End-to-End Outpatient Clinic Modeling for Performance Optimization and Scheduling in Health Care Service

by

Rafael Bello Fricks

Department of Biomedical Engineering
Duke University

Date: October 25, 2018
Approved:

_____
Roger C. Barr, Co-Supervisor

_____
Kishor S. Trivedi, Co-Supervisor

_____
Henry Tseng

_____
Amanda Randles

_____
Lingchong You

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biomedical Engineering in the Graduate School
of Duke University

2018

ABSTRACT

End-to-End Outpatient Clinic Modeling for Performance Optimization and Scheduling

in Health Care Service

by

Rafael Bello Fricks

Department of Biomedical Engineering
Duke University

Date: October 25, 2018
Approved:

_____
Roger C. Barr, Co-Supervisor

_____
Kishor S. Trivedi, Co-Supervisor

_____
Henry Tseng

_____
Amanda Randles

_____
Lingchong You

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Biomedical Engineering in the Graduate School of
Duke University

2018

Abstract

Decisions in health care must often be made under inherent uncertainty; from treating patients, to provisioning medical devices, to operational decisions at an outpatient clinic. The outcomes depend on the health of patients as well as the availability of health care professionals and resources. Complex models of clinic performance allow for experiments with new schedules and resource levels without the time, cost, unfeasibility, or risk of testing new policies in real clinics. Model-based methods quantify the effect of various uncertain factors such as the availability of personnel on health care quality indicators like patient wait times in a clinic.

Despite their purported value, few opportunities have existed to test models from data collection through optimization. This dissertation develops a clinic model from end-to-end, beginning with a description of the medical practice, to data collection, to model validation, to optimization. Specialty medical practice is abstracted into treatment steps, measured electronically, and verified through systematic observation. These data are anonymized and made available for researchers. A validation framework uses the data to develop and test candidate models, selecting one that maximizes predictive accuracy while retaining interpretability and reproducibility. The resulting model is used in improving schedules via heuristic optimization. Clustering the results reveals clinic performance groups that represent different goals in clinic quality.

Dedication

To my mother, father, and brother: close family even at a distance.

Contents

Abstract iv

List of Tables ix

List of Figures x

List of Abbreviations xiii

Acknowledgements xvi

1. Introduction 1

1.1 The Global Need for Efficiency in Health Care 3

1.2 Contributions and Organization of this Dissertation 6

2. Overview of Quality Improvement in Health Care 9

2.1 The Need for Predictive Modeling in Quality Improvement 11

2.2 Progressive Development in Relevant Stochastic Modeling 14

2.2.1 Queuing Models and State-Space Methods 15

2.2.2 Non-Markovian State-Space Models and the "Largeness" Problem 20

2.2.3 Stochastic Petri Nets 24

2.2.4 Solution in Practice through Discrete Event Simulation 28

2.2.5 Validating Models for Generalization Performance 33

2.3 Operations Research in Health Care 38

2.3.1 Quantifying Decision Tradeoffs 43

3. Understanding Outpatient Care Process 48

3.1 Procedural description of glaucoma practice 49

3.1.1 Initial Step: Workup 51
3.1.2 Intermediary Step: Testing or Imaging 55
3.1.3 Final Step: Physician Contact 57
3.1.4 Miscellaneous and Additional Treatment Steps 60
3.2 Description of Duke Eye Center 62
4. Measuring Performance in Outpatient Clinics 66
4.1 Performance Measurement Through Event Logging 66
4.2 Verifying Event Logging Data 68
4.3 Determining Adequate Sample Size for Estimator Consistency 74
4.3.1 Consistency and Moment Estimators 75
4.3.2 Windowed Subsampling 77
4.3.3 Summarized Sample Size Results and Analysis 78
4.3.4 Conclusions on Sample Size for Consistent Estimation 80
4.3 An Anonymized Performance Data Repository for Operations Research, and Anonymization Methods 82
5. Modeling Outpatient Clinics 86
5.1 Challenges in Modeling High Volume Outpatient Practices 88
5.2 Generalized Simulator Logic 92
5.3 Representations of Treatment Time 93
5.4 Modeling Concurrent Clinic Interdependence 100
5.5 Evaluating Generalization and Reproducibility 104
5.6 Accounting for Incongruencies Between Models and Practice 112

5.7 Summarizing Clinic Modeling 116

6. Optimizing Clinic Performance 122

6.1 Dimensionality of Scheduling Input Space 122

6.2 Trade-Off Between Patient Wait Time and Staff Idle Time 125

6.3 Enhanced Quality Metrics 127

6.4 Clustering through Unsupervised Learning of Quality Metrics 130

6.5 Schedule Selection Using Performance Clusters 132

7. Conclussions 136

Appendix A: Analytic Modeling in Chronic Kidney Disease 138

Appendix B: Template for Observation 143

Appendix C: Core of Simulation Program 144

References 150

Biography 160

List of Tables

Table 1: Simulation System Parameters and Results for Entry Process Model..... 102

Table 2: Input parameters and output response of the glaucoma clinic stochastic reward net model (Range of values in parenthesis). 106

Table 3: Test results using the sequestered data, arranged by clinic and check point. ... 109

Table 4: Comparison of scheduling heuristics by performance metrics. 135

List of Figures

Figure 1: SRN Model for ICU Treatment. 45

Figure 2: The probability a device is available for treatment path A when needed. 46

Figure 3: The average number of idle devices. 47

Figure 4: General steps in a visit to a glaucoma clinic, with expected waiting periods explicitly marked. 50

Figure 5: Technicians perform the initial patient interview, confirming the current medications and medical history among other questions 52

Figure 6: Technicians may also perform some eye measurements 53

Figure 7: Acuity testing is a staple of ophthalmology visits, performed during workups using a mirrored setup that increases the path length to vision chart. 54

Figure 8: A variety of imaging procedures and testing may be required to monitor patient health. 56

Figure 9: A doctor may perform an independent examination of the patient. 58

Figure 10: Patients may spend a considerable portion of the visit with the physician discussing a treatment plan. 59

Figure 11: The eye center was dimensioned with various concurrent practices in mind, as well as on-site imaging and testing. 63

Figure 12: Cornea specialties and other departments, as well as operating room space unfinalized as of August 2015. 64

Figure 13: Detailed accounting of the facilities available for Glaucoma practices. 65

Figure 14: Measurement points in event logging, where practitioners agree on a finite set of standard events to record through inputs at EHR. 68

Figure 15: In-person observation was performed in Glaucoma. 70

Figure 16: Histograms compared for in-person observation versus event logging collected from EHR, with a phase-type distribution in the lower frame. 71

Figure 17: Cumulative distribution function graphical comparison for in-person observations, event logging through EHR, and a phase-type distribution model. 72

Figure 18: Diagrammatic representation of the windowed sampling scheme. 78

Figure 19: The range of moment estimates at each window size (Eq. 6). 79

Figure 20: Steps in glaucoma treatment, showing measurement points in the event logging system. 87

Figure 21: Data Partitioning Diagram for Cross-Validation. 88

Figure 22: Entry process for early eye center model. 90

Figure 23: Subsequent steps in eye center treatment. 91

Figure 24: Mean absolute error for phase-type distributions representing workup durations. 97

Figure 25: Absolute error for phase-type distributions with varying number of phases, used to model the duration of generic imaging procedures. 98

Figure 26: Iterations of the concurrent model demonstrated that delays in entry were solely due to staffing demand. 101

Figure 27: Workup prediction accuracy improves significantly when clinic concurrency is explicity modeled. 103

Figure 28: Model for glaucoma service incorporating workup and testing stages with the decision to test modeled as a probabilistic switch. 106

Figure 29: Test set results for SC B, at check point 1. 111

Figure 30: Test set results for SC B, at check point 2 show similar results. 112

Figure 31: Two approaches to modeling doctor activity in a clinic. 114

Figure 32: Comparison of model performance with and without inserted tasks. 115

Figure 33: Complete Glaucoma Clinic Model, based on outpatient practices at Duke Eye Center (DEC). 117

Figure 34: Comparison of check point 1 measurements to model predictions. 118

Figure 35: Comparison of check point 2 measurements to model predictions. 119

Figure 36: End-to-end comparison of simulation outputs versus recorded visit durations at a DEC glaucoma clinic over an 18-month interval. 121

Figure 37: Execution time variability for 41 heuristic schedule generating methods. 124

Figure 38: Schedule diagram, illustrating terminology for templated schedules..... 125

Figure 39: Patient wait time to personnel idle time tradeoff curve, with real schedule performance indicated by the orange circle. 126

Figure 40: Scheduling rules plotted by cost function on logarithmic axes. 130

Figure 41: Preliminary clustering using k-means algorithm with k = 3. 131

Figure 42: Functional grouped determined through iterative re-application of k-means clustering 132

Figure 43: Time-homogenous continuous time Markov chain model of chronic kidney disease progression where cadaveric transplantation is possible..... 139

Figure 44: Mathematica Output for equation solver for A,B,C,D terms as functions of the model parameters 142

Figure 45: Example observation template form used in data collection. 143

List of Abbreviations

AMIA: American Medical Informatics Association

CHOIR: Center for Health Operations Improvement and Research

CKD: Chronic Kidney Disease

COA: Certified Ophthalmic Assistant

COT: Certified Ophthalmic Technician

CTMC: Continuous Time Markov Chain

CVF: Confrontational Visual Field

DEC: Duke Eye Center

DES: Discrete Event Simulation

DUHS: Duke University Health System

EHR: Electronic Health Record

EOM: Extraocular muscles or extraocular movement

ESRD: End-Stage Renal Disease

GDP: Gross Domestic Product

HHS: Health and Human Services (Department)

HIPAA: Health Insurance Portability and Accountability Act

ICU: Intensive Care Unit

IJCAHPO: International Joint Commission on Allied Health Personnel in Ophthalmology

IOP:	Intraocular Pressure
MAE:	Mean Absolute Error
MCF:	Mean Cumulative Function
MLE:	Maximum Likelihood Estimation
MRGP:	Markov Regenerative Process
MRSE:	Mean Root Squared Error
NHS:	National Health Service
NIH:	National Institutes of Health
OCR:	Office for Civil Rights
OCT:	Optical Coherence Tomography
OECD:	Organization for Economic Cooperation and Development
OR:	Operations Research or Operating Room
OT:	Operating Theater
PCAST:	President's Council on Science and Technology
PH:	Phase-type distribution
PHI:	Protected Health Information
PoC:	Point of Care (terminal)
SOAP:	Subjective – Objective – Assessment – Plan (record keeping method)3
SMP:	Semi-Markov Process
SPN:	Stochastic Petri Net

SRN: Stochastic Reward Net

VF: Visual Field (test)

Acknowledgements

First and foremost, I dedicate this work to my family. To my mother, Evelise Bello Fricks, who raised Robson and I to dedicate ourselves to helping others as she did for us. She always taught us to be more disciplined, studious, kind, and appreciative than I am today. To my father, Ricardo Messias Fricks, the constant student and consummate professional I aspire to be. His efforts made my work possible, sometimes quite literally. To my brother, Robson Bello Fricks, whose constant competition molded my formative years and fueled the drive I live by today. I am proud of the person and family man he is, and hope I will be like him when I'm that old.

I am also grateful for my mentors and peers throughout this process. I have been extraordinarily fortunate in both of my advisors, Professor Kishor S. Trivedi and Professor Roger C. Barr, who have sparked my progress in my graduate experience and beyond. I thank the committee members, Professors Henry Tseng, Amanda Randles, and Lingchong You for their guidance and oversight. Finally, all my professors, classmates, friends, mentors and mentees at Duke University. I am proud of my time at this unique institution, and proud to be a Blue Devil.

To the faculty and staff at Duke Eye Center, this research could not have happened without their dedicated efforts. I especially thank Dr. Henry Tseng and Marjorie Veihl, for introducing me to real clinical practice in addition to their collaboration in research. I also send a special thanks to all who volunteered their time,

especially Dr. Mike Kelly, Rochelle Ingram, Kendall Kruszewski, Joseph Griffith, William Chen, and every technician who graciously participated, of which there are too many to list.

A special thank you to my friends in and around Durham, N.C. whom have been like second family to me. I have made a second home here, and aim to stay and continue making fond memories. They can never replace my extended family in Brazil, who I try to honor daily in my actions, but both have lifted me to where I am today.

Finally, I express my gratitude to the agencies that have supported my graduate studies, the Graduate School at Duke University and the National Science Foundation. My reserach throughout this dissertation received generous support from the Duke University Dean's Graduate Fellowship and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGF 1106401 J.

1. Introduction

Advances in medicine open new avenues of treatment for patients. Delivering these treatments to patients requires coordination between physicians and staff, making the best use of available devices, facilities, and other resources. When demand on finite health care resources exceeds capacity, the quality of patient care is affected (Reid et al., 2005). In outpatient clinics, lack of access to or delays in service can greatly affect patient satisfaction and perception of care quality (*Keep Me Waiting: Medical Practice Wait Times and Patient Satisfaction*, 2010; Michael, Schaffer, Egan, Little, & Pritchard, 2013).

Research in optimizing health care delivery, namely operations research, has seen a boon in recent years as organizations face mounting pressure to provide high quality care at lower cost to more people. The challenge in optimizing schedules through models is ensuring models accurately represent medical practice. Various contributions have consistently advanced technical aspects of modeling health care systems. In contrast few contributors disclosed extensive performance data on clinic service. Consequently, studies with independent replication or performance comparisons between competing models of the same practice are virtually non-existent. Several canonical models have been re-applied to functionally similar health care services, contributing to a sense of slow progress in health care operations research (Gunal & Pidd, 2010). Many documented applications begin and end at the model, and

methods achieve differing tradeoffs of model accuracy, computational cost, and ease of use.

This dissertation presents a comprehensive end-to-end approach to quantifying operational decisions in a clinic, beginning with observation and measurement and finishing with modeling and optimization. The end goal of optimization begins with a detailed annotation of clinical practice at high volume outpatient clinics at Duke Eye Centers (DEC), abstracted into generic treatment steps. Stepwise descriptions of patient visits to the center form the basis of a performance measurement set spanning nearly two years of clinical practice. These data serve as a new resource for operations research; a key contribution of this dissertation to replicability in operations research is making an anonymized version of the data available in an online repository.

A simulation model is then proposed to represent the treatment step abstraction of clinical practice, using the data set to parameterize and validate the model. Specifying the model as simulated stochastic reward nets maintains accuracy, ease of use, and the ability to modify system behavior with reasonable computational cost. The model sees iterative improvement which maximizes accuracy in predicting patient flow through clinics at DEC. This iterative approach incorporates cross-validation techniques to estimate prediction accuracy against data withheld from model formulation. The result combines advances in model evaluation to ensure an accurate representation of real clinics. An accurate model of outpatient clinic performance at DEC is produced.

The clinic model is finally used to optimize schedules at an outpatient center. An astronomically vast number of plausible schedules makes exhaustive search ineffective. The high model complexity prevents the use of traditional optimization techniques such as analytical optimization, gradient techniques, or structured search. Recognizing performance boundaries in the studied clinic significantly improves optimization efficiency. A heuristic is proposed for specifying schedules that are easy to implement in practice. Heuristic modifications to existing schedules outperform the current schedule in a multifaceted comparison of many nuanced quality metrics. The results are further distinguished using unsupervised learning techniques to produce functional groupings in scheduling schema that define performance groups for clinics.

By detailing all methods, data, and modeling choices in producing the end result, this dissertation aims to advance reproducibility and objective comparison in health care operations research.

1.1 The Global Need for Efficiency in Health Care

Among all OECD nations, the United States spends the most by far per capita on health care, amounting to 17.2% of GDP in 2016 (OECD, 2017). These massive expenditures, two to three times higher than the average for many comparable health systems, do not translate into significant improvement in population health measures (OECD, 2017). Though health care may benefit globally from engineering approaches, the U.S. health system in particular requires tools for evaluating efficiency and assuring

quality. In urgency consistent with economic data, several national reports highlight a disparity in employing known successful approaches in improving health care delivery (Reid et al., 2005). A joint report by the National Academy of Engineering and Institutes of Medicine details the transformative potential of identifying and advancing tools for managing complex systems (Reid et al., 2005). Two critical developments proposed in the consensus report lay in the realization of these tools as well as integrating information technologies in health care practice (Reid et al., 2005). More recently, recommendations from the President's Council of Advisors on Science and Technology articulated a need for developing methods for gauging health care efficiency (President's Council of Advisors on Science and Technology (PCAST), 2014). Leveraging health information technology systems effectively is key to improving efficiency and reduce the economic burden of health care (President's Council of Advisors on Science and Technology (PCAST), 2010). The expanded adoption of comprehensive electronic health record (EHR) systems hypothetically opens several opportunities for performance evaluation. As few as 6.4% of surveyed health organizations have advanced use of EHR that includes data analytics according to the latest estimates (HIMSS Analytics, EMRAM 2017 Q4).

Although America may be at the forefront, the cost of health care is not a uniquely American problem. Nations with comparable industrial development, such as the G7 countries, expend between 8.9% (Italy) to 11.3% (Germany) of their gross

domestic product (GDP) on health care (OECD, 2017). These monumental expenditures have grown steadily, with OECD countries experiencing an average growth increase of 1.4% since 2009 (OECD, 2017). A desire to curtail these costs is reflected in examples of research output from countries with nationalized health services.

Notably in the Netherlands (10.5% health expenditure as a fraction of GDP), groups of researchers such as the Center for Healthcare Operations Improvement and Research (CHOIR) have consistently examined health care quality from an operations research perspective (CHOIR, 2018). Seminal contributions from CHOIR include doctoral theses such as (Kortbeek, 2012), and the curation of the operations research database ORchestra (Hulshof et al., 2011).

Similarly the United Kingdom (9.7% of GDP on health care), which provides health care through its National Health Service (NHS), has seen a variety of research efforts directed at improving health care efficiency. Systematic reviews (Fone et al., 2003) have documented the use of operations research models in informing health care decisions at an organizational level. Exemplary works span applications from the NHS in the 1960's (Brailsford, 2007) to more broad reviews on the current approach and challenges in health care modeling (Gunal & Pidd, 2010).

Research in improving health care efficiency draws an international effort. Though economic and political factors influence the provision of health care (Brill, 2015), research methods address similar underlying operational efficiency questions regardless

of country. However, as reviews have lamented, developments in operations research still face barriers in affecting health care improvements locally (Gunal & Pidd, 2010; Reid et al., 2005). Challenges remain in ensuring that treatments are broadly available, or that health care systems can meet current demands (Brill, 2015; Reid et al., 2005). Considering the enormous costs of operation and resource limitations in many clinics, even marginal improvements can be significant. International demand continues to grow for methods for evaluating health system performance and quantifying the expected outcomes of administrative decisions.

