;*
- 2. A random seed value; and*
- 3. A name for the current case you have created.*

In the ninth input step, you will describe the simulation parameters. Figure B.14 displays a screen shot of the Simulation sheet. For Step A, indicate the number of Simulation Replications that you wish to run. This value is the number of complete campaigns you would like to simulate. If, for example, you set your campaign to be 3 days long and you declare that you want to run 100 replications, ESCOE will simulate the 3 day campaign 100 times. Note that the run time is proportional to the simulation replications, so doubling the number of replications will double the total time required to run the simulation. If you are not sure how fast ESCOE will run on your computer, start off with only a few replications. If your computer is fairly new and fast, you should have no trouble running 10 to 20 or more replications of ESCOE at one time. The number of

simulation replications should be a positive integer; if you enter a non-integer value, it will be rounded.


 Cornell University	Step 9: Describe the Simulation Experiment												
<div>Main Menu</div> <div>Construct Network</div> <div>Define Lead Times</div> <div>Describe Inventory</div>	<table border="1"><tr><td>Step A</td><td>Enter the Number of Simulation Replications</td><td>1</td></tr><tr><td>Step B</td><td>Enter the Random Seed (Do not change unless you have reason to do so.)</td><td>18</td></tr><tr><td>Step C</td><td>Input Case Name</td><td>Case 1</td></tr><tr><td>Step D</td><td>Return to Menu Page</td><td>Return</td></tr></table>	Step A	Enter the Number of Simulation Replications	1	Step B	Enter the Random Seed (Do not change unless you have reason to do so.)	18	Step C	Input Case Name	Case 1	Step D	Return to Menu Page	Return
Step A	Enter the Number of Simulation Replications	1											
Step B	Enter the Random Seed (Do not change unless you have reason to do so.)	18											
Step C	Input Case Name	Case 1											
Step D	Return to Menu Page	Return											

Figure B.14: Step 9 screen shot.

In Step B, enter any Random Seed that you want to use. This number should be a positive integer. It is used by the Visual Basic random number generator to determine the random numbers used in the simulation. If you use the same random seed with exactly the same set of inputs, you will get the same output. However, if you use the same set of inputs but change the random seed, the simulation output will be different (although the average values will likely be quite similar) because different random numbers will be drawn for the patient arrivals and service capacities. If you want to compare two different sets of inputs and keep the comparison as “fair” as possible, then it is a good idea to use the same random seed for both simulations. This helps eliminate random error as a confounding factor in your comparison.

Step C lets you give your case a name; recall that this “case” is the set of all inputs that you have entered in the previous eight steps. Giving your case a name allows you to save these inputs and reopen this case in the future using

the Select Existing Case dialog box. If you save your case to a name that already exists in the database, the old version of that case will be deleted and the current case will be saved to that case name.

When you are done entering information in each cell, be sure to press the Enter key. If you are in information entry mode for any cell, you will not be able to click the buttons required to move on to the next steps.

Once Steps A - C are complete, proceed to Run the Simulation. There are two ways to do this. One option is to return to the ESCOE Menu page by clicking the Return button in Step D and then clicking the highlighted Run Simulation button on the menu page. The other option is to click the Run Simulation button on the left side of the screen.

B.3.12 Step 10: Run the Simulation

The simulation run time is proportional to the number of simulation replications, the number of time periods, the number of locations in the distribution network and the patient arrival rates. Large simulations may easily take up to 5 or 10 minutes to complete. If you think that the run time is too long, you can stop the simulation by hitting the “Esc” key. Then click End in the error dialog box that pops up. If you do this, you must go to the directory where you have saved ESCOE and delete any new files that have been generated; there may be one called “Simulation_Output.txt.” If you do not delete this file, ESCOE may not run in the future. You can then modify your simulation by decreasing one or more of the parameters mentioned above to decrease the running time.

You will know that the simulation is complete when the ESCOE Menu page is displayed and the font on the output buttons for Step 11 is shown in red.

B.3.13 Step 11: View the Results

ESCOE produces a variety of output tables and graphs that you can use to analyze the performance of your simulated network. From the ESCOE menu, click on either the Output Tables button or Output Graphs button to begin viewing the results of your simulation. You can easily move between the two using the buttons on the left side panel or by returning to the ESCOE menu.

Clicking the Output Tables button will bring you to the Output Tables worksheet, as displayed below in Figure B.15. There are three drop down boxes that allow you to choose the type of data you would like to display. In the first drop-down box, indicate which data you are interested in viewing. You can choose from “Patient Queue at End of Period,” “Patient Demand in Period,” “Number of Patients Served in a Period,” “Number of Patients Arrived in a Period” or “End of Period Inventories.” Once you select which data you want to view, the valid location options will appear in the second drop down box. End of period inventories can be viewed at every location in the network, but the other choices only apply to PODs. After you indicate the location in which you are interested, you can select a single time period or “All Time Periods.” Once you have made a selection for each one of the three drop-down menus, click on the View Table button. The mean, standard deviation, minimum, and maximum values of your chosen inputs will then appear in the table at the bottom of your screen.

Clicking on the Output Graphs button on either the left panel from the Out-

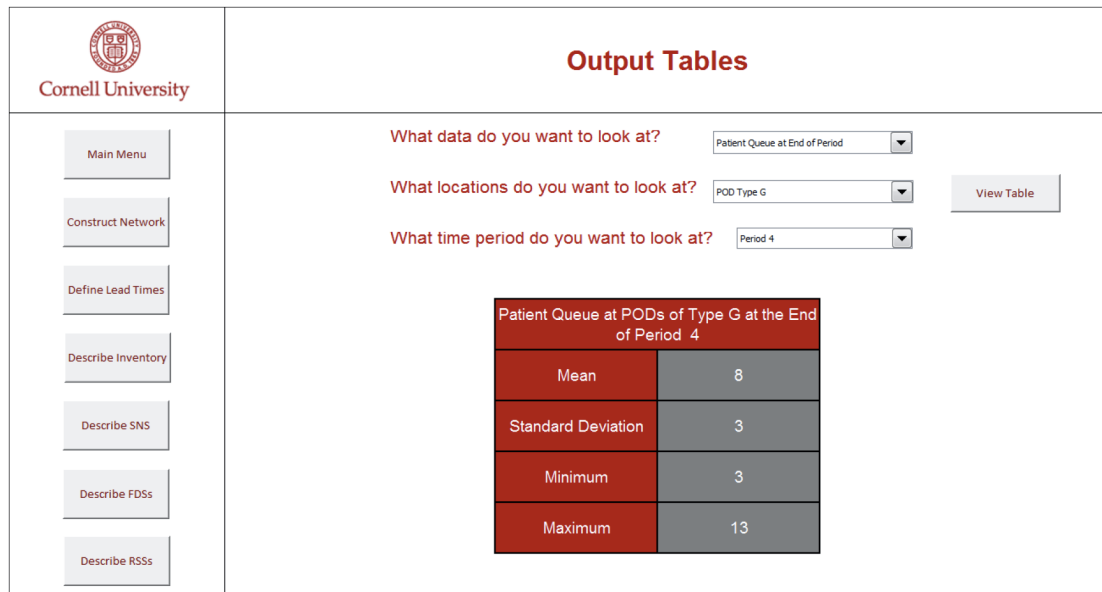


Figure B.15: Output tables screen shot.

put Tables page or on the ESCOE Menu page brings you to the Output Graphs page. A screen shot of this page is displayed below in Figure B.16.

As on the Output Tables page, there are three drop down boxes that allow you to choose the type of data you would display. These drop down boxes are identical to those on the previous page. Once you have made your selection in all three boxes, click on the View Graphs button. Two graphs are displayed at a time. The first graph shows the data for the particular location over all time periods. The second graph shows a histogram of the values for the particular time period you selected. An example of the graphs is displayed in Figure B.17.

The data shown in these plots are drawn from the Access database. In order to plot the data in different ways or view the raw numbers, you must explore the database directly. This is discussed in the next section.

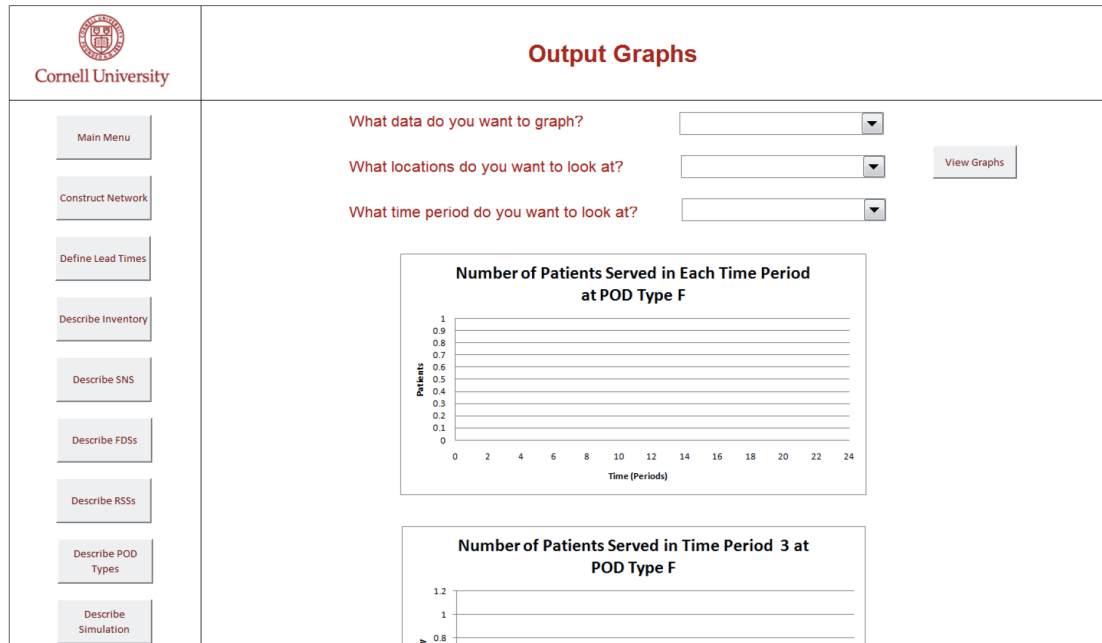


Figure B.16: Screen shot of output graphs page, before data to graph has been selected.

B.4 Navigating the Access Database

When you click the Run button to start your simulation, ESCOE immediately saves all input data to the database, called “SNS-CapacitiesModel.mdb” . The data is stored in the tables titled “Case List,” “fds_rss,” “fdsSimTable,” “fdsTable,” “fdsTimeTable,” “podSimTable,” “podTable,” “podTypeTable,” “podTypeTimeTable,” “rss_pod,” “rssSimTable,” “rssTable,” “rssTimeTable,” “SingleInputs,” “snsSimTable,” “snsTable” and “snsTimeTable.”

The “CaseList” table stores the case name specified on the “Describe the Simulation Experiment” page. This table also assigns the unique CaseListID to each CaseName. Some tables refer to a Case by its CaseName and others refer to its CaseListID.