1.2 Contributions and Organization of this Dissertation

The remainder of this dissertation is organized as follows:

Chapter 2 presents an overview of theoretical developments in stochastic modeling and their applications in operations research and health care. The progressive presentation positions the model developments in later chapters, which expand the application of Petri net formalisms.

Chapter 3 details clinical practices in ophthalmology, with an emphasis on glaucoma clinics and the specialty outpatient practices studied in this dissertation research. This chapter condenses evidence-based practices into a procedural description aimed at non-clinicians. Treatment at these clinics is abstracted into distinct treatment steps, which form the basis of subsequent measurement and modeling efforts. These steps are chosen to delineate operational decisions from medical science.

Chapter 4 describes the collection methods for performance data in large practices. Electronically collected performance data is independently verified through in-person observation. Statistical estimators are then used in a detailed examination of sample size adequacy in modeling clinic performance. The collected database is published to a repository. A detailed documentation of the anonymization process is included as a template for reproducing these efforts at other institutions.

Chapter 5 iteratively builds on models of clinical description using a cross-validation framework. Several common assumptions in modeling are evaluated for accuracy, producing a 'white-box' model of clinic performance. Evaluations include the effect of independence assumptions in approximate models of concurrently operating clinics. Emulating the activity and resource demands of adjacent clinics improves accuracy in the clinic model. Similarly, adjustments for work-conserving assumptions implicit in nearly all queuing models are introduced to accommodate human practitioner behavior. The iterative approach produces a model of outpatient service with extensive validation in the studied practice and details for adapting the model for other services.

Chapter 6 presents a heuristic method for specifying schedule templates that are actionable by patients and staff. The performances of several plausible schedules are evaluated using the clinic performance model from the preceding section. These schedules indicate a fundamental tradeoff between wait times and staff idle time at a

given patient load. This chapter defines new clinic performance metrics that allow for a more nuanced discrimination between scheduling schema. The metrics are clustered using unsupervised learning algorithms, resulting in clinic performance groups.

Chapter 7 concludes the discussion with some final observations.

2. Overview of Quality Improvement in Health Care

Before modeling a clinic or proposing changes to health care delivery, quality improvement begins with establishing what is quality in health care and the metrics by which it will be measured. Health care organizations vary in defining performance and quality based on the setting or perspective (Ahmadi-Javid, Jalali, & Klassen, 2017; Cayirli & Veral, 2003; Kortbeek, 2012). This discussion defines health care performance as the ability to serve patients at a given quality of care. Example metrics for quantifying quality of care include patient wait times, staff idle time, or resource utilization. Quality improvement endeavors in health care seek to identify changes to service which improve performance, as measured by quality indicators. Changes to a clinic or other health care system can be considered a performance improvement based on the effect on quality of care, such as lowering patient wait times or raising the number of patients served. All subsequent techniques, irrespective of formulation, are attempts at identifying changes that yield quality improvements in health care.

Administrators today can draw upon considerable theory to guide quality improvements in a health care system. These theories vary greatly the expertise required to apply them, and in their ability to predict the response of changes or quantify quality. The enormity of techniques applied in quality improvement are beyond the scope of any single discussion. While the focus here is model-based quality improvement, this

discussion briefly introduces some non-modeling quantitative methods in common use to motivate the stochastic modeling approach developed in subsequent chapters.

The remainder of the chapter presents an overview of stochastic modeling for quantitatively evaluating proposed improvements to health care and introduces the theoretical basis of many model-based methods. Core stochastic modeling techniques that form the theoretical basis of many operations research approaches are presented in progressive order to illustrate iterative developments in modeling techniques. This progression is then used to organize a sampling of publications in the vast domain of operations approaches to health care. Finally, a hybrid approach is demonstrated in a hypothetical system model to illustrate the ability of technical advancements to quantify nuanced tradeoffs for decision making. Later models developed to predict clinic performance at Duke Eye Center (DEC) are based on the hybrid approach.

This discussion relies on terminology from stochastic processes, probability modeling, and statistics. Concepts such as random variables, distribution functions, event algebra, and sample space among others are fundamental to understanding models of health care service. The discussion assumes readers are familiar with core concepts in probability modeling. Several texts have been written on this topic and are recommended for readers looking to refresh their knowledge (Bertsekas, 2008; Çınlar & Sollenberger, 2013; Durrett, 2012; Manolakis, Ingle, & Kogon, 2005; Rosner, 2011; Ross, 1997; Trivedi, 2002).

2.1 The Need for Predictive Modeling in Quality Improvement

In general, a model is an approximate representation of a system (Cobelli & Carson, 2008). Models of health care performance provide a simplified representation used to evaluate the system studied (Gunal & Pidd, 2010). Abstraction, the process by which a modeler selects details to represent, is subjective. Consequently several viable models may describe the same system with comparable accuracy yet serve different purposes by virtue of the details represented (Breiman, 2001). The need for model-based approaches that introduce a degree of subjectivity can be motivated through contrasting with alternatives for system improvement in common use.

In discussing statistical modeling, Leo Breiman grades case studies on two aspects of the models used, referred to as prediction and interpretability (Breiman, 2001). A predictive model produces an accurate response to future inputs when compared to the future system outputs (Breiman, 2001; Hastie, Tibshirani, & Friedman, 2009). An interpretable model makes clear the association between inputs and outputs (Breiman, 2001). Interpretability relates to the concept of black-box versus white-box modeling (Breiman, 2001; Cobelli & Carson, 2008), where a black-box model may make accurate predictions while providing no information about how the system associates inputs to outputs. Consistent, ongoing quality improvement in health care requires white-box model-based approaches that are interpretable and predictive.

For contrast, consider managerial science practices for quality improvement that are common in health care. Many health systems adopt techniques from manufacturers for continual improvement, in particular Lean management principles (Lawal et al., 2014). Lean principles are practices for identifying and eliminating eight categories of waste, based on experiences in automotive manufacturing at Toyota Motor Corporation (Liker, 2004). A detailed treatment of management concepts behind Lean is provided by (Womack & Jones, 2003). Rarely are large health systems without continual improvement efforts, often overlapping at various levels of organization. Facilities following Lean methods train staff to re-evaluate practices and look for everyday waste (Kennedy, 2018). These efforts lead to various ad hoc adjustments of health care systems. In practice Lean methods do not predict the response of changes to service, rather improvements are adopted by trial and error. Without adequate predictive support, decisions may be difficult to evaluate or lead to inaccurate conclusions on their effect (Fenton & Neil, 2012). The ability to assess a change may be limited, as not all facilities measure performance, and with continual improvement few instances of the changed process may occur before another change is implemented. Given the inherent variability in health care, it may not be feasible to distinguish successful improvements from fortunate circumstances. While the widespread adoption of Lean principles indicates net improvements to health care delivery, models of health care with high predictive

accuracy may further eliminate waste by reducing the number of ineffective changes that are enacted.

Predictive accuracy without interpretability is insufficient for evaluating most changes prior to implementation. Several black-box models are in use in health care to represent measurements with the potential for random data censoring. These models may accurately predict response to an input without explaining the underlying processes in the system. Notable in non-parametric methods is the Kaplan-Meier or product-limit estimator for survival functions, which represent the time until an event occurrence based on relative frequency within an observed cohort (Miller, Gong, & Muñoz, 1981; Rosner, 2011). Kaplan-Meier estimators in the clinic performance context may provide a probabilistic description of the time until a patient completes a treatment, but the response time may be conditional to operating conditions such as staffing conditions or other patients in the clinic. Similarly the mean cumulative function (MCF) is a technique for representing the average number of occurrences at a given time (Nelson, 2003). Graphing the MCF of a measured quantity can show trends in system throughput and has been used in studying medical imaging device repair (R. B. Fricks & Trivedi, 2017). Black-box models may be intuitive representations of a system, but they lack the interpretability necessary to explore new configurations of the system.

In accurately predicting clinic response to a change in service, interpretable models must account for variability in patients and service time while explicitly

representing interactions in a clinic system. The remainder of this chapter elaborates on developments in white-box models of health care predominantly in the field of operations research, as well as validation techniques for estimating predictive accuracy.

2.2 Progressive Development in Relevant Stochastic Modeling

Developments in stochastic modeling have represented systems in many disciplines, from health care, to operations research, to dependability engineering, to cybersecurity and software development (Gunal & Pidd, 2010; Hulshof et al., 2011; Jun, Jacobson, & Swisher, 1999; Kortbeek, 2012; Rigdon & Basu, 2000; Trivedi, 2002; Trivedi & Bobbio, 2017). Model-based analysis of health care service at the outpatient level centers around how patients move through a system, such as a clinic. Such a system is an example of a 'system of flow,' where in corresponding terminology, commodities proceed through the system through finite-capacity channels (Kleinrock, 1975). The analysis method depends on the predictability of flow, whether it is steady *deterministic* flow or unsteady *stochastic* flow (Kleinrock, 1975).

Steady-flow analysis treats patient movement as a continuous, deterministic flow and greatly simplifies the required mathematics when many patients are represented. Examples of steady-flow analysis are system dynamics techniques, which are viable for some quality improvement decisions at higher levels of health care organization (Brailsford, 2007). System dynamics approximations are not generally adequate for modeling flow at the individual patient level.

Representing individual patients at the clinic level, flow tends to be unsteady. Numerous factors—e.g. patient mobility, to case complexity, to the individualism with which personnel treat patients— result in random visit durations. Clinic flow requires stochastic modeling approaches.

Stochastic modeling methods treat health care service as a stochastic process, where the system performing a service progresses non-deterministically through defined states (Cınlar & Sollenberger, 2013; Durrett, 2012; Ross, 1997; Trivedi, 2002). Most efforts model the process using a quantifiable formalism and use the generated model to test changes to clinic service, typically patient or staff scheduling. From these commonalities, applications vary greatly in the quality of data collection and analysis, the formalism chosen, the degree to which models are validated, the completeness of the solution, and degree of follow-up. In this progressive presentation, modeling paradigms are introduced in escalating ability to represent systems, terminating in Petri net formalisms. Once introduced, special considerations for parameterization, solution, and model validation in practice are discussed.

2.2.1 Queuing Models and State-Space Methods

Queues present the simplest model of unsteady flow through a system. In a queue, customers arrive at uncertain times, wait until a server is available if necessary, and depart after being served for a randomly distributed duration by the server. Variations on queuing systems are compactly described by Kendall's notation,

attributed to D.G. Kendall (Kendall, 1951), and adopted nearly universally by authors in describing queues such as (Durrett, 2012; Kleinrock, 1975; Kortbeek, 2012; Ross, 1997; Hideaki Takagi, Kanai, & Misue, 2017; Trivedi, 2002). In its current form, Kendall's notation is typically written as:

$$A/D/s/L/x$$

Where

- A: an encoded value describing the inter-arrival distribution, such as 'M' for memoryless (exponentially distributed), or 'G' for generally distributed, amongst others
- D: the service time distribution, encoded similarly to the arrival distribution
- s: the number of independent servers available
- L: the queue length
- x: additional features, such as the scheduling discipline, such as 'FCFS' for first come, first served

For an example using this notation, the default system is the M/M/1 queue, where inter-arrival time is exponentially distributed, service time is exponentially distributed, and there is one server that services arriving customers. The omission of further values (L/x) implies default values of an infinite potential queue length, and first come first served scheduling discipline. M/M/1 is a commonly used example queue (Kleinrock, 1975; Trivedi, 2002).

Elementary queuing model solutions are well-known (Çınlar & Sollenberger, 2013; Kleinrock, 1975; Trivedi, 2002), From a health care perspective, an independent queue is likely to adequately fit highest priority departments that share few resources with other departments, such as intensive care units (ICU) (McManus, Long, Cooper, & Litvak, 2004). Queuing models are only analytically tractable under well-defined conditions, which are the default specification or usually assumed during modeling. These conditions include limitations on the arrival and service distributions, which in the simplest case are assumed to be independent and identically distributed (i.i.d.), and exponentially distributed.

Suppose it is insufficient to assume a queue operates independently. The networks of queues (or queuing networks) theoretical framework exploits narrowly defined dependence to obtain compactly specified models (Bolch, Greiner, de Meer, & Trivedi, 2006; Trivedi, 2002). Queuing network solutions require constraints on system representation such as restriction to exponential distributions, or that the network sees no net loss of customers (Trivedi, 2002). (Kortbeek, 2012) formulated several queuing network models and concluded the constraints in their formulation unsuitable for representing health care service. The expertise required to formulate or adjust queuing network models tends to restrict the systems that can be represented (Kortbeek, 2012; van Dijk & Kortbeek, 2009), and many authors rely on canonical models and their solutions (Hideaki Takagi et al., 2017; H. Takagi, Misue, & Kanai, 2014).

Alternatively, interdependence between queues can be modeled using state-space methods. Broadly speaking, state-space method is a blanket term for analysis techniques that retain some notion of the states in which a system can exist, and specifies (potentially) stochastic transition timings between states (Trivedi & Bobbio, 2017). State-space models can be formulated to represent arbitrary and complicated dependence between system states (Trivedi, 2002; Trivedi & Bobbio, 2017). A vast majority of state-space techniques restrict transition time distributions to the exponential distribution for their desirable analytic properties, which are now elaborated.

Properties of the exponential distribution facilitate analytical solution when a system can be modeled as a set of exponential transitions. Chiefly, this is due to the memoryless property of an exponential distribution, derived in (Kleinrock, 1975; Trivedi, 2002). When a duration X is said to be exponentially distributed with parameter λ, given that t_1 has elapsed, the conditional distribution for the remaining time is also exponentially distributed with parameter λ. This result is expressed as a conditional probability,

$$\Pr[X > t + t_1 | X > t_1] = \Pr[X > t]$$

The memoryless property implies that the time spent waiting for duration X to elapse is irrelevant to the probability distribution of the remaining duration. When the number of patients in a clinic at time t (N(t)) are modeled as a queue with exponential service and interarrival times, the state of the system and future evolution at any time is

fully described by the number of patients present and the transition rate parameters. This stochastic process model is defined for continuous time points with discrete states. A finite number of states correspond to the number of patients present, and an arrival or departure is considered a discrete event occurring at a specific time, causing the system to transition between states. When all state transitions are memoryless, the Markov property (Çınlar & Sollenberger, 2013; Kleinrock, 1975; Trivedi, 2002) holds at all time points, where the Markov property is defined as

$$\Pr[N(t_{k+1}) = n_{k+1} | N(t_0) = n_0, \ldots, N(t_k) = n_k] = \Pr[N(t_{k+1}) = n_{k+1} | N(t_k) = n_k]$$

When a system is Markovian, progression from the current state depends only on the current state, not the preceding states. State-space methods define a system with stochastic behavior by the probability of existing in one of various states. State descriptions must be complete such that state descriptions, in combination with the model itself, sufficiently describe future evolution of the system. When the Markov property holds at all moments in time, the stochastic process is a continuous time Markov chain (CTMC). Most queues are analytically tractable when all transitions can be decomposed into various exponentially distributed phases (Fackrell, 2008; Kleinrock, 1975; Neuts, 1981), resulting in a more complex definition of state which can be modeled as a CTMC (Trivedi, 2002; Trivedi & Bobbio, 2017).

2.2.2 Non-Markovian State-Space Models and the "Largeness" Problem

Not all random intervals in health care service are exponentially distributed. This section provides an overview of strategies for modeling non-Markovian systems. Recall that state-space models describe a system that exists in one of several states and specifies transitions between those states. Suppose a model specifies that three patients are independently serviced by three technicians, with no other patients or personnel present. Two cases are considered, (A) the service time distribution for a technician attending to a patient is exponentially distributed, (B) identical to (A), however a Weibull distribution is used to model service time distribution.

In case (A), the system state is fully specified by the number of patients present and can be said to be in one of four discrete states, where 0-3 patients are under treatment. It is unnecessary to track the elapsed treatment time when each patient leaves as, due to the memoryless property of the exponential distribution, the elapsed time is inconsequential to the remaining time until completion. Case (A) is the basis of many elementary queue solutions (Kleinrock, 1975).

In case (B), the distribution of the remaining time until treatment for each patient must be conditioned on the elapsed time. Whereas before the system state had three possibilities, this new system has continuous states which transition with each infinitesimal moment of elapsed time.

Case (B) provides a cursory glance at the problems in formulating non-Markovian models directly. Two primary strategies in modeling this case involve explicitly modeling the non-exponential distribution or approximating the non-exponential distribution as a combination of exponential stages.

The first strategy explicitly formulates models with non-exponential distribution by exploiting instances when the system probabilistically resets. Theoretical developments in Semi-Markov Processes (SMP) and Markov Regenerative Processes (MRGP) allow for formulating these models under strict constraints (Trivedi & Bobbio, 2017). Queues with either non-exponential service or inter-arrival times employ similar constraints in their canonical solutions (Kleinrock, 1975). However, one of the primary limitations is that only one non-exponentially distributed transition may be active in a given state (R. M. Fricks et al., 1998; Trivedi & Bobbio, 2017). Queues with an arbitrary number of concurrent, non-exponential transitions have defied analytical solution (Bertsimas, 1990). Given the tremendous expertise and unique considerations for formulating each state-space model with non-exponential transitions, SMP and MRGP have seen limited use in health care models. Concurrent service situations such as case (B) are frequent in health care.

In the alternative strategy, non-exponential distributions can be explicitly modeled or well-approximated as a combination of exponential distributions. This method is variably referred to as stage expansion, Markovization, or the method of

stages (Kleinrock, 1975; Trivedi, 2002). Often non-exponential distributions are replaced with phase-type distributions, which are represented by a Markov process composed of several exponentially distributed 'phases'. Phase-type (often abbreviated PH) distributions provide a class of causal distributions with scalable complexity that were previously found to well represent the duration of patient visits (Fackrell, 2008; R. B. Fricks, Tseng, Veihl, Trivedi, & Barr, 2018; Horvath & Telek, 2017; Kleinrock, 1975; Thummler, Buchholz, & Telek, 2005; Trivedi & Bobbio, 2017). The visit duration PH distribution with D phases can be compactly expressed in matrix form using the matrix exponential (Al-Mohy & Higham, 2010) as

$$F(t) = 1 - \boldsymbol{\alpha} e^{At} \mathbf{1}$$

Where A is a DxD matrix of real-valued numbers, α is a $1xD$ row vector of non-negative values that satisfies $\sum \alpha = 1$, and $\mathbf{1}$ is a $Dx1$ column vector of ones. A PH distribution is itself an absorbing CTMC. Phase-type distributions can theoretically fit a stochastic process to an arbitrary degree of precision (Fackrell, 2008; Kleinrock, 1975; Marshall, Vasilakis, & El-Darzi, 2005; Trivedi & Bobbio, 2017). In effect PH approximations may convert non-Markovian models into Markovian models, at the cost of a more complicated state definition which must now account for phase progression.