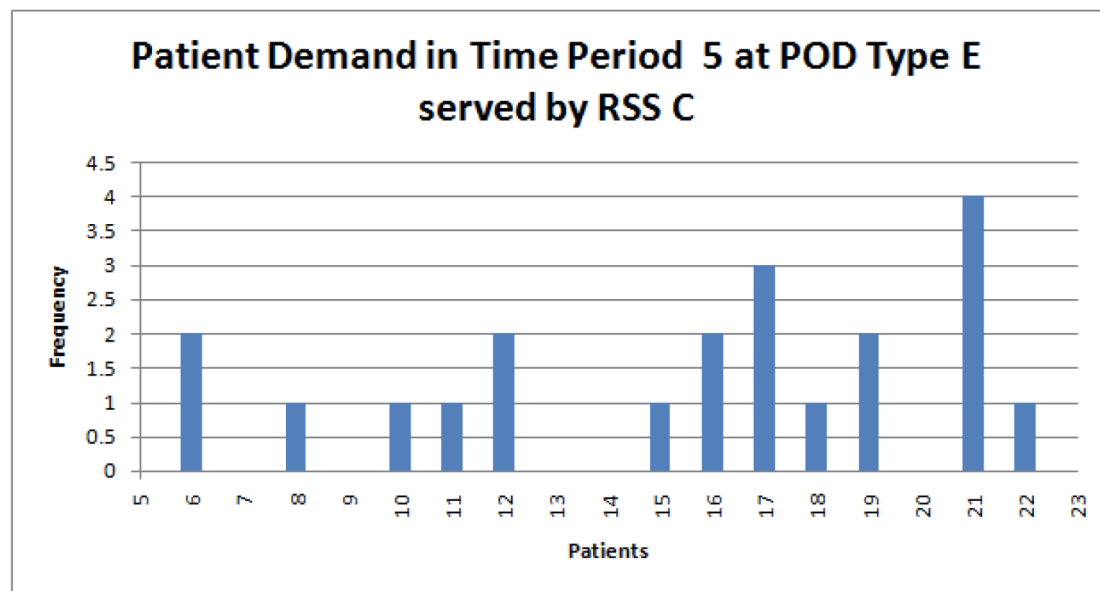
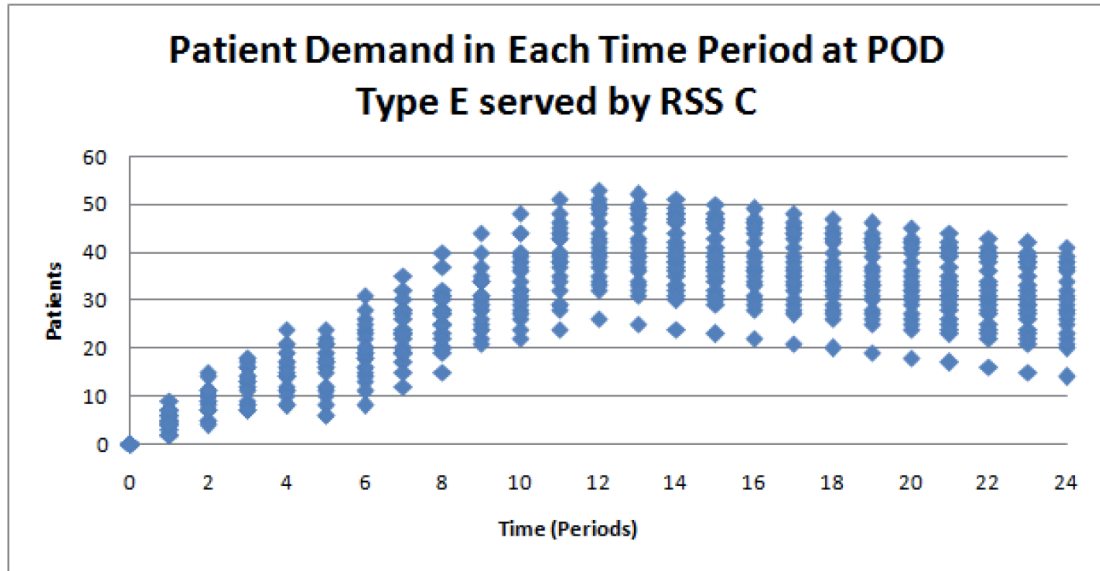


Figure B.17: Screen shot of output graphs of an patient demand at PODs of a particular type that are served by one of the RSSs.

The “fds_rss” table stores which FDSs serve which RSSs, according to their ID’s. These relationships were specified on the “Construct the Network” page.

The “fdsSimTable” table stores the data about the FDSs that varies over time

and simulation repetition. It stores the CaseName, FDS ID, simulation number, time period, how much inventory is sent out and the amount of inventory left at the FDS at the end of the time period.

The “fdsTable” table stores information about each FDS in each saved case, including its ID, its name, and its maximum capacity. The FdsID is generated by the simulator. The name of each FDS is acquired on the “Construct the Network” page and the maximum capacity is an input on the “Describe the FDSs” page.

The “fdsTimeTable” table stores whether each FDS is open during each time period. It lists each fds by its case and ID. For each time period there is a box that is checked if the FDS is open during that time period.

The “podSimTable” table stores the data about each POD that varies over time and simulation repetition. It stores the CaseName, POD ID, simulation number, time period, the number of patients who arrived, the number of patients served, the number of patients unserved and the patients demand. The “Patient Demand” is the number of patients who arrived in this time period and the number of patients unserved in the previous time period. It also stores information about inventory, including how much inventory is received by the POD, how much inventory is used on patients, and the amount of inventory left at the POD at the end of the time period. The “podTable” table stores the PODs ID and its Type ID. Each of these ID’s is generated by the simulator.

The “podTypeTable” table stores each POD Type by its ID and its name. The ID is generated by the simulator and its name was an input on the “Construct the Network” page. The “podTypeTimeTable” table stores the average patient

arrival rate, the average patient service rate for each POD Type during each time period. It also stores whether or not each POD Type is open during each time period.

The “rss_pod” table stores which RSSs serve which PODs, according to their ID’s. These relationships were specified on the “Construct the Network” page. The “rssSimTable” table stores the data about the RSSs that varies over time and simulation repetition. It stores the CaseName, RSS ID, simulation number, time period, how much inventory is received by the RSS, how much inventory is sent out from the RSS, and the amount of inventory left at the RSS at the end of the time period.

The “rssTable” table stores the RSS’s ID, its name and its maximum capacity. The ID is generated by the simulator. The RSS’s name is specified on the “Construct the Network” page, and its maximum capacity is determined from the “Describe the RSSs” page.

The “rssTimeTable” table stores whether or not each RSS is open during each time period. It lists each RSS by its case and ID. For each time period there is a box that is checked if the RSS is open during that time period.

The “SingleInputs” table contains all input values that are entered through ESCOE. The values are stored based on the CaseListID, the variable name (varName), and the value the variable holds in the program (varValue). There is also a boolean value, “SingleInput,” in the table. “SingleInput” = ‘True’ if the input value is not entered through a chart within ESCOE, such as the number of simulation replications or the number of days in the prophylaxis campaign. These inputs are stored into “SingleInputs” each time you run the simulation.

The “snsSimTable” table stores the data about the SNS that varies over time and simulation repetition. It stores the CaseName, SNS ID, simulation number, time period, how much inventory is received by the SNS, how much inventory is sent out from the SNS, and the amount of inventory left at the SNS at the end of the time period.

The “snsTable” table stores the SNS’s ID, which is generated by the simulator, and its maximum capacity, which was specified on the “Describe the SNS” page. The “snsTimeTable” table stores whether or not the SNS is open during each time period. It lists each SNS by its case and ID. For each time period there is a box that is checked if the SNS is open during that time period

BIBLIOGRAPHY

- [Aaby *et al.*, 2006] Aaby, K., Abbey, R.L., Herrmann, J.W., Treadwell, M., Jordan, C.S., & Wood, K. 2006. Embracing Computer Modeling to Address Pandemic Influenza in the 21st Century. *Journal of Public Health Management Practice*, **12**(4), 365–372.
- [Baccam & Boechler, 2007] Baccam, P., & Boechler, M. 2007. Public Health Response to an Anthrax Attack: An Evaluation of Vaccination Policy Options. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, **5**(1), 26–34.
- [Bajardi *et al.*, 2011] Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V., & Vespignani, A. 2011. Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic. *PLoS ONE*, **6**(1), e16591.
- [Balcan *et al.*, 2009] Balcan, D., Colizza, V., Goncalves, B., Hu, H., Ramasco, J.J., & Vespignani, A. 2009. Multiscale Mobility Networks and the Spatial Spreading of Infectious Diseases. *PNAS*, **106**(51), 21484–21489.
- [Barr *et al.*, 2010] Barr, I.G, McCauley, J., Cox, N., Daniels, R., Engelhardt, O.G., Fukuda, K., Grohmann, G., Hay, A., Kelso, A., Klimov, A., Odagiri, T., Smith, D., Russell, C., Tashiro, M., Webby, R., Wood, J., Ye, Z., Zhang, W., & Writing Committee of the World Health Organization Consultation on Northern Hemisphere Influenza Vaccine Composition for 2009-2010. 2010. Epidemiological, Antigenic and Genetic Characteristics of Seasonal Influenza A(H1N1), A(H3N2) and B Influenza Viruses: Basis for the WHO Recommendation on the Composition of Influenza Vaccines for Use in the 2009-2010 Northern Hemisphere Season. *Vaccine*, **28**, 1156–1167.
- [Berman *et al.*, 2011] Berman, O., Gavius, A., & Huang, R. 2011. Location of Response Facilities: A Simultaneous Game Between State and Terrorist. *International Journal of Operational Research*, **10**(1), 102–120.
- [Birge & Louveaux, 1997] Birge, J.R., & Louveaux, F. 1997. *Introduction to Stochastic Programming*. Springer.
- [Braithwaite *et al.*, 2006] Braithwaite, R.S., Fridsma, D., & Roberts, M.S. 2006. The Cost-Effectiveness of Strategies to Reduce Mortality from an Intentional Release of Aerosolized Anthrax Spores. *Medical Decision Making*, **26**(1), 182–193.

- [Brandeau *et al.*, 2009] Brandeau, M.L., McCoy, J.H., Hupert, N., Holty, J.E., & Bravata, D.M. 2009. Recommendations for Modeling Disaster Responses in Public Health and Medicine: A Position Paper for the Society for Medical Decision Making. *American Journal of Disaster Medicine*, **29**(4), 438–460.
- [Bravata *et al.*, 2006] Bravata, D.M., Zaric, G.S., Holty, J.C., & Brandeau, M.L. 2006. Reducing Mortality from Anthrax Bioterrorism: Strategies for Stockpiling and Dispensing Medical and Pharmaceutical Supplies. *Biosecurity and Bioterrorism*, **4**(3), 244–262.
- [Brookmeyer *et al.*, 2003] Brookmeyer, R., Johnson, E., & Bollinger, R. 2003. Modeling the Optimum Duration of Antibiotic Prophylaxis in an Anthrax Outbreak. *PNAS*, **100**(17), 10129–10132.
- [Brookmeyer *et al.*, 2004] Brookmeyer, R., Johnson, E., & Bollinger, R. 2004. Public Health Vaccination Policies for Containing an Anthrax Outbreak. *Nature*, **432**, 901–904.
- [Brownstein *et al.*, 2006] Brownstein, J.S., Wolfe, C.J., & Mandl, K.D. 2006. Empirical Evidence for the Effect of Airline Travel on Inter-Regional Influenza Spread in the United States. *PLoS Medicine*, **3**(10), e401.
- [Buzacott & Shanthikumar, 1993] Buzacott, J.A., & Shanthikumar, J.G. 1993. *Stochastic Models of Manufacturing Systems*. New Jersey: Prentice-Hall.
- [Caggiano & Muckstadt, 2010] Caggiano, K.E., & Muckstadt, J.A. 2010. RASCAL. School of Operations Research and Information Engineering, Cornell University, Ithaca, NY.
- [Carrat *et al.*, 2010] Carrat, F., Pelat, C., Levy-Bruhl, D., Bonmarin, I., & Lapidus, N. 2010. Planning for the Next Influenza H1N1 Season: A Modelling Study. *BMC Infectious Diseases*, **10**(201).
- [CDC, 2004] CDC. 2004. *Cities Readiness Initiative Guidance, Appendix 3*. Centers for Disease Control and Prevention, Atlanta, GA. <http://www.bt.cdc.gov/planning/guidance05/pdf/appendix3.pdf> [Accessed September 2, 2011].
- [CDC, 2009] CDC. 2009. *Influenza-Like Illness (ILI) Reports*. Centers for Disease Control and Prevention, Atlanta, GA. <http://www.cdc.gov/flu/weekly/fluactivitysurv.htm> [Accessed December 15, 2011].