One of the challenges of in state-space modeling is that as more detail is added, the number of permutations of system state possible grows exponentially. For example, if clinic operations depend on the availability of M physicians, T technicians, and can have up to N customers present, there are at minimum (M+1)(T+1)(N+1) possible states.

Stage expansion exacerbates this problem as the system state must be subdivided to account for transition phases, which has a multiplicative effect on the number of overall states. The rising number of enumerable states leads can be problematic for many solution methods in practice where computational resources may be limited, particularly as some solution techniques may enumerate all possible states prior to solution. This is referred to as the "Largeness" problem (Haas, 2002; Trivedi & Bobbio, 2017). There are two generally recognized approaches for modeling in the face of largeness, which depend on the nature of the problem:

- *Largeness avoidance:* combine states or otherwise avoid generating all possible state permutations. This approach is indicated when largeness prohibits model solution on existing hardware.
- *Largeness tolerance:* use higher level formalisms that allow model specification to automatically generate the underlying state-space. This approach is advisable when state definitions are patterned, such as a queue with PH distribution service time.

Solutions to specific largeness problems may have aspects of both tolerance and avoidance. Modeling in the presence of largeness is a central theme in (Rahul Ghosh, 2012; R. Ghosh, Longo, Frattini, Russo, & Trivedi, 2014). Largeness, and largeness tolerance in particular is relevant to health care modeling and influences model specification and solution, and will be revisited in context in subsequent sections.

2.2.3 Stochastic Petri Nets

Petri nets were introduced in the doctoral thesis of Carl Adam Petri in 1962 (Gianfranco Ciardo, Blakemore, Chimento, Muppala, & Trivedi, 1993; Silva, 2012). Extensions to petri nets have introduced timing, stochastically timed transitions, additional modeling power, and several modeling conveniences to the formalism such as guards and arc multiplicity (Gianfranco Ciardo et al., 1993; Gianfranco Ciardo & Trivedi, 1993; Trivedi, 2002). Stochastic Petri nets and related formalisms are classical examples of largeness tolerance. They provide a specification formalism for Markov chains that avoids manually defining each permutation of state in a large model, which is tedious and error-prone when performed by a human operator (Gianfranco Ciardo et al., 1993; Trivedi, 2002; Trivedi & Bobbio, 2017).

Petri net and subsequent iterations provide powerful modeling tools for addressing large concurrent systems (Gianfranco Ciardo & Trivedi, 1993; Haas, 2002) such as in health care. The stochastic Petri net formalism (SPN) provides a graphical representation for precisely specifying arbitrarily large CTMCs while retaining overall a coherent graph representation of key system features. Stochastic Reward Nets (SRN) further extend the SPN definition, adding further modeling capabilities described incrementally in (Gianfranco Ciardo et al., 1993), and reiterated in (Trivedi, 2002; Trivedi & Bobbio, 2017). The SRN state-space method can theoretically describe any CTMC, and by implication any system of Markovian queues (Gianfranco Ciardo & Trivedi, 1993;

Haas, 2002). The culmination of analytic modeling from several researchers have indicated SPN and subsequent developments of the formalism such as SRN for modeling health care (Kortbeek, 2012), however few studies have proceeded. While SPN derivatives can model very large clinical systems, they do not eliminate technical barriers such as computational requirements when solved numerically (Gianfranco Ciardo et al., 1993; Trivedi & Bobbio, 2017). The use of SRNs directly also require that all timed transitions are decomposed into combinations of exponential transitions, which may be out of reach for most users. Consequently, SPN clinic models are presented unsolved (Kopach-Konrad et al., 2007; Kortbeek, 2012), as a qualitative tool (Alfonso, Xie, Augusto, & Garraud, 2012), or with minimal analysis (Leite et al., 2010). Use of SPN in health care is rare compared to broader use of queuing models or discrete event simulation (Gunal & Pidd, 2010; Hulshof et al., 2011). SPN, SRN, and related Petri net formalisms are precise techniques for specifying state space models with a unique interpretation for generating the state space (Gianfranco Ciardo et al., 1993).

The definition of the stochastic reward net formalism is critical in specifying SRN models, and is reproduced here from (Gianfranco Ciardo et al., 1993; Gianfranco Ciardo & Trivedi, 1993) with added commentary; a stochastic reward net is defined as an 11-tuple

$$\boldsymbol{A} = [P, T, D^{-}, D^{+}, D^{o}, e, >, \mu_0, \lambda, w, M]$$

- P = $[P_{lobby}, \dots, P_{open}]$ is the finite set of places, each containing a non-negative number of tokens representing entities in a system such as patients.
- $T = [T_{entry}, \dots, T_{asssess}]$ is the finite set of transitions such that ($P \cap T = \emptyset$), which can be used to model steps with finite duration or flow decisions.
- $D^- \in [0,1]^{|P \times T|}, D^+ \in [0,1]^{|P \times T|}$ describe the input and output arcs. An arc exists iff $D^*_{ij} \neq 0$
- $D^o \in [0,1]^{|P \times T|}$ are marking-dependent inhibitor arcs, used for conditional disabling of transitions.
- e is the set of enabling functions, also known as guards. One guard controls the arrival transition as a function of the Open place, tracking hours of operation for this model clinic
- $>$ is the set of transitive and irreflexive relations imposing priorities among transitions. Priorities in SRN are static, and can be used to enforce the occurrence of immediate transitions (which are assigned maximal priority) over timed transitions.
- μ_0 is the initial marking of all states. In the outpatient models depicted all solutions assume a definite initial state, where the assigned number of staff and facilities are available, the clinic is initially open, and no patients are initially in the clinic.

- λ is the partial function defining transition rates where $\forall t \in T, \lambda_t: \mathbb{N}^{|P|} \rightarrow \mathbb{R}^+ \cup [\infty]$ is the rate of the exponential distribution for the firing time of transition t. For infinite rates, the firing time is zero, giving rise to the set of immediate transitions in the set T.
- w is the partial function defining weights assigned to enabled transitions, where $\forall t \in T, w_t: \mathbb{N}^{|P|} \rightarrow \mathbb{R}^+$. When the transition rate evaluates to infinity, the probability taking one transition is given by

$$\frac{w_t(\mu)}{\sum_{t_i: \mu \xrightarrow{t_i}} w_{t_i}(\mu)}$$

- $M = [(p_1, r_1, \psi_1), \ldots, (p_{|M|}, r_{|M|}, \psi_{|M|})]$ is a finite set of measures specifying the computation of a single real value. This imparts the notion of rewards in a stochastic reward net in three components (p, r, ψ). These are respectively, a reward rate accumulated as a function of marking (p), an impulse reward applied at the firing of a specified transition (r), and a function computing a real value from the stochastic process (ψ).

The 11-tuple, combined with the interpretation rules specified in (Gianfranco Ciardo et al., 1993), form the basis of the SRN modeling formalism. Any model specified using SRN has a precise mathematical interpretation as defined by these formalisms,

where the solution can be determined by adhering to the execution policies in the original specification.

2.2.4 Solution in Practice through Discrete Event Simulation

In the preceding discussion each modeling paradigm has been presented assuming that solution of a specified model is possible. Solutions also vary based on dynamic properties of the specified model and desired result. State-space models can be solved for transient or steady-state behavior. Irreducible state-space models, such as most queues, will converge to a steady-state distribution of state probabilities, irrespective of the initial conditions (Cınlar & Sollenberger, 2013; Durrett, 2012; Trivedi, 2002; Trivedi & Bobbio, 2017). However, in many applications a transient solution may be more appropriate, such as clinics with finite operating hours.

Solutions to state-space models depend on the model specified, and result in differing tradeoffs in precision, computational cost, and assurance that improbable events are represented. Solution methods fall under one of three general approaches; analytic solution, analytic-numeric approximation, and discrete event simulation. Discrete event simulation is clearly preferred by most health care modeling applications (Brailsford, 2007; Gunal & Pidd, 2010); however, the other two methods are worth noting for their relative strengths. The solution method, whether analytic, numeric, or simulative, impacts the difficulty of seeking a transient or steady-state solution, but often the application dictates which is needed.

Analytic solution benefits are apparent in cases where closed-form mathematical expressions can be readily derived, such as queues and some queuing networks (Bolch et al., 2006). Analytic methods are sometimes referred to as exact methods (Bolch et al., 2006), which distinguishes this solution method from analytic-numeric methods which approximate ODE solution through a variety of numeric methods, or discrete event simulation which samples the solution through Monte Carlo methods. Closed-form solutions are less computationally expensive when compact forms are available, and the expressions can be further interrogated mathematically. For this subset of models, quality metrics of interest such as the mean number of patients remaining in a queue can be derived directly, however may require significant rederivation if the model specification is changed. The simplest class of state-space stochastic model, homogeneous continuous time Markov chains (CTMC) are desirable for their ease in solution. Steady state solutions for several non-absorbing CTMCs of arbitrary specification can be solved with simple algebraic operations (Trivedi, 2002). Many queuing model solutions are solved under these restrictions (Kleinrock, 1975). In general it is not advisable to pursue analytic solutions of complicated models unless patterns in the model formulation can be exploited in facilitating solutions, such as in product-form queuing networks (Bolch et al., 2006; Kortbeek, 2012), or models with notable generator matrix patterns (Neuts, 1981).

Analytic solutions for transient system behavior in particular require more expertise in solution. Rarely is it effective to solve Markov chains analytically through direct integration. Solution strategies often make use of transform methods such as the Laplace transform to avoid integration (Trivedi, 2002). Relatively innocuous models can generate unwieldy solutions. For even few states, closed form solutions may be impractical if there is little symmetry in the CTMC structure (R. B. Fricks, Bobbio, & Trivedi, 2016) (See Appendix A). Adjusting the structure would also entail rederiving expressions, which may not be practical for all end-users.

A more reasonable approach over analytic solution for most applications is to employ analytic-numeric approximation. Analytic-numeric approaches perform employ numeric methods to find high-accuracy approximate solutions for most CTMC models. A variety of targeted solution methods have been developed for solving continuous time Markov chains (Bolch et al., 2006) , as implemented in software packages such as SHARPE (Sahner, Trivedi, & Puliafito, 1996; Trivedi & Sahner, 2009). While not an exact method, analytic-numeric solutions exhibit minor error when compared to exact solutions (Bolch et al., 2006; Rahul Ghosh, 2012; R. Ghosh et al., 2014).

Particularly in using Petri net formalisms, the main obstacles in seeking analytic-numeric solutions are twofold. First, largeness is a frequent problem in analytic-numeric solution. Recall that Petri net formalisms, for instance, impart largeness tolerance rather than largeness avoidance. Solving a Petri net through analytic-numeric means begins

with constructing a reachability graph (Gianfranco Ciardo et al., 1993), wherein the state space is enumerated in its entirety for analytic-numeric solution as a CTMC. This step will fail in solving basic queues such as M/M/1 or open queuing networks, where the state-space is technically unlimited and thus cannot be fully enumerated. Logical work-arounds exist in either forcing a closed system as was done in (R. B. Fricks & Trivedi, 2016). Alternatively, some Petri net solution packages may include programmatic features such as halting conditions, which limit state space exploration (G. Ciardo, Muppala, & Trivedi, 1989; Hirel, Tuffin, & Trivedi, 2000; Trivedi & Bobbio, 2017). This latter approach deliberately truncates the state-space and is a form of largeness avoidance by effectively ignoring typically rare possibilities (Trivedi & Bobbio, 2017). More elaborate largeness avoidance techniques continue to expand the solvable state space size and is an active area of recent research such as (Rahul Ghosh, 2012; R. Ghosh et al., 2014; Trivedi & Bobbio, 2017). However analytic-numeric solutions of models based on SRN or other Petri net will intrinsically require consideration and application-specific adjustment for largeness.

The second obstacle in applying analytic-numeric solution techniques to SRN models for health care service lays in strict adherence to the SRN formalism. Clinics frequently pace activities around deterministic intervals, such as a lunch break (R. B. Fricks, H. H. Tseng, M. Pajic, & K. S. Trivedi, 2017). Similarly, schedules are templated in rational increments for the benefit of patient and staff. Concretely, clinics may want to

specify schedules such as “two patients every 15 minutes, except for a lunch break from 11AM to 1PM.” It is possible to represent this type of schedule using PH approximations for deterministic transitions, which are derived in (Kleinrock, 1975). For the end-user however, formulating a petri-net representation of a schedule input, for every variation in schedule, is a cumbersome process when rational constraints are applied.

In light of the complexities and limitations of other methods, overwhelmingly researchers have turned to discrete event simulation (DES) (Fone et al., 2003; Gunal & Pidd, 2010; Hulshof et al., 2011; Jun et al., 1999). Authors cite the wide availability of DES packages that are can be solved with conventionally available computers (Gunal & Pidd, 2010). Discrete event simulation naturally extends state-space methods by programmatically enforcing a system state (Haas, 2002; Law & Kelton, 2000). Solutions to the stochastic model and metrics of interest are estimated by iteratively generating realizations of a specified stochastic process, then computing statistics over the realizations (Law & Kelton, 2000). This technique places no restrictions on transition distributions and does not require determining all state permutations a priori, at the cost of requiring several simulative trials to sample the output. Since reachability analysis is not necessary, DES avoids largeness problems altogether. The resulting model response is not an exact solution of the underlying petri net, but rather a sampling of common-case behavior. Advanced sampling schema such as Metropolis-hasting (Durrett, 2012), importance sampling, or Gibbs sampling (Koller & Friedman, 2009; Ross, 1997) may

better represent rare events. Arguably for models of health care service based on imperfect data, rare events may be uninformative when the true goal is improved common-case behavior.

A distinct drawback in DES lays in the variety of implementations without formal specification. How a system is specified or solved varies between simulation programs, which complicates comparing model results. Discrete event simulation can be used to solve SPN or SRN models by implementing SPN rules (Gianfranco Ciardo et al., 1993), demonstrated in ICU models (R. B. Fricks & Trivedi, 2016), and relaxing Markovian restrictions to model specification (Alfonso et al., 2012; R. B. Fricks, H. Tseng, M. Pajic, & K. S. Trivedi, 2017). This hybrid approach uses SRN to specify DES models with precision, while avoiding largeness problems and allowing for more open schedule specification. The hybrid approach employed by this dissertation work uses DES to solve complex SRN models, and is introduced in Section 2.4.

2.2.5 Validating Models for Generalization Performance

Applications of the stochastic modeling methods in the preceding discussion have traditionally relied on goodness-of-fit techniques and hypothesis testing for validating models when measurements are available. These statistical tests and validation techniques are described in various texts (D'Agostino & Stephens, 1986; Rosner, 2011; Trivedi, 2002). When validation is performed, these tests are applied to evaluate if a model is a suitable representation of the data set used to parameterize and

formulate the model, with examples of this practice in (Alfonso et al., 2012; Hribar et al., 2018; Hideaki Takagi et al., 2017). Hypothesis tests are used to approve a modeling procedure, rather than to quantify predictive performance.

The use of cross-validation in other disciplines has shown that separating the available data into independent sets in some manner are vital in preventing overfitting (Hastie et al., 2009). When such evaluation is not performed, such as if the same training data is used for model construction and model validation, improvements in model training performance may lead to loss of generalization performance or added model complexity with no appreciable performance improvement. In general, the training error will optimistically underestimate the model error. Evaluations based on training error alone may substantially overestimate the model goodness-of-fit (Hastie et al., 2009).

The validation methods in use in health care literature are in sharp contrast to more recent developments for incorporating data and evaluating new models for generalization performance (Breiman, 2001; Fenton & Neil, 2012; Goodfellow, 2016; Hastie et al., 2009). Cross-validation techniques allow for quantitative selection between potentially several model alternatives based on their expected performance on new data, a far more rigorous conception than previously used in validation. Evaluating the generalization performance of a given model estimates the predictive value of that model (Hastie et al., 2009). This section introduces adaptations of cross-validation for evaluating predictive performance of clinic flow models.

Consider the duration of patient visits to a clinic, denoted by Z. Patient i is said to take z_i time to complete treatment. Stochastic modeling approaches treat Z as a random variable with unknown true distribution $F_Z(t)$. The distribution of Z may depend on several factors, and alternative modeling processes may present different hypotheses for $F_Z(t)$ through abstraction from data and system details (R. B. Fricks et al., 2018). Any hypothesized model $H_{\hat{Z},\Theta}(t)$ with parameters θ to estimate the true distribution of visit durations $F_Z(t)$ from a set of empirical measurements $\hat{Z}$ can be evaluated quantitatively by defining a loss function L with respect to the empirical distribution computed from a sample, $\hat{F}_{\hat{Z}}(t)$. This evaluation selects the absolute error function as the loss measure

$$L\left(\hat{F}_{\hat{Z}}(t), H_{\hat{Z}}(T)\right) = \left|\hat{F}_{\hat{Z}}(t) - H_{\hat{Z},\Theta}(t)\right|$$

The expected or average absolute error for a model formulated from a given data set $\hat{Z}$ of M observations occurring at time t_m, where $m = 1, 2, 3 \ldots M$ is

$$Err_{\hat{Z}} = E\left[L\left(\hat{F}_{\hat{Z}}(t), H_{\hat{Z},\Theta}(t)\right)\middle|\hat{Z}\right] = \frac{1}{M}\sum_{1}^{m}\left|\hat{F}_{\hat{Z}}(t_m) - H_{\hat{Z},\Theta}(t_m)\right|$$

Where both $\hat{F}_{\hat{Z}}(t), H_{\hat{Z},\Theta}(t)$ are assumed to be sampled from continuous distributions, and $H_{\hat{Z},\Theta}(t)$ is defined at all M points where $\hat{F}_Z(t)$ is measured. We can now distinguish between **parameterization** and **model selection**, using this notation. **Parameterization** is the process of optimizing the hypothesized model to minimize the expected error with respect to its training set. For a model with a q-dimensional set of parameters θ, this training error can be written as

$$\min_{\theta \in \mathbb{R}^q} Err_{\hat{Z}} = E\left[L\left(\hat{F}_{\hat{Z}}(t), H_{\Theta,\hat{Z}}(t)\right) \middle| \hat{Z}\right]$$

For various models, it can be assumed the parameterization is possible and subsequently use $H_{\hat{Z}}(t)$ to refer to the optimally parameterized variation of model H with respect to $\hat{Z}$, omitting the parameter set θ. Presented with a set of optimally parameterized models **H**(t), irrespective of formulation, **model selection** is selecting the model that minimizes the expected loss function with respect to a new sample, or conversely optimizes generalization performance. To motivate the need for cross-validation in quantifying generalization performance, note that minimal error can be trivially achieved in the above expression by setting $H_{\hat{Z}}(t) = \hat{F}_{\hat{Z}}(t)$, effectively using the empirical distribution function of one data sample for prediction. Consequently, the model selection procedure thus far is prone to overfitting. Model selection can now be specified more formally using cross-validation. Suppose we have two data samples $\widehat{Z_1}$ (size M_1) and $\widehat{Z_2}$ (size M_2) independently and randomly sampled from the same population. If we denote $\widehat{Z_1}$ the training sample, used for formulating and parameterizing the model we can designate $\widehat{Z_2}$ the validation sample, used for evaluating the model's ability to generalize. The training error and validation error are thus defined as

$$Err_{trn,(\widehat{Z_1},\widehat{Z_1})} = \frac{1}{M_1}\sum_{1}^{m_1}\left|\hat{F}_{\widehat{Z_1}}(t_{m_1}) - H_{\widehat{Z_1}}(t_{m_1})\right|$$

$$Err_{val,(\widehat{Z_1},\widehat{Z_2})} = \frac{1}{M_2}\sum_{1}^{m_2}\left|\hat{F}_{\widehat{Z_2}}(t_{m_2}) - H_{\widehat{Z_1}}(t_{m_2})\right|$$

A model generalizes well if the error is consistent when presented with a new sample, i.e.

$$0 \leq Err_{trn} \approx Err_{val}$$

Consider this error as a function of model complexity, generically referred to as χ. As more details are represented via a more complex model it is expected (Breiman, 2001)

$$\lim_{\chi \to \infty} Err_{trn}(\chi) \to 0, Err_{val}(\chi) \to 0$$

Given three successively more complex models at complexities χ_1, χ_2, χ_3, where $\chi_1 < \chi_2 < \chi_3$, overfitting is characterized by a progression such as

$$Err_{trn}(\chi_1) \geq Err_{trn}(\chi_2) \geq Err_{trn}(\chi_3)$$

$$Err_{val}(\chi_1) \geq Err_{val}(\chi_3) \geq Err_{val}(\chi_2)$$

Notably there is not a monotonic decrease in validation error, even as training error may continue to decrease. Cross-validation is used to find the hyperparameter set that minimizes validation error, given that parameters already minimize training error. In contrast to parameters, hyperparameters are model features optimized with respect to the generalization performance. Measures of complexity, such as the number of phases in a phase-type distribution, are frequently optimized as a hyperparameter.