- [CDC, 2010] CDC. 2010. *Clinical Signs and Symptoms of Influenza*. Centers for Disease Control and Preparedness. <http://www.bt.cdc.gov/publications/2010phprep/pdf/2010phprep.pdf> [Accessed April 19, 2012].
- [CDC, 2011] CDC. 2011. *Strategic National Stockpile*. Centers for Disease Control and Prevention, Atlanta, GA. <http://www.bt.cdc.gov/stockpile/> [Accessed September 2, 2011].
- [CDC, 2012] CDC. 2012 (May). *Alternative Methods for Antiviral Dispensing and Distribution During an Influenza Pandemic Meeting*. Centers for Disease Control and Prevention, Atlanta, GA.
- [Chao *et al.*, 2011] Chao, D.L., Matrajt, L., Basta, N.E., Sugimoto, J.D., Dean, B., Bagwell, D.A., Ojulfstad, B., Halloran, M.E., & Longini, I.M. 2011. Planning for the Control of Pandemic Influenza A (H1N1) in Los Angeles County and the United States. *American Journal of Epidemiology*, **173**(10), 1121–1130.
- [Cieslak & Eitzen, 1999] Cieslak, T.J., & Eitzen, E.M. 1999. Clinical and epidemiologic principles of anthrax. *Emerging Infective Diseases*, **5**(4), 552–555.
- [Clark & Scarf, 1960] Clark, A.J., & Scarf, H. 1960. Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science*, **6**(4), 475–490.
- [Colizza *et al.*, 2007] Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.J., & Vespignani, A. 2007. Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLoS Medicine*, **4**(1), e13.
- [Craft *et al.*, 2005] Craft, D.L., Wein, L.M., & Wilkins, A.H. 2005. Analyzing Bioterror Response Logistics: the Case of Anthrax. *Management Science*, **51**(5), 679–694.
- [Diks & de Kok, 1998] Diks, E.B., & de Kok, A.G. 1998. Optimal Control of a Divergent Multi-Echelon Inventory System. *European Journal of Operations Research*, **111**(1), 75–97.
- [Diks & de Kok, 1999] Diks, E.B., & de Kok, A.G. 1999. Computational Results for the Control of a Divergent *N*-echelon inventory system. *International Journal of Production Economics*, **59**(1-3), 237–336.
- [Dogru *et al.*, 2004] Dogru, M.K., de Kok, A.G., & van Houtum, G.J. 2004. *Optimal Control of One-Warehouse Multi-Retailer Systems with Discrete Demand*. Beta

Working Paper WP 122. Department of Technology Management, Technische Universiteit Eindhoven.

- [Dogru *et al.*, 2005] Dogru, M.K., de Kok, A.G., & van Houtum, G.J. 2005. *A Numerical Study on the Effect of the Balance Assumption in One-Warehouse Multi-Retailer Inventory Systems*. Tech. rept. Department of Technology Management, Technische Universiteit Eindhoven.
- [Eppen & Schrage, 1981] Eppen, G., & Schrage, L. 1981. Centralized Ordering Policies in a Multi-Warehouse System with Lead Times and Random Demand. In: Schwartz, L.B. (ed), *Multi-level Production/Inventory Control Systems: Theory and Practice*.
- [Epstein *et al.*, 2007] Epstein, J.M., Goedecke, D.M., Yu, F., Morris, R.J., Wagener, D.K., & Bobashev, G.V. 2007. Controlling Pandemic Flu: The Value of International Air Travel Restrictions. *PLoS ONE*, **5**, e401.
- [Federgruen & Zipkin, 1984a] Federgruen, A., & Zipkin, P. 1984a. Allocation Policies and Cost Approximations for Multilocation Inventory Systems. *Naval Research Logistics*, **31**(1), 97–129.
- [Federgruen & Zipkin, 1984b] Federgruen, A., & Zipkin, P. 1984b. A Combined Vehicle Routing and Inventory Allocation Problem. *Operations Research*, **32**(5), 193–207.
- [Ferguson *et al.*, 2005] Ferguson, N.M., Cummings, D.A.T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D.S. 2005. Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia. *Nature*, **437**, 209–214.
- [Ferguson *et al.*, 2006] Ferguson, N.M., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C., & Burke, D.S. 2006. Strategies for Mitigating an Influenza Pandemic. *Nature*, **442**, 448–452.
- [Fiore *et al.*, 2011] Fiore, A.E., Fry, A., Shay, D., Gubareva, L., Bresee, J.S., & Uyeki, T.M. 2011. *Antiviral Agents for the Treatment and Chemoprophylaxis of Influenza: Recommendations of the Advisory Committee on Immunization Practices (ACIP)*. Tech. rept.
- [Fowler & Shafazand, 2011] Fowler, R.A., & Shafazand, S. 2011. Anthrax Bioterrorism: Prevention, Diagnosis and Management Strategies. *Journal of Bioterrorism and Defense*, **2**(2).

- [Fowler *et al.*, 2005] Fowler, R.A., Sanders, G.D., Bravata, D.M., Nouri, B., Gastwirth, J.M., Peterson, D., Broker, A.G., Garber, A.M., & Owens, D.K. 2005. Cost-Effectiveness of Defending against Bioterrorism: A Comparison of Vaccination and Antibiotic Prophylaxis against Anthrax. *Annals of Internal Medicine*, **142**(8), 601–611.
- [Fraser *et al.*, 2009] Fraser, C., Donnelly, C.A., Cauchemez, S., Hanage, W.P., Kerkhove, M.D. Van, Hollingsworth, T.D., Griffin, J., Baggaley, R.F., Jenkins, H.E., Lyons, E.J., Jombart, T., Hinsley, W.R., Grassly, N.C., Balloux, F., Ghani, A.C., Ferguson, N.M., Rambaut, A., Pybus, O.G., Lopez-Gatell, H., Alpuche-Aranda, C.M., Chapela, I.B., Guevara, E.P. Zavala D.M.E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., & Collaboration, The WHO Rapid Pandemic Assessment. 2009. Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. *Science*, **324**(19), 1557–1561.
- [Gerdil, 2003] Gerdil, C. 2003. The Annual Production Cycle for Influenza Vaccine. *Vaccine*, **21**, 1776–1779.
- [Germann *et al.*, 2006] Germann, T.C., Kadau, K., I.M. Longini, Jr, & Macken, C.A. 2006. Mitigation Strategies for Pandemic Influenza in the United States. *PNAS*, **103**(15), 5935–5940.
- [Glass *et al.*, 2006] Glass, R.J., Glass, L.M., Beyeler, W.E., & Min, H.J. 2006. Targeted Social Distancing Design for Pandemic Influenza. *Emerging Infectious Diseases*, **12**(11), 1671–1681.
- [Graves & Willems, 2008] Graves, S.C., & Willems, S.P. 2008. Strategic Inventory Placement in Supply Chains: Nonstationary Demand. *MSOM*, **10**(2), 278–287.
- [Halder *et al.*, 2010] Halder, N., Kelso, J.K., & Milne, G.J. 2010. Analysis of the Effectiveness of Interventions Used During the 2009 A/H1N1 Influenza Pandemic. *BMC Public Health*, **10**(168).
- [Halloran *et al.*, 2008] Halloran, M.E., Ferguson, N.M., Eubank, S., Longini, I.M., Cummings, D.A.T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T.C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C.A., Burke, D.S., & Cooley, P. 2008. Modeling Targeted Layered Containment of an Influenza Pandemic in the United States. *PNAS*, **105**(12), 4639–4644.
- [Henderson, 1999] Henderson, D.A. 1999. The Looming Threat of Bioterrorism. *Science*, **283**, 1279–1282.

- [HHS, 2005] HHS. 2005. *HHS Pandemic Influenza Plan Supplement 7: Antiviral Drug Distribution and Use*. Department of Health and Human Services.
- [HHS, 2009] HHS. 2009 (April). *Secretary SEbelius Takes Two Key Actions on Strategic National Stockpile [News Release]*. U.S. Department of Health and Human Services, Washington, DC. <http://www.hhs.gov/news/press/2009pres/04/20090430a.html> [Accessed April 20, 2011].
- [Hupert, 2011] Hupert, N. 2011 (August). *SNS POD Plan*. Personal Communication.
- [Hupert *et al.*, 2002] Hupert, N., Mushlin, A.I., & Callahan, M.A. 2002. Modeling the Public Health Response to Bioterrorism: Using Discrete Event Simulation to Design Antibiotic Distribution Centers. *Medical Decision Making*, **22**, S17–S25.
- [Hupert *et al.*, 2009] Hupert, N., Wattson, D., Cuomo, J., Hollingsworth, E., Neukermans, K., & Xiong, W. 2009. Predicting Hospital Surge after a Large-Scale Anthrax Attack: A Model-Based Analysis of CDC's Cities Readiness Initiative Prophylaxis Recommendations. *Medical Decision Making*, **29**, 424–437.
- [Inglesby, 2002] Inglesby, T. V. et al. 2002. Anthrax as a Biological Weapon, 2002: Updated Recommendations for Management. *Journal of the American Medical Association*, **287**(17), 2236–2252.
- [Koonin, 2012] Koonin, L. 2012 (January). *Estimated Number of Antivirals Needed for Treatment - Excel Spreadsheet*. Personal Communication.
- [Koonin *et al.*, 2011] Koonin, L.M., Beauvais, D.R., Shimabukuro, T., Wortley, P.M., Palmier, J.B., Stanley, T.R., Theofilos, J., & Merlin, T.L. 2011. CDC's 2009 H1N1 Vaccine Pharmacy Initiative in the United States: Implications for Future Public Health and Pharmacy Collaborations for Emergency Response. *Disaster Medicine and Public Health Preparedness*, **5**(4), 253–255.
- [Kunnumkal & Topaloglu, 2008] Kunnumkal, S., & Topaloglu, H. 2008. A Duality-Based Relaxation and Decomposition Approach for Inventory Distribution Systems. *Naval Research Logistics Quarterly*, **55**(7), 612–631.
- [Kunnumkal & Topaloglu, 2010] Kunnumkal, S., & Topaloglu, H. 2010. Linear Programming Based Decomposition Methods for Inventory Distribution Systems. *European Journal of Operational Research*, **forthcoming**.