Lastly, we wish to estimate the model performance unconditional to the training or validation sets, known as expected prediction error, denoted as

$$Err = E[Err_{\hat{\chi}}]$$

This error estimates model performance on an independent data by averaging expected error over several data sets, minimizing the influence of set selection on the error estimate. The k-fold cross-validation technique (Goodfellow, 2016; Hastie et al., 2009) provides an estimate of expected prediction error by subdividing the data set $\hat{Z}$ into k equally sized random samples. By dividing the data set into k subsets, we evaluate the validation error k times by computing validation error when the kth fold is used as a validation set, and all other subsets are used as a training set for parameterizing the model:

$$Err = E[Err_{\hat{Z}}] = E\left[\frac{1}{M_k}\sum_{1}^{m_\kappa}\left|\hat{F}_\kappa\left(t_{m_\kappa}\right) - H_{-\kappa}\left(t_{m_{-\kappa}}\right)\right|\right]$$

Where $-\kappa$ above indicates the set of data made by excluding data in fold κ using notation from (Hastie et al., 2009). We can now select the hyperparameter configuration that minimizes expected prediction error, and therefore provides optimal generalization performance. Finally, it is possible to set a validation threshold by defining a minimum acceptable error threshold, though not strictly necessary.

2.3 Operations Research in Health Care

Model-based quality improvement in health care dates back to seminal papers by Welch and Bailey in 1952 (Welch & Bailey, 1952). Since then pivotal reviews have covered the proliferation of discrete event simulation (Fone et al., 2003; Gunal & Pidd, 2010; Jun et al., 1999) and analytic approaches (well summarized in (Kortbeek, 2012)) to

model health care. Similarly, optimization of outpatient scheduling and appointment systems is covered in (Ahmadi-Javid et al., 2017; Cayirli & Veral, 2003). Reviews cite an abundance of similar applications in the literature (Gunal & Pidd, 2010), with limited technical innovation or comparison of methods.

Efforts tend to focus on the steady state solutions of inpatient wards (Brailsford, 2007; R. B. Fricks & Trivedi, 2016; McManus et al., 2004; Hideaki Takagi et al., 2017). The steady state solution for clinics with finite hours is typically uninformative; all patients eventually vacate the clinic during standard hours. For the outpatient setting, transient solutions are necessary to accurately represent clinic status as patients arrive and leave during finite hours of operation (R. B. Fricks et al., 2017). Regardless of the application setting, some key features for representing health service are consistent and dictate partially the choice of model formalism. Consistently, the approaches feature stochastic arrival times and service times for patients; constraints on resources for treatment; and service flows where patients queue for a variable number of treatment phases. Queues and queuing networks have been shown to reasonably model service particularly in inpatient treatment departments (Griffiths, Price-Lloyd, Smithies, & Williams, 2005; McManus et al., 2004; Hideaki Takagi et al., 2017; H. Takagi et al., 2014; van Dijk & Kortbeek, 2009).

Health care modelers frequently use a variety of causal distributions to represent the duration of treatment. Exponentially distributed transitions are highly desirable for

their analytic properties (Gianfranco Ciardo et al., 1993; Trivedi & Bobbio, 2017). Markovian models have seen several applications in health care, such as in (R. B. Fricks & Trivedi, 2016; Kortbeek, 2012; McManus et al., 2004; Hideaki Takagi et al., 2017; van Dijk & Kortbeek, 2009). Often however the exponential distribution and associated assumptions may be an inadequate representation for treatment durations. Other distributions may be more accurate or easier to use, for instance the uniform distribution is sometimes used for its ease of parameterization (Alfonso et al., 2012). A variety of named distributions such as Weibull or lognormal have seen use (e.g. (R. B. Fricks et al., 2017)) but provide few parameters for model refinement. Alternatively, phase-type distributions can fit a stochastic process to an arbitrary degree of precision (Fackrell, 2008; Kleinrock, 1975; Marshall et al., 2005; Trivedi & Bobbio, 2017). Several tools are now readily available for fitting complex stochastic processes (Horvath & Telek, 2017; Okamura, Dohi, & Trivedi, 2011; *Principles of Performance and Reliability Modeling and Evaluation [electronic resource] : Essays in Honor of Kishor Trivedi on his 70th Birthday*, 2016; Trivedi & Bobbio, 2017) such as phase-type distributions, Markovian Arrival Processes, or Random Arrival Processes (Horvath & Telek, 2017). Traditionally most Petri net formalisms struggle with non-exponentially distributed transition times, which in turn requires modelers to either restrict the use of non-exponential transitions (Choi, Kulkarni, & Trivedi, 1994; Trivedi & Bobbio, 2017) or solve the net using discrete event simulation (Alfonso et al., 2012; R. B. Fricks et al., 2017).

Currently, there is no consensus on stochastic modeling approaches for clinic flow, particularly in high-volume outpatient settings. Researchers have proposed a variety of analytic approaches in modeling health care service, ranging from queuing networks (Kortbeek, 2012; Hideaki Takagi et al., 2017; van Dijk & Kortbeek, 2009) to stochastic Petri nets (Alfonso et al., 2012; R. B. Fricks & Trivedi, 2016; R. B. Fricks et al., 2018; R. B. Fricks et al., 2017; Kortbeek, 2012; Leite et al., 2010) to discrete event simulation (Fone et al., 2003; R. B. Fricks et al., 2018; Gunal & Pidd, 2010; Hribar et al., 2016; Hribar et al., 2018; Hribar et al., 2015; Jun et al., 1999). Efforts tend to focus on the steady state solutions of inpatient wards (Brailsford, 2007; R. B. Fricks & Trivedi, 2016; McManus et al., 2004; Hideaki Takagi et al., 2017). In clinics with finite hours, all patients eventually vacate the clinic. For the outpatient setting, transient solutions are necessary to accurately represent clinic status as patients arrive and leave during finite hours of operation (R. B. Fricks et al., 2017). Regardless of the application setting, some key features for representing health service are consistent and dictate partially the choice of stochastic model. Consistently, the approaches feature stochastic arrival times and service times for patients; constraints on resources for treatment; and service flows where patients queue for a variable number of treatment phases.

Although underused, Petri Net formalisms provide an effective paradigm for modeling stochastic systems where concurrent use of resources is a key feature (Gianfranco Ciardo & Trivedi, 1993; R. B. Fricks & Trivedi, 2016; R. B. Fricks et al., 2017;

Haas, 2002). Since SRN allows for a concise description of a clinic as a graph with unambiguous mathematical interpretation (Gianfranco Ciardo et al., 1993), using SRN facilitates comparison between models, providing a standard language for describing complex clinic systems. Current use of Petri net models in the literature is scarce (Alfonso et al., 2012; R. B. Fricks & Trivedi, 2016; R. B. Fricks et al., 2018; R. B. Fricks et al., 2017; Kopach-Konrad et al., 2007; Kortbeek, 2012; Leite et al., 2010), with rare adherence to a strict formalism.

While the use of discrete event simulation (DES) for clinic planning has long been documented in health care (Fone et al., 2003; Gunal & Pidd, 2010; Jun et al., 1999), secondary use of EHR alleviates earlier difficulties in data collection. Previous studies often relied on manually collected data (Alfonso et al., 2012; Welch & Bailey, 1952), parameters estimated through population statistics (Adan, Bekkers, Dellaert, Vissers, & Yu, 2008; Christodoulou & Taylor, 2001; McManus et al., 2004), or expert input (Leite et al., 2010). More recent studies benefitted from secondary use of EHR data in simulating patient flow through clinics (R. B. Fricks et al., 2018; R. B. Fricks et al., 2017; Griffiths et al., 2005; Hribar et al., 2016; Hribar et al., 2018; Hribar et al., 2015; Hideaki Takagi et al., 2017; H. Takagi et al., 2014), although approaches to data handling or model validation vary. Sample size in most cases appears constrained by feasibility, where however much data is attainable is used. Since various methods for validating models factor sample size into quantitative metrics (D'Agostino & Stephens, 1986; Goodfellow, 2016; Hastie, 2009),

sample size alone may have greater-than-expected influence on model selection and robustness. The ability to achieve consistency in various estimators, in various clinics, at similar sample size demonstrates that the moment-based approach may apply to other practices to empirically determine when collected samples are adequate.

Many approaches to clinic optimization implicitly or explicitly represent visit durations using stochastic methods (Cayirli & Veral, 2003; R. B. Fricks & Trivedi, 2016; R. B. Fricks et al., 2018; R. B. Fricks et al., 2017; Goodfellow, 2016; Hribar et al., 2016; Hribar et al., 2018; Hribar et al., 2015; Kortbeek, 2012; Marshall et al., 2005; McManus et al., 2004; Hideaki Takagi et al., 2017; H. Takagi et al., 2014; van Dijk & Kortbeek, 2009). Perspectives on how to predict future behavior of a measured stochastic system have seen considerable development in recent decades (Breiman, 2001; Goodfellow, 2016; Hastie, 2009; Hastie et al., 2009; Trivedi & Bobbio, 2017).

2.3.1 Quantifying Decision Tradeoffs

The advantage of white box model-based analysis of health care operations is in the ability of models to quantify the effect of many variables on a decision tradeoff under potentially overwhelming uncertainty. In (R. B. Fricks & Trivedi, 2016), the model in Figure 1 was proposed for evaluating intensive care unit (ICU) performance when dependent on fallible medical devices.

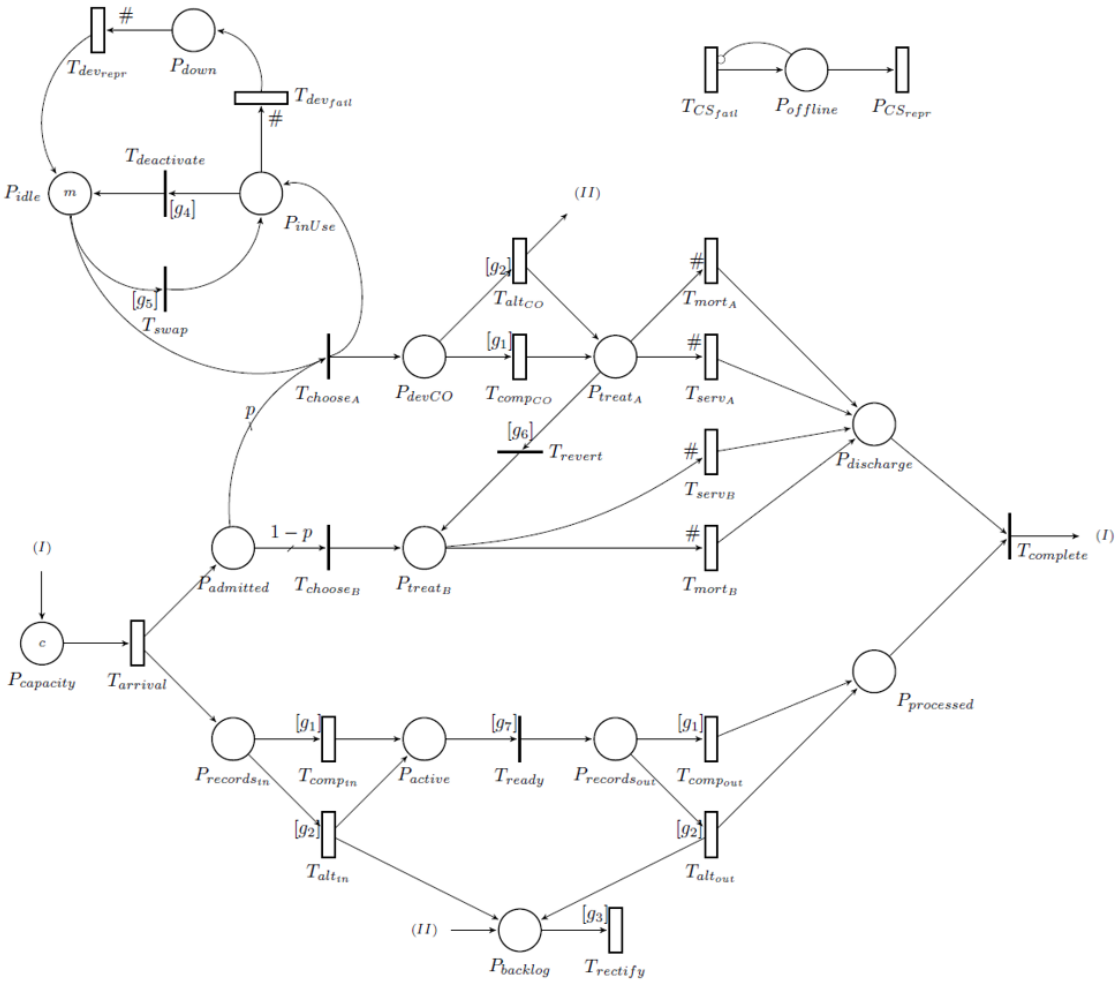
Various studies have validated the use of variations on the Erlang loss (EL) queuing model in describing patient flows in an independent ICU (Griffiths et al., 2005;

McManus et al., 2004). The EL model is a M/M/c/c queue in Kendall's queuing notation (Kleinrock, 1975): a queue experiencing Poisson arrivals; independent, exponentially distributed service times; first-come, first-served scheduling discipline; and finite capacity c with dedicated servers for each accepted entry.

The model in Figure 1 expands traditional ICU queuing models using the SRN formalism to explicitly represent treatment options, case mix, administrative record keeping, and reliance on medical devices and EHR infrastructure. This model was detailed in (R. B. Fricks & Trivedi, 2016). Several dependencies of intensive care service are modeled in general form to illustrate how modeling extensions in SRN can be used to describe the potential for complex control in health care delivery.

Notable in this model, a place labeled $capacity$ is designated which enforces bounds on reachability; at any moment there can be no more than c tokens in total between the $capacity$ place and the places $(admitted, devCO, treat_A, treat_B, discharge)$ as specified. This construct is a means of largeness avoidance.

Even though this model adheres to strict SRN specification, using only exponentially distributed transitions, the solution is problematic. The moderate number of places combined with the potential for many tokens specified by the allotted capacity (c) and number of devices (m) results in immediate largeness problems when realistic values for ICU capacity are represented.

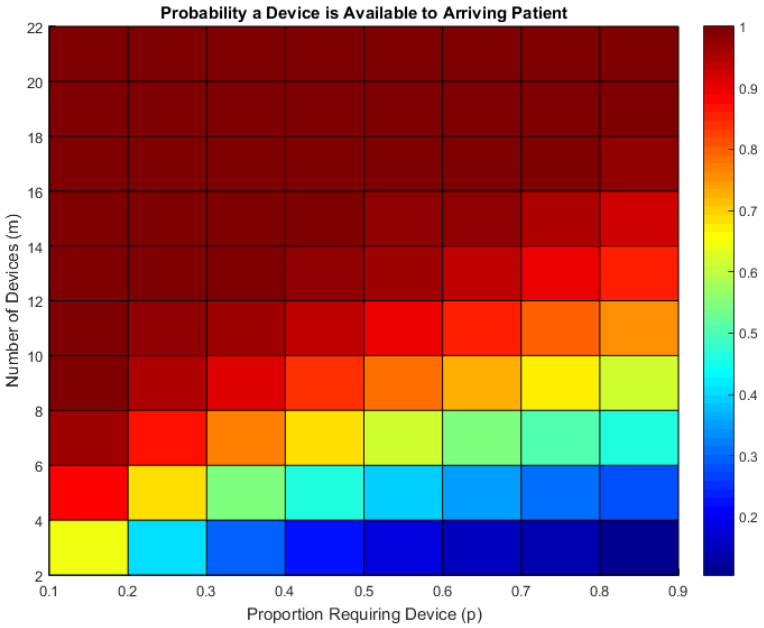


In this SRN two treatment paths are available, contingent on available medical devices which initially are represented by tokens in the idle place. Based on an elementary queue model, device failure, record keeping, and reliance on a central computer are represented to fully showcase modeling capability using SRN. For a complete description, see detail in (Fricks & Trivedi 2016).

Figure 1: SRN Model for ICU Treatment.