- [Larson, 2007] Larson, R.C. 2007. Simple Models of Influenza Progression within a Heterogeneous Population. *Operations Research*, **55**(3), 399–412.
- [Lee *et al.*, 2010a] Lee, B.Y., Brown, S.T., Korch, G., Cooley, P.C., Zimmerman, R.K., Wheaton, W.D., Zimmer, S.M., Grefenstette, J.J., Bailey, R.R., Assi, T.M., & Burke, D.S. 2010a. A Computer Simulation of Vaccine Prioritization, Allocation, and Rationing During the 2009 H1N1 Influenza Pandemic. *Vaccine*, **28**(31), 4875–4879.
- [Lee *et al.*, n.d.] Lee, E.K., Smalley, H.K., Zhang, Y., Pietz, F., & Benecke, B. Facility Location and Multi-Modality Mass Dispensing Strategies and Emergency Response for Biodefense and Infectious Disease Outbreaks. *International Journal of Risk Assessment and Management*, **12**(2-4, pages =).
- [Lee *et al.*, 2006a] Lee, E.K., Maheshwary, S., Mason, J., & Glisson, W. 2006a. Decision Support System for Mass Dispensing of Medications for Infectious Disease Outbreaks and Bioterrorist Attacks. *Annals Operations Research*, **148**(1), 25–53.
- [Lee *et al.*, 2006b] Lee, E.K., Mason, J., & Glisson, W. 2006b. Large-Scale Dispensing for Emergency Response to Bioterrorism and Infectious Disease Outbreak. *Interfaces*, **36**(6), 591–607.
- [Lee *et al.*, 2010b] Lee, E.K., Chen, C.H., Pietz, F., & Benecke, B. 2010b. Disease Propagation Analysis and Mitigation Strategies for Effective Mass Dispensing. *AMIA 2010 Symposium Proceedings Page*, 427–431.
- [Lee, 2008] Lee, Y.M. 2008 (December). Analyzing Dispensing Plan for Emergency Medical Supplies in the Event of Bioterrorism. In: Mason, S.J., Hill, R.R., Monch, L., Rose, O., Jefferson, T., & Fowler, J.W. (eds), *Proceedings of the 2008 Winter Simulation Conference*.
- [Lien *et al.*, 2006] Lien, O., Maldin, B., Franco, C., & Gronvall, G.K. 2006. Getting Medicine to Millions: New Strategies for Mass Distribution. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, **4**(2), 176–182.
- [Lipsitch *et al.*, 2011] Lipsitch, M., Finelli, L., Heffernan, R.T., Leung, G.M., & Redd, S.C. 2011. Improving the Evidence Base for Decision Making During a Pandemic: The Example of 2009 Influenza A/H1N1. *Biosecurity and Bioterrorism*, **9**(2), 89–115.
- [Lu *et al.*, 2010] Lu, H.M., Zeng, D., & Chen, H. 2010. Prospective Infectious

- Disease Outbreak Detection Using Markov Switching Models. *IEEE Transactions on Knowledge and Data Engineering*, **22**(4), 565–577.
- [Medlock & Galvani, 2009] Medlock, J., & Galvani, A.P. 2009. Optimizing Influenza Vaccine Distribution. *Science*, **325**(5948), 1705–1708.
- [Milne *et al.*, 2008] Milne, G.J., Kelso, J.K., Kelly, H.A., Huband, S.T., & McVernon, J. 2008. A Small Community Model for the Transmission of Infectious Diseases: Comparison of School Closure as an Intervention in Individual-Based Models of an Influenza Pandemic. *PLoS ONE*, **3**(12), e4005.
- [Mitchell-Blackwood *et al.*, 2011] Mitchell-Blackwood, J., Gurian, P.L., & ODonnell, C. 2011. Finding Risk-Based Switchover Points for Response Decisions for Environmental Exposure to *Bacillus anthracis*. *Journal of Human and Ecological Risk Assessment*, **17**(2), 489–509.
- [Montjoy & Herrmann, 2010] Montjoy, A., & Herrmann, J.W. 2010. Adaptive Large Neighborhood Search for the Inventory Slack Routing Problem. In: Johnson, A., & Miller, J. (eds), *Proceedings of the 2010 Industrial Engineering Research Conference*.
- [Muckstadt & A., 2010] Muckstadt, J.A., & A., Sapra. 2010. *Principles of Inventory Management*. Springer.
- [Neale & Willems, 2009] Neale, J.J., & Willems, S.P. 2009. Managing Inventory in Supply Chains with Nonstationary Demand. *Interfaces*, **39**(5), 388–399.
- [Nelson *et al.*, 2008] Nelson, C., Chan, E.W., Chandra, A., Sorensen, P., Willis, H.H., Comanor, K., Park, H., Ricci, K.A., Caldaroni, L.B., Shea, M., Zambrano, J.A., & Hansell, L. 2008. *Recommended Infrastructure Standards for Mass Antibiotic Dispensing*. Tech. rept. RAND Corporation.
- [Nigmatulina & Larson, 2009] Nigmatulina, K.R., & Larson, R.C. 2009. Living with Influenza: Impacts of Government Imposed and Voluntarily Selected Interventions. *European Journal of Operational Research*, **195**, 613–627.
- [Nishiura *et al.*, 2010] Nishiura, H., Chowell, G., Safan, M., & Castillo-Chavez, C. 2010. Pros and Cons of Estimating the Reproduction Number from Early Epidemic Growth Rate of Influenza A (H1N1) 2009. *Theoretical Biology and Medical Modelling*, **7**(1).
- [Presanis *et al.*, 2009] Presanis, A.M., Angelis, D. De, Team, The New York City

- Swine Flu Investigation, Hagy, A., Reed, C., Riley, S., Cooper, B.S., Finelli, L., Biedrzycki, P., & Lipsitch, M. 2009. The Severity of Pandemic H1N1 Influenza in the United States, from April to July 2009: A Bayesian Analysis. *PLoS Medicine*, **6**(12), e1000207.
- [Riley, 2007] Riley, S. 2007. Large-Scale Spatial-Transmission Models of Infectious Disease. *Science*, **316**, 1298–1301.
- [Savachkin & Uribe, 2011] Savachkin, A., & Uribe, A. 2011. Dynamic Redistribution of Mitigation Resources During Influenza Pandemics. *Socio-Economic Planning Sciences*.
- [Schmitt *et al.*, 2007] Schmitt, B., Dobrez, D., Parada, J.P., Kyriacou, D.N., Golub, R.M., Sharma, R., & Bennett, C. 2007. Responding to a Small-scale Bioterrorist Anthrax Attack: Cost-effectiveness Analysis Comparing Preattack Vaccination with Postattack Antibiotic Treatment and Vaccination. *Archives of Internal Medicine*, **167**(7), 655–662.
- [Schuchat *et al.*, 2011] Schuchat, A., Bell, B.P., & Redd, S.C. 2011. The Science behind Preparing and Responding to Pandemic Influenza: The Lessons and Limits of Science. *Epidemiologic Perspectives & Innovations*, **52**(S1), S8–S12.
- [Shaman *et al.*, 2011] Shaman, J., Goldstein, E., & Lipsitch, M. 2011. Absolute Humidity and Pandemic Versus Epidemic Influenza. *American Journal of Epidemiology*, **173**(2), 127–135.
- [Shrestha *et al.*, 2011] Shrestha, S.S., Swerdlow, D.L., Borse, R.H., Prabhu, V.S., Finelli, L., Atkins, C.Y., Owusu-Edusei, K., Bell, B., Mead, P.S., Biggerstaff, M., Brammer, L., Davidson, H., Jernigan, D., Jhung, M.A., Kamimoto, L.A., Merlin, T.L., Nowell, M., Redd, S.C., Reed, C., Schuchat, A., & Meltzer, M.I. 2011. Estimating the Burden of 2009 Pandemic Influenza A (H1N1) in the United States (April 2009–April 2010). *Clinical Infectious Diseases*, **52**, S75–S82.
- [SK&A, 2011] SK&A. 2011 (April). *National Pharmacy Market Summary*. Irvine, California.
- [SNS, 2008] SNS. 2008. *Point of Dispensing (POD) Standards*. Coordinating office for Terrorism Preparedness and Emergency Response, Division of Strategic National Stockpile, Atlanta, GA. <http://health.mo.gov/emergencies/sns/pdf/12-PODStandards.pdf> [Accessed September 2, 2011].

- [Starr, 2012] Starr, D. 2012 (February). *New York City POD Modeling Meeting*. Personal Communication.
- [Teytelman & Larson, 2012] Teytelman, A., & Larson, R.C. 2012. Modeling Influenza Progression within a Continuous-Attribute Heterogeneous Population. *European Journal of Operational Research*, **220**, 238–250.
- [Tuite *et al.*, 2010] Tuite, A.R., Greer, A.L., Whelan, M., Winter, A.L., Lee, B., Yan, P., Wu, J., Moghadas, S., Buckridge, D., Pourbohloul, B., & Fisman, D.N. 2010. Estimated Epidemiologic Parameters and Morbidity Associated with Pandemic H1N1 Influenza. *Canadian Medical Association Journal*, **182**(2), 131–136.
- [USCB, 2012] USCB. 2012. *2012 Statistical Abstract: Resident Population By Sex and Age*. United States Census Bureau. <http://www.census.gov/compendia/statab/cats/population.html>.
- [Washington, 2009] Washington, M.L. 2009. Evaluating the Capability and Cost of a Mass Influenza and Pneumococcal Vaccination Clinic via Computer Simulation. *Medical Decision Making*, **29**(July-August), 414–423.
- [Wein & Atkinson, 2009] Wein, L.M., & Atkinson, M.P. 2009. Assessing Infection Control Measures for Pandemic Influenza. *Risk Analysis*, **29**(7), 949–962.
- [Wein *et al.*, 2003] Wein, L.M., Craft, D.L., & Kaplan, E.H. 2003. Emergency Response to an Anthrax Attack. *PNAS*, **100**(7), 4346–4351.
- [White & Pagano, 2010] White, L.F., & Pagano, M. 2010. Reporting Errors in Infectious Disease Outbreaks, with an Application to Pandemic Influenza A/H1N1. *Epidemiologic Perspectives & Innovations*, **7**(12).
- [White *et al.*, 2009] White, L.F., Finelli, J., Wallinga L., Reed, C., Riley, S., Lipsitch, M., & Pagano, M. 2009. Estimation of the Reproductive Number and the Serial Interval in Early Phase of the 2009 Influenza the Current Influenza A/H1N1 Pandemic in the USA. *Influenza and Other Respiratory Viruses*, **3**(6), 267–276.
- [Wilkening, 2008] Wilkening, D.A. 2008. Modeling the Incubation Period of Inhalational Anthrax. *Medical Decision Making*, **28**(4), 593–605.
- [Yang *et al.*, 2009] Yang, Y., Sugimoto, J.D., Halloran, M.E., Basta, N.E., Chao, D.L., Matrajt, L., Potter, G., Kenah, E., & Longini, I.M. 2009. The Transmissi-

bility and Control of Pandemic Influenza A (H1N1) Virus. *Science*, **326**(5953), 729–733.

[Zaric *et al.*, 2008] Zaric, G.S., Bravata, D.M., Holty, J-E.C., McDonald, K.M., Owens, D.K., & Brandeau, M.L. 2008. Modeling the Logistics of Response to Anthrax Bioterrorism. *Medical Decision Making*, **28**, 332–250.