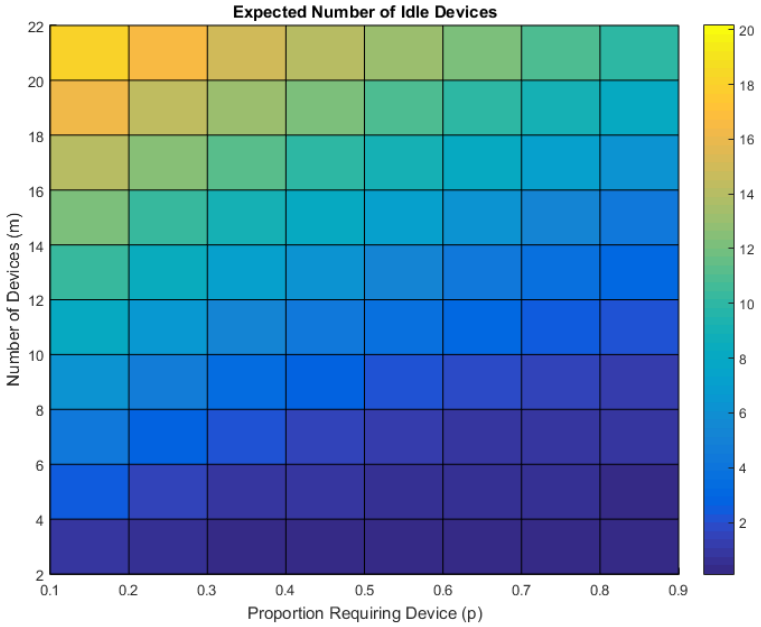
Simulating this SRN is unproblematic however, and the model is readily simulated for the equivalent of 10,000 years of ICU operation. Computing the steady-state probability that a device is available (Figure 2) and average number of idle devices (Figure 3), it is apparent in this system that there is a reciprocal relationship between

ensuring a device is available and having many copies of that device idling. The use of the generic term device is deliberate; the optimal tradeoff between overstocking and unavailability depends on the clinical needs and device in question.



Steady state probabilities for when the device is available are shown as the probability of requiring the device (p) and number of devices available are varied relative to a fixed ICU capacity of 20.

Figure 2: The probability a device is available for treatment path A when needed.



The expected number of idle devices as the number of intially available devices (m) and probability the device is needed (p) are varied relative to fixed parameters and ICU capacity of 20. Ensuring a device is available with high probability (Figure 2) has an inevitable tradeoff in raising the average number of idle devices. Such tradeoffs are recurrent in quality improvement decisions.

Figure 3: The average number of idle devices.

3. Understanding Outpatient Care Process

Models of health care service must accurately reflect clinical procedure. This chapter describes the practice environment where data were collected for modeling purposes. Considerable attention is devoted to decomposing visits to an outpatient clinic into constituent treatment steps. The result is a general description of standard practices in the clinics studied with a focus on glaucoma specialties. The discussion emphasizes the granularity required for measuring performance and modeling operations.

For ease and efficiency, visits to a clinic tend to follow an orderly pattern. The SOAP method (Willis, 2008) for record keeping is insightful in understanding how patient visits are managed and documented. SOAP stands for

Subjective: symptoms, as reported by the patient

Objective: observable signs, including test results and health data

Assessment: evaluate the patient's status and effectiveness of current treatment (if any), note any newfound problems or diagnoses

Plan: Subsequent actions, such as prescribing medications, procedures, planning surgery, or return visits.

SOAP provides a rough outline of clinical practice for non-clinicians. First technical staff examine the patient, noting intangible symptoms such as sensations reported by the patient and measuring signs like intraocular (eye) pressure. A physician may then further examine the patient and interpret any previously collected information

to determine the next course of action. This process is referred to as evidence based medicine (Masic, Miokovic, & Muhamedagic, 2008), and incorporates years of clinical experience and the latest research in the assessment and plan stages of SOAP. For operational decisions it is often sufficient or necessary to simplify clinical process into observable steps and resource requirements.

The challenge for health care organizations is ensuring all data, including patient measurements and tests, are efficiently recorded and available during the patient's visit. SOAP does not consider capacity or process questions such as where measurements are performed, by whom, and how long each step may last. Nonetheless the SOAP steps are useful in understanding the correct order of procedures and verifying the specification of fine grain models.

3.1 Procedural description of glaucoma practice

The objective in this section is to delineate the steps in a visit to a glaucoma practice, at a granularity conducive to measurement and model formulation. For the purposes of modeling clinic flow, visit time is considered as the time from check-in to check-out. Appointment times and arrivals, which occur prior to check-in, are discussed in later sections.

Various process diagram conventions are in use broadly. Flowcharts require little specialized knowledge, but have been used in understanding business process, developing software, and planning discrete event simulation algorithms (Law & Kelton,

2000). It is sensible to apply a relatively universal method that corresponds to algorithmic descriptions. Additionally, practitioners at the observed clinic have also adopted flowcharts for mapping common operations, such as procedural aspects of receiving patients at the clinic. Flowcharts are accessible to health care professionals with little additional training. The accessibility, universality, and correspondence of flow charts indicates their use for diagramming patient flow. A patient flow diagram will identify treatment phases, which can be measured and statistically characterized in subsequent steps. Glaucoma practice, according to early data exploration of the timestamped events in the electronic health records, proceeds in a specific pattern in most cases, illustrated in Figure 4.

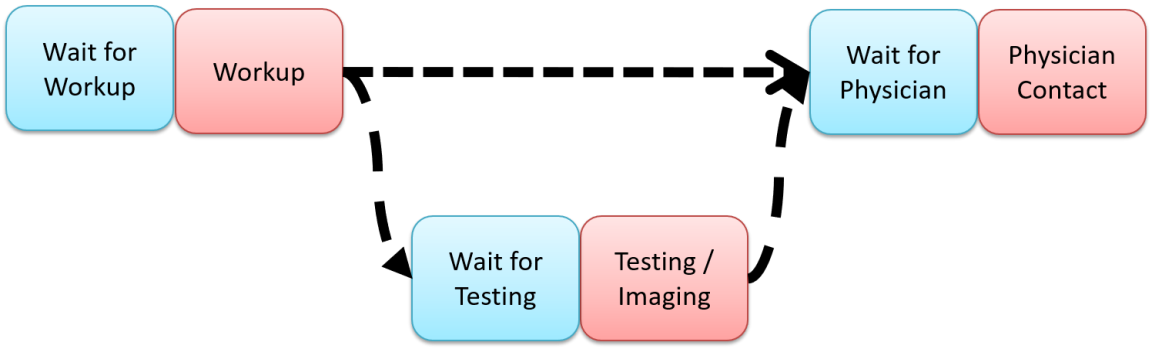


Figure 4: General steps in a visit to a glaucoma clinic, with expected waiting periods explicitly marked.

These three steps constitute patient visits to a glaucoma clinic, described at a granularity that emphasizes demands on clinical resources; a workup requires a technician, imaging or testing requires a specialized technician assigned to the required medical device, and patients wait for their doctor to finish the visit with physician

contact. These steps are now further elaborated in terms of SOAP (Willis, 2008), resource requirements, and finer grain procedural terminology.

3.1.1 Initial Step: Workup

After check-in, the first step in treatment is referred to as the workup. In large ophthalmology practices workups are performed by a technician. Several technician certifications are recognized by the Joint Commission on Allied Health Personnel in Ophthalmology; workups are predominantly performed by certified ophthalmic assistants (COA) or certified ophthalmic technicians (COT) (International Joint Commission on Allied Health Personnel in Ophthalmology (IJCAHPO), 2018). Use of the term "technician" refers to personnel with either certification in this discussion.

A technician will locate the next patient in the waiting area then guide the patient and any family members to an examination room for a workup. Workups begin with a patient interview where the technician confirms the patient's identity, medical history, existing prescriptions, and the reason for a visit (sometimes referred to as the primary complaint (Willis, 2008)). These first questions among others primarily collect subjective information and are registered in the electronic health record system at point-of-care (PoC) stations located in examination rooms (Figure 5).

Figure 5: Technicians perform the initial patient interview, confirming the current medications and medical history among other questions

Next, several objective tests may be performed (Figure 6). Technicians at the practice observed have required proficiency in the following standard tests:

- Visual acuity testing (standard vision tests)
- Refraction (manifest or automated)
- Lensometry (manual or automatic)
- Retinoscopy
- Pupil examinations
- Extraocular Muscle (EOM) or Sensorimotor exams

- Confrontation Visual Field (CVF)
- Slit-lamp testing
- Tonometry (eye pressure measurement)
- Drop instillation (such as pupil dilation eye drops)

This list covers many common procedures, but is not exhaustive. Other advanced tests may also be performed by technicians such as Pachymetry, brightness acuity testing, or Schirmer testing.

Figure 6: Technicians may also perform some eye measurements

Some tests are relatively standard in an ophthalmic visit, such as visual acuity testing (Figure 7). Other tests may be previously ordered, or necessary for certain referral or diagnoses. For instance tonometry and pachymetry are key in diagnosing and monitoring glaucoma (National Institutes of Health (NIH), 2015).

Figure 7: Acuity testing is a staple of ophthalmology visits, performed during workups using a mirrored setup that increases the path length to vision chart.

Many permutations of workup are possible by combining the available tests to meet patient needs or the purposes of a visit. From an operational standpoint however workups require a technician, an examination room, and serve the primary purpose of collecting subjective and objective information from the patient.

Once a workup is completed, a patient either proceeds to testing or imaging (3.1.2) or visit with the physician (3.1.3). The technician may move the patient to a waiting area, another exam room, or the patient may wait in the exam room used during workup, depending on the practice and patient needs.

3.1.2 Intermediary Step: Testing or Imaging

Additional tests including imaging may be necessary beyond the workup (Figure 8). What distinguishes the next step, Testing or Imaging, from workups are that these tests require specialized equipment that is stationary in a dedicated testing or imaging room. This step is still conducted by a technician, although some facilities may devote technicians to specialization with one type of test or imaging station.

Testing or Imaging may refer to:

- Optical Coherence Tomography (OCT)
- Optic disc photography
- Stereo photography
- Visual Field Testing (Humphrey, Octopus, Goldmann varieties)
- Optical Biometry
- Corneal Topography

A larger variety of specific orders can be addressed, such as ordering contrast agents, or specific imaging protocols such as autofluorescence or angiography. Many enumerable permutations of orders are possible to fit the medical needs of a patient. From an operations standpoint this step collects more objective measurements, requiring a technician and specialized imaging equipment or room, and is typically performed prior to evaluation by a physician.

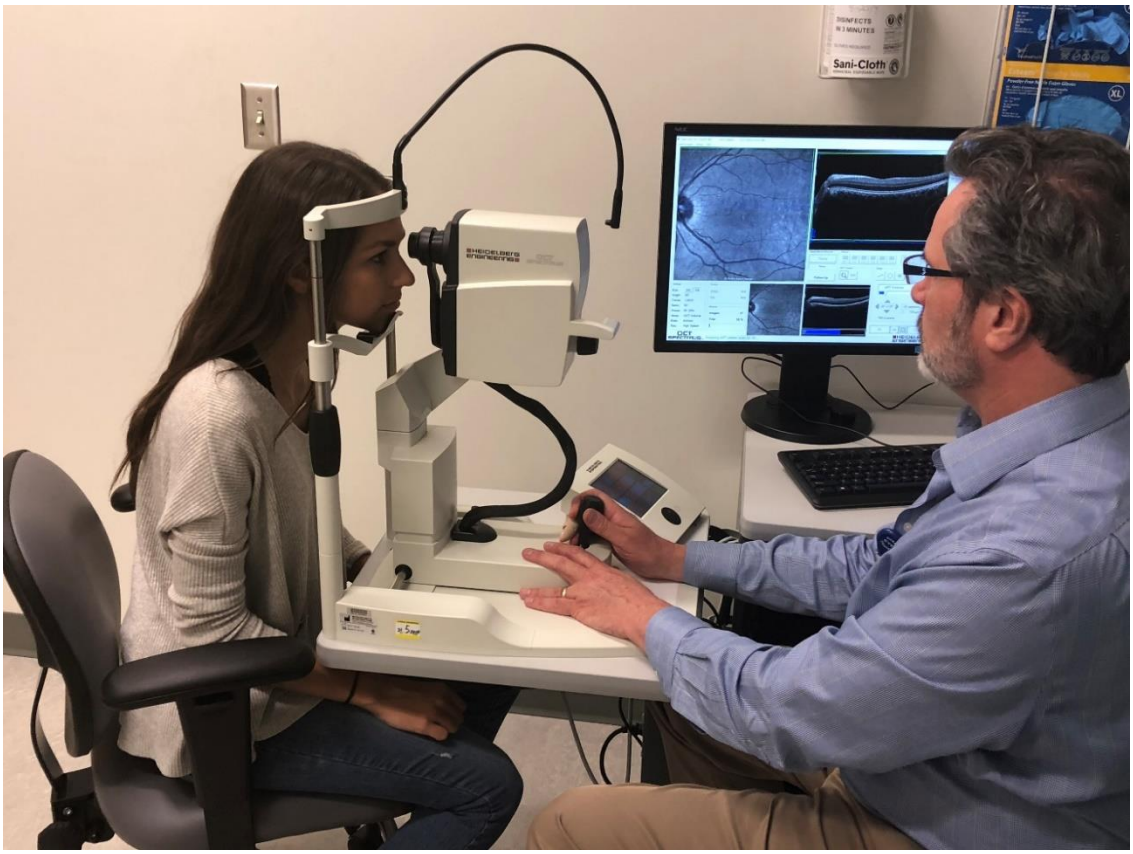


Figure 8: A variety of imaging procedures and testing may be required to monitor patient health.

3.1.3 Final Step: Physician Contact

The final step in a clinic visit is contact with the physician. Once referred, one doctor oversees a patient's care in the physician's specialty, and patients will visit the same doctor during each appointment. Each doctor often sets the pace for his or her clinic; additional technicians and equipment can be added to service a larger volume of patients in a given clinic, however all patients eventually see the same doctor.

During this last step, a doctor will usually follow up on previous visits as well as information collected earlier in the visit. A doctor may clarify issues raised by the patient, perhaps revisiting the medical history with more targeted questions. Conversations with the patient may raise vital subjective information. Additionally, a doctor frequently performs an independent examination to collect or confirm additional objective information (Figure 9). This examination may be more specific than previously or involve specific tests that require a trained physician.

After evaluating all available information, a doctor will assess the patient's health status and determine a care plan. The transition through the third SOAP step is often not be externally observable. The assessment incorporates clinical experience, research findings, and patient specifics, and is beyond the scope of most interpretable modeling or simulation. Rather it is useful to consider the possible decisions as generalized categories of treatment plan.

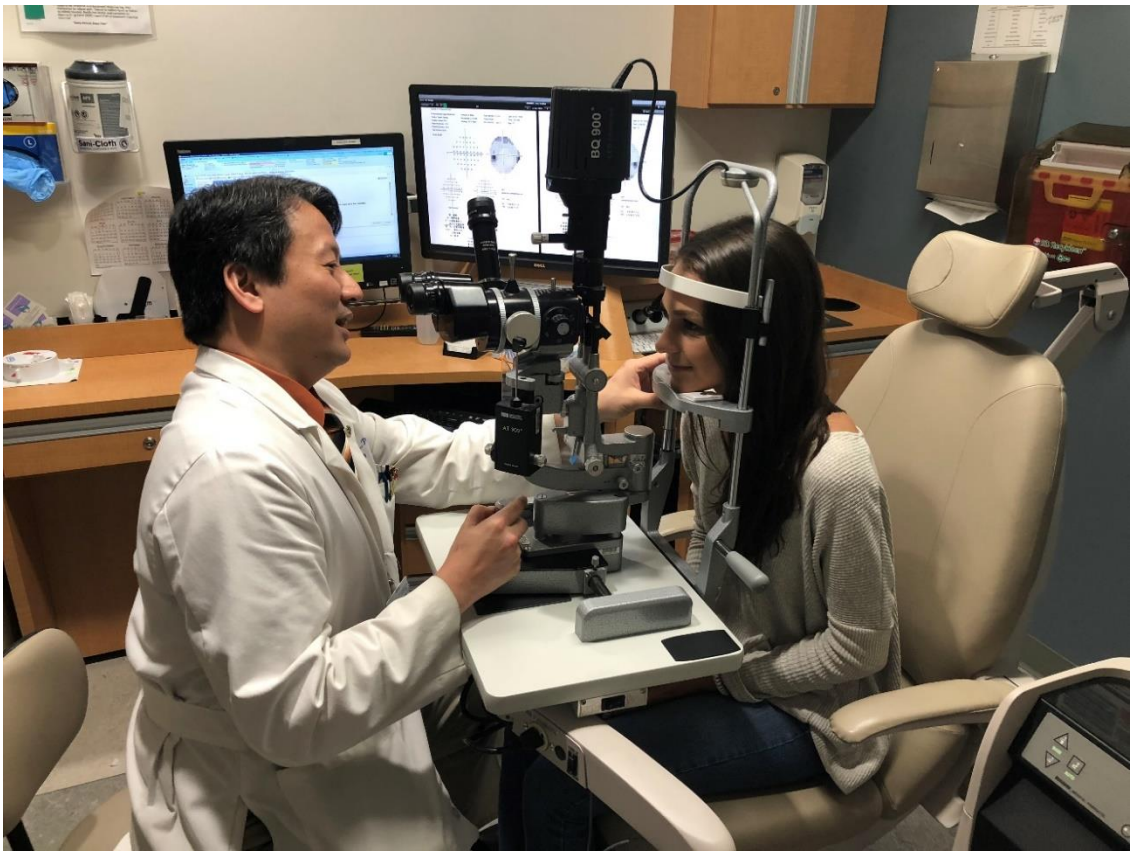


Figure 9: A doctor may perform an independent examination of the patient.

The physician will form a treatment plan based on patient needs. Through interviewing doctors, this may involve ordering additional tests; scheduling a procedure or surgery (3.1.4); referring a patient to another doctor; assigning a treatment plan such as an eye drop regimen for the patient to adhere to at home; or plan follow-up visits. Clear communication between doctor and patient is critical in ensuring high quality patient outcomes (Ha & Longnecker, 2010), and a doctor may spend a considerable portion of the visit discussing treatment options with the patient or ensuring that directions are well understood (Figure 10).

Figure 10: Patients may spend a considerable portion of the visit with the physician discussing a treatment plan.

From an operational standpoint, this final step requires an examination room and the physician. Visits typically conclude after this step from a system standpoint, as most patients are ready to check out afterwards and no longer require facilities, personnel, or other resources. Treatment plan options that require clinical personnel, such as procedures or surgeries, are typically scheduled outside of typical clinic days and are discussed in section (3.1.4).

3.1.4 Miscellaneous and Additional Treatment Steps

Most patient visits to a physician on clinic days follow the common case behavior in Figure 4. Some alternative steps in patient treatment do not significantly impact online operational performance, either due to unique personnel or resource requirements, or infrequent occurrence. These additional steps are discussed in this subsection.

Visit with Resident or Fellow: At academic institutions, part of the mission is training new physicians in residency or fellowship programs. Physician trainees may conduct an additional examination, either independently or under the supervision of an attending physician. These examinations are usually conducted after workup but before contact with the attending physician. Trainees handle tests requiring specialized expertise, such as gonioscopy in glaucoma practices. Trainees may also be qualified to perform procedures. Trainee-led procedures on clinic days has the added benefit of expediting clinic flow, as procedures may be time consuming. The attending physician determines the scope of each of his or her assigned trainee's responsibilities and independence.

Surgery: some ophthalmic conditions require surgical intervention. In glaucoma, examples include surgical trabeculectomy which may be performed to alleviate intraocular pressure (IOP) (National Institutes of Health (NIH), 2015). Alternatively, a shunt may be placed to manage IOP. Lens replacement is also routinely

performed to treat cataracts, which is comorbid with conditions such as glaucoma (Pham, Wang, Rochtchina, Maloof, & Mitchell, 2004). Surgery requires dedicated facilities, and personnel. Physicians who perform surgeries typically schedule surgeries on non-clinic days. Scheduling itself may occur on clinic days, but usually after discussing options with a physician during the visit. An independent scheduling office coordinates with a patient after they have effectively checked out of the clinic.

Procedure: in the health system studied, procedures refer to more minor interventions with fewer facility requirements than surgery. An example in glaucoma is laser trabeculoplasty (National Institutes of Health (NIH), 2015). Procedures can be performed during clinic days in dedicated procedure rooms, as opposed to an operating room. Procedures can be performed with simpler aseptic requirements when appropriate rather than the sterile requirements for surgery. In ophthalmology the distinction between surgery and procedure is further blurred as many surgeries are performed on an outpatient basis, and the eye is accessible without incision. In the practices studied, procedures may occur on clinic days, and may be performed by fellows.

Re-entrant patients and emergency visits: visits to specialty practices in ophthalmology are primary by referral; clinics do not typically treat emergency or critical cases. While some clinics offer a limited number of urgent appointments, an

appointment schedule is rarely modified by an online operational decision in the current practice.

Similarly, step repetition is rarely noted. Workups are repeated typically if an oversight occurred. Additional imaging or testing can be scheduled and evaluated by the physician outside of clinic days. Patients do not re-enter the clinic on the same day.

Miscellaneous: additional patient services are provided, such as speaking to a financial counselor or alternative check-out procedures. These meetings are infrequently recorded and occur outside of the core clinical visit.

3.2 Description of Duke Eye Center

The procedural description of clinic practices is based on observation at Duke Eye Center (DEC). Features of DEC further motivate the need for model-based analysis, as well as modeling decisions critical to accurate representation.

Several ophthalmic clinics run concurrently at the DEC, ranging from general eye care to specialized clinics focused on retinal diseases, glaucoma, or ophthalmic oncology. Over 85,000 patients visit DEC clinics annually, receiving treatment in as many as fifteen ophthalmic specialties (Duke University Health System (DUHS), 2018). Four specialty departments in Glaucoma, Vitreous Retinal, Cornea, and Comprehensive Ophthalmology account for the majority of patient visits and have relatively well-defined case mixes. Together these four specialties generated 96,360 visit records attributed to 80 physicians on record from January 2016 to July 2017.

DEC was designed as a multi-service center for ophthalmic specialties, where clinical practices share on-site imaging and testing facilities. Several clinics may operate concurrently, sharing available resources and personnel. Figure 11 shows designated areas for glaucoma and retina practices on the second floor (ground level), which share check-in and imaging areas placed in between department spaces. Similarly, a third floor houses exam rooms for other departments such as Cornea and operating rooms (Figure 12). These plans reflect strategic planning as of August 2015, yet are largely consistent with current use of the facilities.

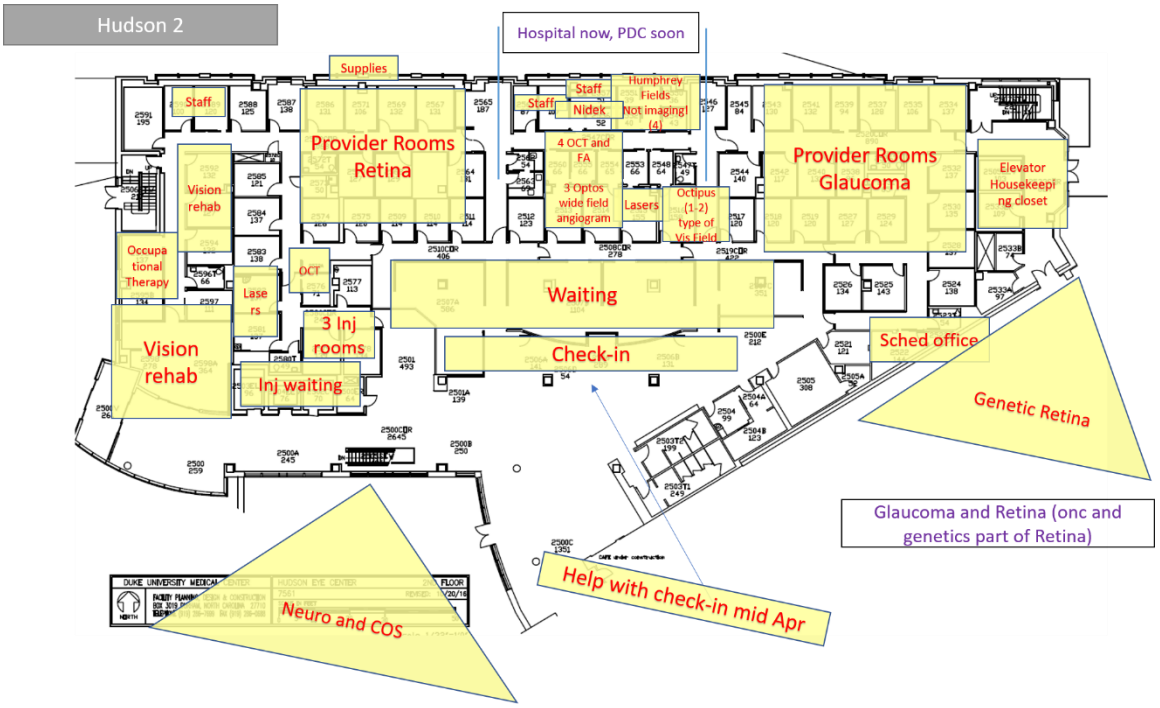


Figure 11: The eye center was dimensioned with various concurrent practices in mind, as well as on-site imaging and testing.

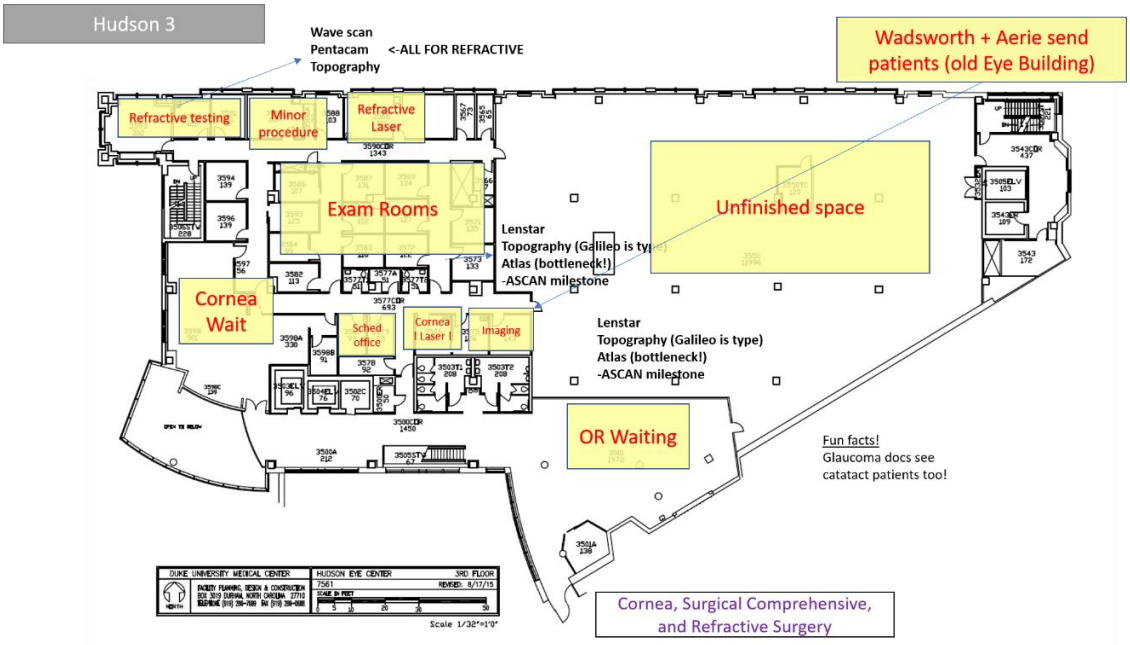


Figure 12: Cornea specialties and other departments, as well as operating room space unfinalized as of August 2015.

To focus on facilities available to a specialty clinic, Figure 13 marks the location of various facilities for glaucoma practice, reflecting more recent allocations. While imaging is shared with adjacent retina practices, DEC was dimensioned to sustain various practices sharing clinical space, with current room for expansion. The tradeoff is a physically extensive location which may be challenging for some patients. As organized, glaucoma practices combined may account for 100 or more patient visits to DEC on peak days.

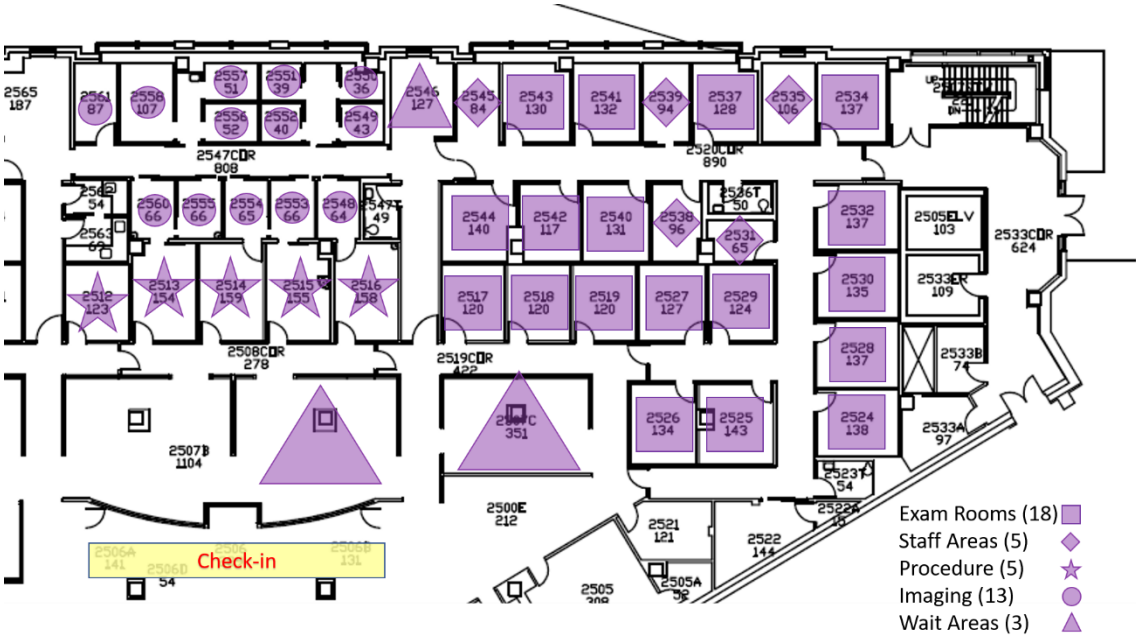


Figure 13: Detailed accounting of the facilities available for Glaucoma practices.

4. Measuring Performance in Outpatient Clinics

Time-motion studies can be traced from early optimization efforts (Welch & Bailey, 1952) to present simulations (Alfonso et al., 2012; Hribar et al., 2016; Hribar et al., 2018). Timing the movement of patients through a clinic bridges procedural understanding and quantitative evaluation.

The physical extent of centers introduces difficulties in measuring performance. Smaller practices are more amenable to observe (Vakili, Pandit, Singman, Appelbaum, & Boland, 2015), but inherently produce fewer observations that represent different practice conditions. Large academic centers provide more treatment options and accordingly paths through the system. A substantial devotion of measurement effort must be devoted to achieve full coverage, which can interfere with clinic service. Too few observers however may be overwhelmed with the frequency of events when several patients are proceeding through their visits simultaneously.

This chapter discusses measurement strategies for timing clinic performance in large academic centers, as well as findings from collection efforts at Duke Eye Centers (DEC). Anonymized timing data from more than 120,000 patient visits are introduced from electronic collection efforts and compared to in-person observations.

4.1 Performance Measurement Through Event Logging

Since the introduction of electronic health records (EHR), health care organizations now generate massive volumes of data during normal operations. New

initiatives apply data science techniques to gain insights from routine practice (Brennan, Chiang, & Ohno-Machado, 2018). Health care operations research in particular now benefits from "secondary use" of EHR, where patient data is used to improve practice outside direct use during a patient visit (Safran et al., 2007). Secondary use naturally leverages existing health information technologies such as EHR systems and point-of-care terminals, in line with national recommendations (President's Council of Advisors on Science and Technology (PCAST), 2010). The 'holy grail' of time motion studies is to collect visit duration information from keystrokes on existing systems and no additional collection effort on the part of caregivers.

Implementations currently find a compromise, where personnel must indicate the start and end of mutually defined events, corresponding to treatment steps. Circa late 2015, administrators at the eye center implemented an event logging system using extensions to the electronic health records (EHR) interface that can be accessed at point-of-care systems in each room. In point-of-care event logging, clinic staff use additional inputs added to their typical workflow in EHR interface to log the initiation or completion of locally defined phases of treatment. In contrast to direct observation, this system introduces no additional staff. Staff providing treatment are responsible for logging initiation and completion times of activities. There is no need for additional observer training or introduced ambiguity in start and end times as staff recognize when treatment has been performed.

This form of entry is however particularly subject to staff cooperation. Logging activities tend to be omitted particularly during critical peak periods when patient care takes priority, and paradoxically measuring these critical periods are highly relevant to quality improvement decisions. Records with inaccurate start, finish, or other omissions are difficult to identify without additional context. Other forms of context, such as patient complexity, may be eliminated in anonymization steps performed before data can be analyzed. Lastly, staff compliance and administrative mandates influence completeness of records. The event logging system yielded approximately 22,000 observations across all specialties in the first three-month period available, January through March of 2016. Event labels are mapped to treatment steps in Figure 14.

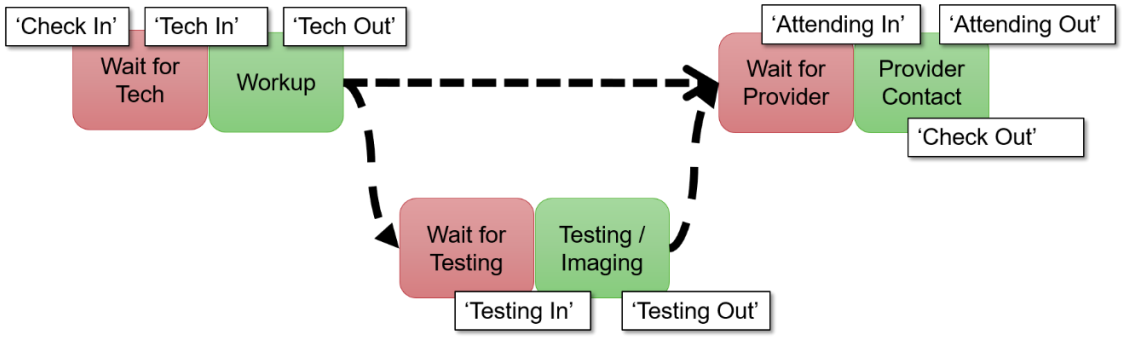


Figure 14: Measurement points in event logging, where practitioners agree on a finite set of standard events to record through inputs at EHR.

4.2 Verifying Event Logging Data

Direct observation provides the most potential for control over single observations, as dedicated human observers may take corrective actions in recording the duration of treatment phases. However, drawbacks include the volume of labor required

to capture high volume operations on a continuing basis, particularly when observers are paired with individual patients. This introduces several new personnel to a clinic, which can hamper efficiency or raise privacy concerns.

Alternatively, observations can be performed from a vantage point (Figure 15), but this strategy introduces uncertainty about when a treatment phase has definitively concluded versus, for instance, other reasons for patients or staff to leave an examination room such as restroom breaks or equipment procurement. A single vantage point may also have obstructed views of examination rooms, further decreasing the efficiency of this approach. In general observation also introduces human errors and inconsistencies between observers, even with a regimented observation protocol.

An observation program implemented at Duke Eye Center from April 2017 to November 2017 produced 43 patient visits using volunteer observers. Volunteers recorded event timings using a templated form (Appendix B). Continual retraining of volunteers to ensure patient privacy yielded unsatisfactorily low yield of data for the effort committed to this program, and observations were ceased in November amid a reorganization of Eye Center administration. Difficulty in measuring clinics in person is consistent with results in literature, where manual collection methods are typically applied for brief periods and potentially uninformative (Alfonso et al., 2012; Hribar et al., 2018; Vakili et al., 2015).

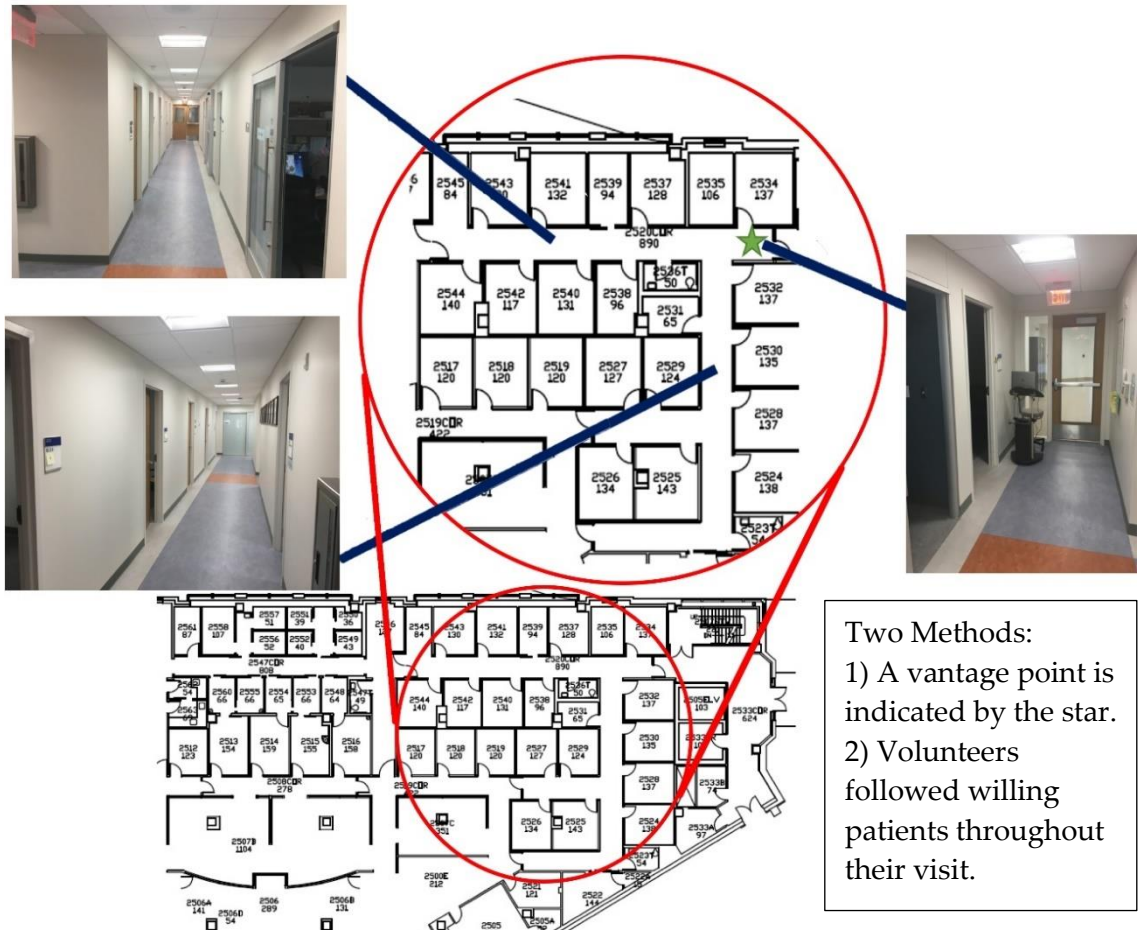


Figure 15: In-person observation was performed in Glaucoma.

The initial concerns about consistency between logged times and in-person observations, assuming that in-person observations provide more opportunity for verifying records, were examined using the 43 observations produced by volunteers combined with 54 observations using the single-vantage point method for the workup treatment phase. Histograms for the combined observation data set are compared to an 18-month period in Figure 16.

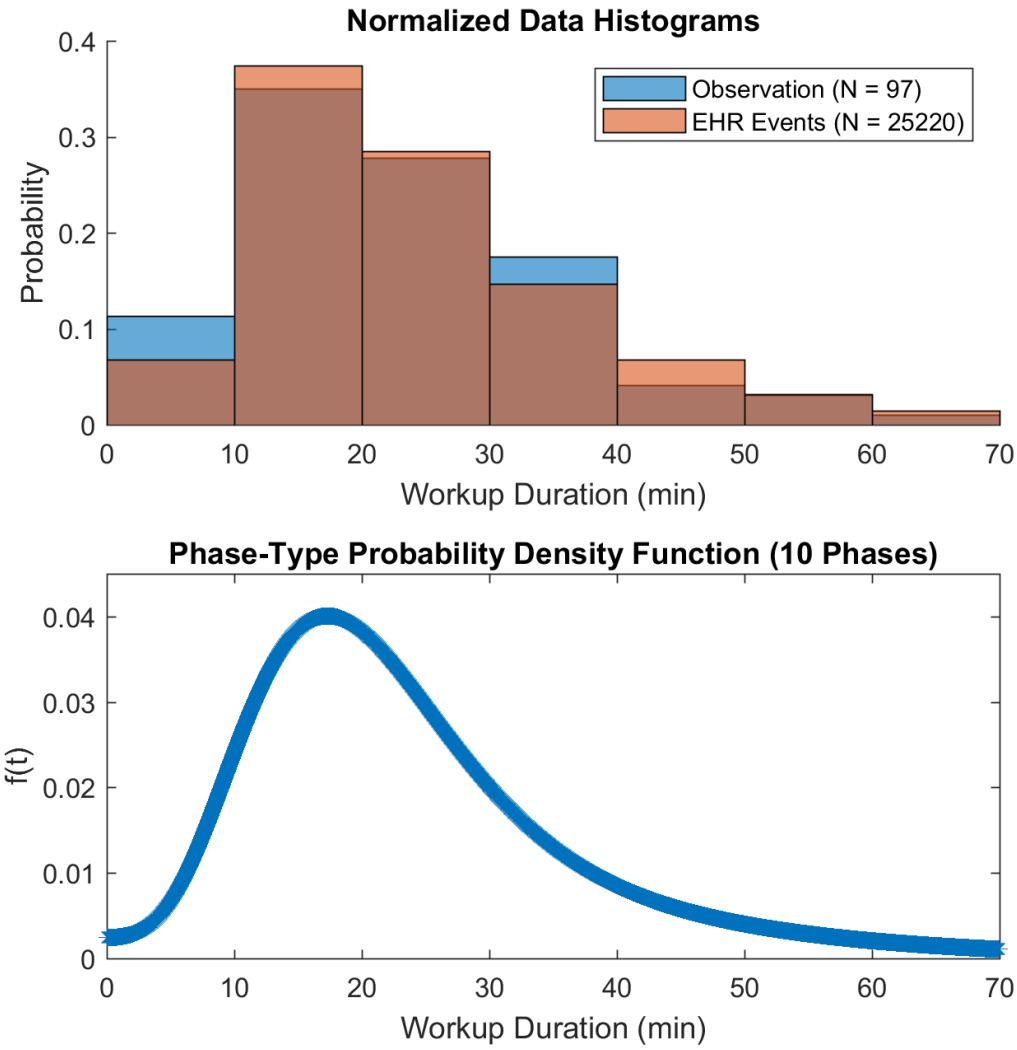


Figure 16: Histograms compared for in-person observation versus event logging collected from EHR, with a phase-type distribution in the lower frame.

A comparison of empirical distribution functions is plotted in Figure 17. High correspondence is observed between the measured durations of the same treatment phase under two independent measurements. The addition of in-person observation therefore does not deviate substantially from event logging measurements that

introduce no additional personnel. Further, while additional independent measures of these treatment phases may better distinguish questions of practical importance (i.e. the use of one type of imaging over another), these results do not indicate that additional measurement will drastically alter subsequent modeling procedures or provide information that cannot be derived from EHR. The EHR sample is adequate to demonstrate validation or optimization techniques.

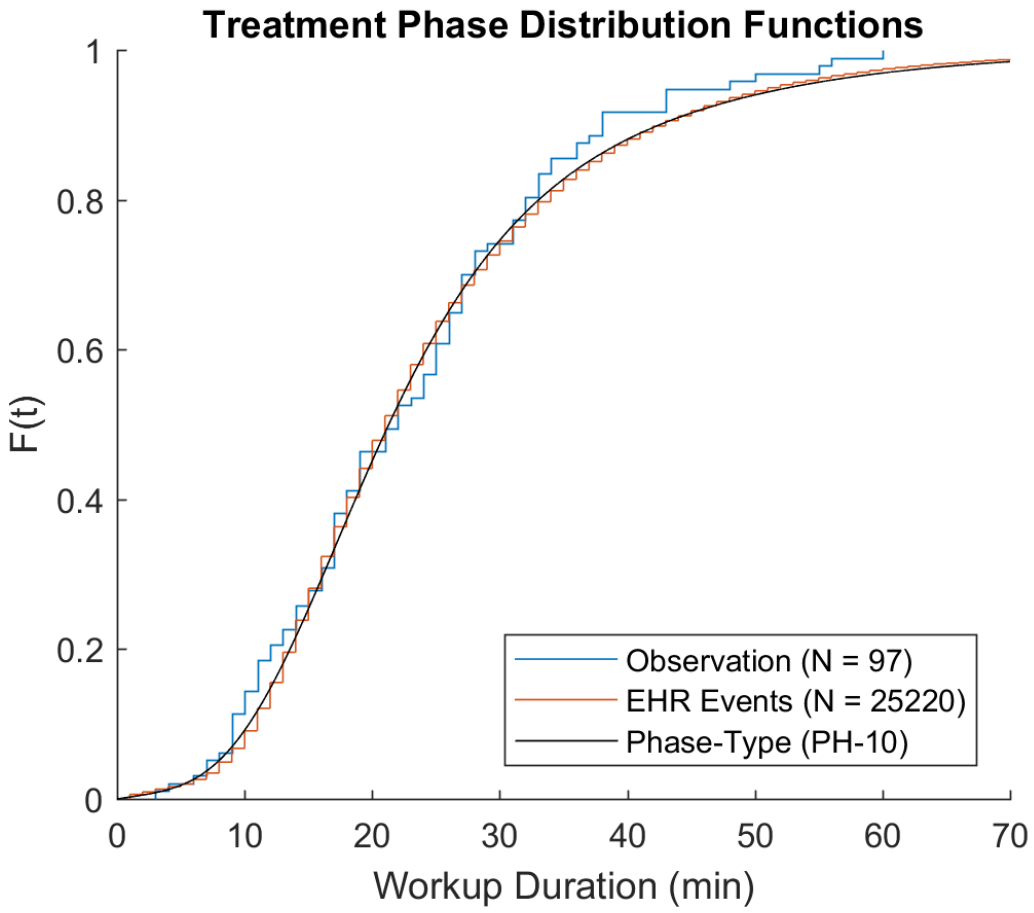


Figure 17: Cumulative distribution function graphical comparison for in-person observations, event logging through EHR, and a phase-type distribution model.

These findings are now further tested by quantifying goodness-of-fit. EHR event data through point-of-care terminals yielded 25,220 valid duration measurements for glaucoma workups during the period of Jan. 1, 2016 to Oct. 31, 2017. In-person observation yielded 97 measurements over approximately a nine-month window. We evaluate the statistical similarity between the data sets using two methods:

1. Two-sample Kolmogorov-Smirnov Test (D'Agostino & Stephens, 1986)

h_0 : the samples are from the same continuous distribution.

h_1 : the samples are from different continuous distributions.

Result: At the 99% confidence level (alpha of 0.01), we fail to reject the hypothesis h_0 (p = 0.8009, KS stat= 0.0647). The data does not suggest the measurements are significantly different.

2. Evaluating predictions by mean root squared error (MRSE)

A phase-type distribution with 10 phases is fit to the EHR event data (Thummler et al., 2005), then we compute the MRSE in predicting both sets of data. For a set $\hat{Z}$ of M observations occurring at time t_m where $m = 1,\ 2\ \ldots M$, the MSRE is given by

$$\text{MRSE} = \frac{1}{M}\sum_{1}^{M}\sqrt{\left(\hat{F}_{\hat{Z}}(t_m) - H(t_m)\right)^2}$$

Where $\hat{F}_{\hat{Z}}(t)$ is the empirical cumulative distribution function of sample $\hat{Z}$, and $H(t)$ is the phase-type distribution parameterized from EHR data. MSRE in this case can be read as a percentile error of the estimate.

Result: The phase-type distribution predicts the original EHR data with **1.29%** error, and similarly the observations with **2.85%** error. A model fit using EHR event data predicts in-person observations with similar accuracy.

4.3 Determining Adequate Sample Size for Estimator Consistency

As new clinics begin to collect secondary use data for optimizing services through simulation, each practice must determine when they have measured enough to produce robust predictions. Smaller samples may be insufficient to represent natural variations in patient visit durations, confounded by factors such as imperfect measurements and the range of treatments a practice offers (often referred to as the case mix (Kortbeek, 2012)). However larger sample sizes may provide no additional information; techniques such as learning curves (Goodfellow, 2016) have shown diminishing returns from continued data collection. Sample size selection has received much consideration in other domains such as clinical trials (Sozu, 2015). No such guidelines have been determined for clinic simulation, or other such estimators of clinic flow behavior. It is worthwhile to determine an "adequate sample size" tradeoff point between these two extremes as collecting samples often requires a commitment of time, personnel, and resources, but is vital to ensure accurate predictions.

In this section we describe quantitative experiments to determine sample size adequacy. Experiments use data collected from records of check-in and check-out event

times to clinics at Duke Eye Center (DEC) during a period between January 2016 and June 2017. The data are preprocessed into three fields;

1) *appointment time* - the check-in time evaluated as time elapsed since an arbitrary common time point in 2015,

2) *duration* - the visit duration as elapsed time between check-in and check-out times,

3) *department* - which one of four outpatient specialties within ophthalmology was visited.

Data are subject to random censoring, for instance due to delays or omissions by staff in manually logging events. No further distinctions are made from visit records. Time measurements are discretely quantized in the electronic health record systems but interpreted as continuous values with minute precision. We evaluate each department independently.

4.3.1 Consistency and Moment Estimators

Visits to each clinic vary in duration due to several factors; the utilization of staff and facilities, whether additional imaging or testing is required, the specialty visited, and variations due to the patients themselves. We treat the duration of a visit to a given department as a random variable X to account for differences between individual visits. A frequentist interpretation assumes that if enough samples of X are collected, estimators of the underlying sampled population will converge to theoretical true

values. As the sample size N approaches sufficiently large numbers, an estimator based on N samples (denoted $\hat{\theta}_N$) approximates the population response Θ. Consistency is thus defined as

$$\operatorname*{plim}_{N\to\infty} \hat{\theta}_N = \Theta$$

where for any $\epsilon > 0$, $P\left(\left|\hat{\theta}_N - \Theta\right| > \epsilon\right) = 0$ as $N \to \infty$.

Estimators are broadly defined, however we focus on four simple but informative estimators based on sample moments. These estimators are the sample mean, standard deviation, coefficient of skewness, and coefficient of kurtosis, which we denote $\overline{m_1}$ through $\overline{m_4}$ respectively.

$$(mean) \qquad \overline{m_1} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$(standard\ deviation) \qquad \overline{m_2} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \overline{m_1})^2}$$

$$(skewness) \qquad \overline{m_3} = \frac{1}{N(\overline{m_2})^{\frac{3}{2}}}\sum_{n=1}^{N}(x_n - \overline{m_1})^3$$

$$(kurtosis) \qquad \overline{m_4} = \frac{1}{N(\overline{m_2})^2}\sum_{n=1}^{N}(x_n - \overline{m_1})^4$$

In discussing randomly varying values, moments often serve as valuable summary statistics when a full characterization of the distribution of values may be unattainable. Moments are measures of central tendency. In practice, we must estimate moments based off a finite sample of size N. For samples from a homogenous

population of visit durations, these estimates are assumed to converge to a static value for sufficiently large sample sizes (Klenke, 2014). For unbiased estimators, the static value approximates the true population values as the sample size tends to infinity (Goodfellow, 2016).

4.3.2 Windowed Subsampling

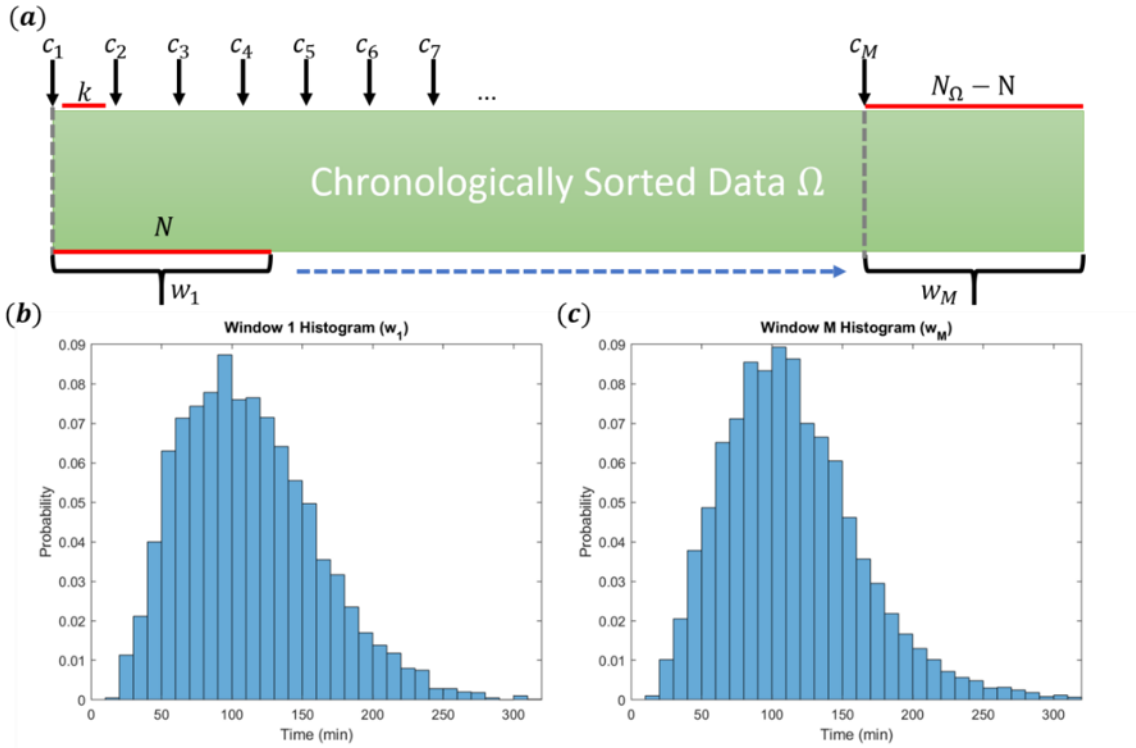
To estimate the sample size at which estimators stabilize, we take size N subsamples of the chronologically ordered set Ω of all N_Ω sample clinic visit durations using a sliding window, producing a set of windows $W = (w_1, w_2, w_3, \dots, w_M)$. Each window contains N data and is indexed by an offset value c_i so that window w_i contains samples $[c_i, c_i + N - 1]$. We increment c_i by a static value k for P windows so that $c_1 = 0, c_2 = c_1 + k, \dots c_M = c_{M-1} + k \leq N_\Omega - N$. This sampling procedure is shown diagrammatically in Figure 18. Values for the estimators $\overline{m_1}, \overline{m_2}, \overline{m_3}, \overline{m_4}$ are calculated in each window as N is varied.

We also assess the normalized range $R_{j,N}$ of estimator j values at sample size N, where $\overline{\boldsymbol{m}_{j,N}}$ is the vector of estimated moment values at size N. We define $R_{j,N}$ as

$$R_{j,N} = \left[\frac{\min\left(\overline{\boldsymbol{m}_{j,N}}\right)}{E\left[\left(\overline{\boldsymbol{m}_{j,N}}\right)\right]}, \frac{\max\left(\overline{\boldsymbol{m}_{j,N}}\right)}{E\left[\left(\overline{\boldsymbol{m}_{j,N}}\right)\right]}\right]$$

Where $E[\cdot]$ is the expectation operator, $\overline{m_1}$.

$$E\left[\left(\overline{\boldsymbol{m}_{j,N}}\right)\right] = \frac{1}{N}\sum_{l=1}^{N} \overline{\boldsymbol{m}_{j,N}}(l)$$



(a) For each fixed window size N, a sample window is incremented by k values across the entire chronologically ordered sample, producing overlapping windowed sets w_1,w_2,w_3,...,w_M, where w_1 contains the first N samples, w_2 contains the samples from k to k+N-1 and so on. Windowed resampling can be repeated for various window sizes; (b-c) plot histograms for w_1 and w_M in the Glaucoma practice, and are shown for N = 6000, k = 15.

Figure 18: Diagrammatic representation of the windowed sampling scheme.

4.3.3 Summarized Sample Size Results and Analysis

Figure 19 plots the normalized range metric $R_{j,N}$ for all departments. Each plot shows the range of moment values for all windows at a given sample size, normalized by the mean of those values. A line is indicated at 6000 samples. Mean and standard deviation vary by less than 10% of the mean of all estimated values for these moments once approximately 6000 samples are included, for all departments except for Vitreous

Retinal. The normalized moments Skewness and Kurtosis saw a wider range of values and were significantly more sensitive to sampling effects. All computed moments do appear to asymptotically converge by around the 6000-sample threshold for mean and standard deviation.

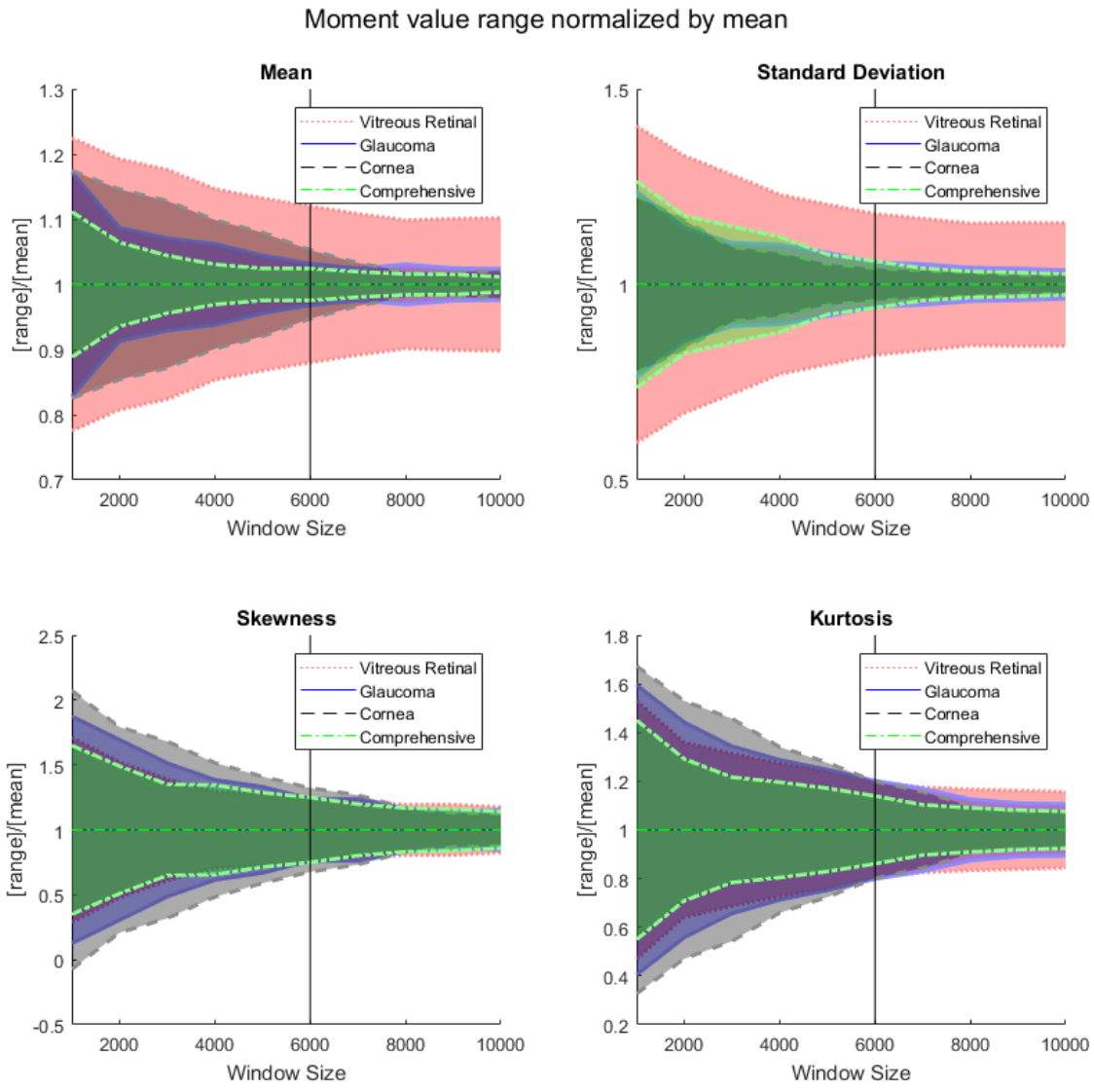


Figure 19: The range of moment estimates at each window size (Eq. 6).

These central tendencies indicate the variations in visit duration are finite; 6000 samples are sufficient for consistent estimates of moments. Samples past 6000 do not further narrow the range of moment estimates.

4.3.4 Conclusions on Sample Size for Consistent Estimation

Estimating sample moments from various time periods in the total sample provided two informative findings on sample stability.

First and notably, we see no seasonality influence on the overall visit duration. In practices that include trainees such as physician-fellows with one-year appointments, we do not encounter periodic variations in visit duration. If there were a substantial acclimation period for these trainees, we would expect longer visits shortly after the trainees start and resulting effects such as increased mean durations around the fellowship start period. Rather, the visit duration is primarily influenced by technician measurements and process wait times during busy periods, as well as individual needs of the patient. Since technicians do not join or leave practices en masse as with limited traineeships, the technician contribution to visit duration is consistent and predictable with enough samples. These results indicate the patient population similarly contributes bounded variation to visit durations. Busy periods do significantly impact visit durations within one day, but sample sizes exceeding 100 visits span more than one days' worth of collection at this practice. Sample sizes at the chosen consistency point of 6000 provide adequate sampling of busy and idle periods.

Second, we find that convergence of the sample moments is a useful indicator of when a feature has been sufficiently measured for testing more elaborate models. This convergence occurs in all departments in sample sizes that are 18.6-33.4% of the total available sample size in this study. Reducing the collection requirements may save substantial time and work hours for clinics beginning to collect visit timings, and lead to earlier modeling efforts. Unless a significant change is enacted in clinic practices, the additional information does not improve model selection or selected model performance.

Sample sizes of approximately 6000 adequately account for natural variation in a sample of measurements of the duration of outpatient appointments. This approach for sample size evaluation may generalize to practices in other facilities and specialties. Outpatient practices with similar workflow may find consistent results. The estimates are consistent once this sample size is achieved, independent of when the sampling period occurs.

Analysis in each department determined a sample size for estimator consistency that is beyond most in-person collection efforts, but feasible for automated systems. Considering the mean visit duration to a glaucoma clinic was found to be approximately two hours, collecting a sample size of 6000 visit durations through direct observation requires at minimum 12,000 hours of labor, a prohibitive demand.

4.3 An Anonymized Performance Data Repository for Operations Research, and Anonymization Methods

The formation of repositories for performance data on clinic operations is essential for independent replication of studies. The ability to replicate findings or compare results on identical data is necessary for objectively improving methods and advancing a standard approach to clinic modeling.

(R. Fricks, Veihl, Tseng, Trivedi, & Barr, 2018) references a repository containing 122,204 anonymized visit records collected through event logging at Duke Eye Centers. These data were used throughout this dissertation, and collected from the practices and measurement methods described in previous sections.

To inspire similar repositories, this section outlines the precautions taken in de-identifying this protected health information (PHI). In American health care systems the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule designates as PHI any "individually identifiable health information" (*Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* 2012).

Two de-identification standards are defined: Safe Harbor and Expert Determination ("45 CFR 164.514 - Other requirements relating to uses and disclosures of protected health information," 2000). Safe Harbor removes any potentially sensitive information from one of several defined categories and is too stringent for operations research applications where appointment date provides significant information. The

alternative, expert determination, defines an expert as "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" (*Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* 2012). This expert applies documented methods to render the risk of re-identification of the anonymized information improbable.

The methods employed with the (R. Fricks et al., 2018) data set are as follows:

1. Categorical fields, such as personnel names, are encoded by assigning all unique entries a number, then replacing the occurrence of each entry in that field with the number. For example, if Dr. X is assigned number 17, Dr. X appears as '17' in every record associated with Dr. X. Categorical fields are: 'Provider', 'Fellow Name', 'Resident Name', 'Ascan Tech', 'Injection Tech', 'Rooming Tech', 'VF Tech', 'Work Up Tech', 'Imager (OCT Retina)', 'Imager (Testing)', 'Orthoptist Name'.
2. Appointment dates were encoded with the following procedure:
    a. Initialize random number generation with a pre-selected seed value, for auditing purposes
    b. Convert all data into dates relative to Dec. 31, 2015 (all records are originally 2016 & 2017).

c. Calculate how many unique dates show up in the records designated as Y, and the maximum relative date as X.
d. Each date is now uniquely assigned a number between X+1 and X+Y+1, which is randomly sorted
e. The entire data set, which is originally in date order, is first sorted by the calculated length of stay (LoS) field. LoS is effectively uncorrelated with appointment date.
f. It is then sorted again by the assigned physician number.
g. A column named 'month' is added, which has a number 1-12 indicating which month the patient visited during, to retain some ability to search for seasonality

For example, if an appointment occurs on Jan. 1, 2016, its relative date is 1, and day 2 is Jan. 2, 2016. If the max relative date is 400, and there are 300 unique appointment dates, entries for Jan. 1 will be assigned a new date which is uniformly likely to be between 401 and 701, or in other words, have a 1/300 probability of being 459 for instance. 460 is no longer guaranteed to be Jan. 2, 2016.

The de-identification processes should retain enough information to replicate studies, but not trace data back to patient identity. The randomization and sorting is irreversible with the information provided. Given these data do not contain health status

info besides a 'primary diagnosis' field that is itself categorical, these procedures should satisfy requirements for de-identifying PHI.

5. Modeling Outpatient Clinics

Beginning with an abstraction of system behavior, and corresponding performance measurements, modeling aims to produce a representation of crucial system elements that informs or allows for analysis. This chapter describes iterative development of a predictive white-box model for outpatient glaucoma clinics. The model capitalizes on the precision and power of stochastic reward net (SRN) formalisms, which are underused in modeling health care performance and underdeveloped in the literature. Combining SRN with discrete event simulation (DES) for solution overcomes limitations inherent in state-space methods, such as largeness concerns or the requirement that models are fully specified as an SRN, facilitating the specification of schedules in later quality improvement analysis. The development of a hybrid SRN-DES model is presented iteratively, with successive refinements removing common clinic modeling fallacies through rigorous validation. Cross-validation techniques are adapted for use in this domain and employed throughout to assess generalization performance, selecting models that minimize the expected error in predicting independent data (Hastie et al., 2009).

Traditional validation of discrete event simulation models employs 'input-output' statistics techniques, where the model outputs are compared to data inputs via a suitable statistical test as in examples in (Alfonso et al., 2012; Hribar et al., 2018; Hideaki Takagi et al., 2017). The presence of random censoring in patient visit records restricts

the ability to validate model performance using average occupancy, as missing records may skew estimates of patients or staff status in the clinic unpredictably. To ensure comprehensive accuracy in the resulting model, intermediary endpoints were defined corresponding to different steps in the clinic visit model (Figure 20). These checkpoints can be used to validate a model for visits to the glaucoma clinic by comparing available measurements of the time elapsed from "check in" to each check point with simulation trace results at equivalent steps in treatment.

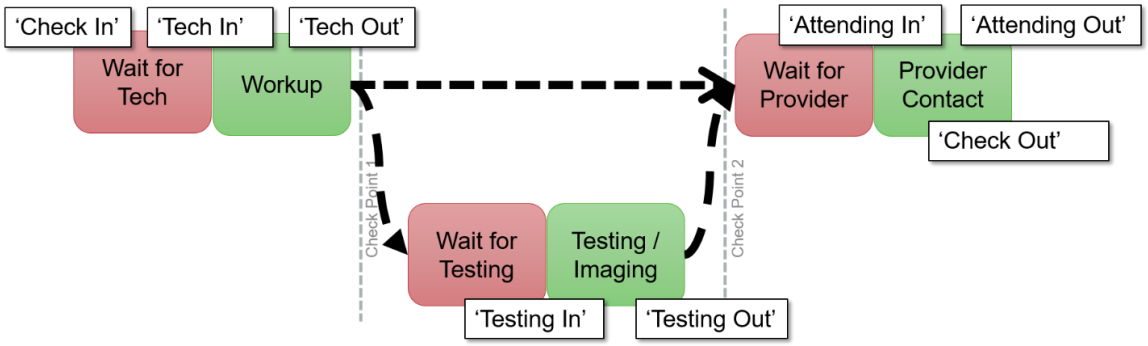
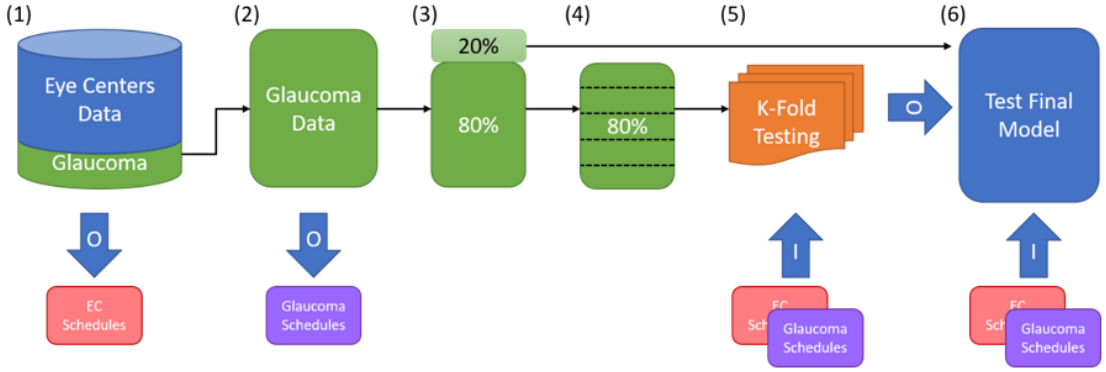


Figure 20: Steps in glaucoma treatment, showing measurement points in the event logging system.

In addition to censoring, the validation protocol must also account for inherent flaws in data collection and potential variations or outliers in the measurand. One interpretation of overfitting phenomena is that accuracy has been improved with respect to a sample by representing measurement noise. Adding more complexity that improves accuracy with respect to one sample may diminish accuracy when evaluated versus a similar, independent sample. Independent evaluations in cross-validation provide an indication of when overfitting may occur and are necessary to produce robust models that

remain predictive outside of the initial sample. Figure 21 diagrams the data partitioning that is used in cross-validation during iterative refinement phases.



(1) Contiguous schedules of Duke Eye Center days are extracted from the entirety of the study period sample. (2) Glaucoma-specific demands on technicians are extracted from the subset of glaucoma clinic visit records. (3) 20% of the data is partitioned into a test set using simple random sampling; the test set is excluded from modeling until the final evaluation in (6). (4) The remainder of samples account for approximately 26000 visits to glaucoma clinics, noting the individual duration of patient visits, and is subdivided into 5 folds. (5) K-fold cross-validation is used to refine models for generalization performance using the schedules as inputs. (6) Once all model features have been selected, the model is parameterized from the 80% sample, and tested against the 20% sample sequestered in (3), with schedules for patient arrivals, concurrent technician and imaging demand as inputs.

Figure 21: Data Partitioning Diagram for Cross-Validation.

5.1 Challenges in Modeling High Volume Outpatient Practices

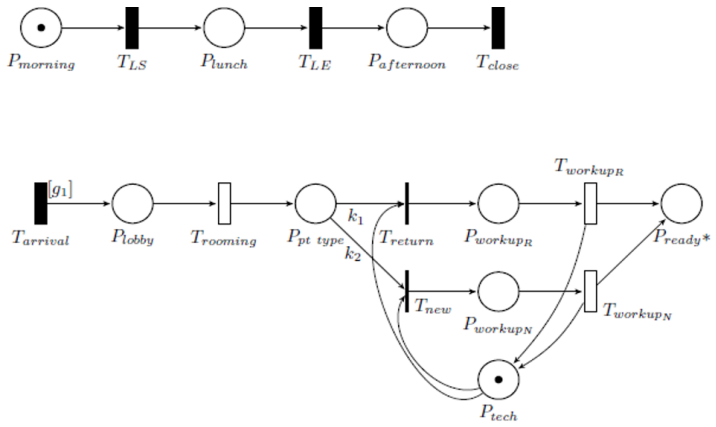
Compared to the inpatient setting modeled in (R. B. Fricks & Trivedi, 2016), the outpatient setting introduces distinct modeling considerations and technical challenges to the solution of large Stochastic Reward Net (SRN) models. (R. B. Fricks et al., 2017) introduced the complications in adapting SRN methods to outpatient clinics with finite schedules. Examining differences between the generic ICU model to the clinic model in

(R. B. Fricks et al., 2017) elucidates modeling considerations necessary for accurate representation in outpatient settings.

Dynamic behavior critically distinguishes the ICU from an outpatient clinic. A general assumption in inpatient critical care is that the ward will operate indefinitely, at all hours in a given day, every day. In modeling such a system, features of interest include the instantaneous availability of capacity in a ward, referred to as beds. When the ICU is filled to capacity, incoming patients must be diverted elsewhere, which negatively impacts patient outcomes (R. B. Fricks & Trivedi, 2016). Suppose the number of available beds is modeled as a stochastic process. One approach is to represent the number of beds as a queue and solve for the steady-state distribution of bed availability. Modeling the ICU as a queue in the steady-state asserts that the distribution of available beds in the real-world ICU converges to a stationary distribution. Steady-state queue behavior was found to be a reasonable approximation for ICU occupancy (McManus et al., 2004) and queuing models are widely accepted in this domain (Ahmadi-Javid et al., 2017).

In contrast, outpatient clinics operate on finite schedules, requiring a transient solution. Using the Petri net package SPNP (G. Ciardo et al., 1989; Hirel et al., 2000), the SRN in Figure 22-23 emulates a templated schedule with a defined lunch period. This emulation is accomplished by using the upper nodes in Figure 22 to modulate a guard function $[g_1]$ which controls arrivals. When the $(close)$ transition ultimately fires,

arrivals are permanently shut off for that simulated trial, and the clinic model eventually clears all patients from the system.



This first proposed model presents one potential abstraction of the observed clinical practices at Duke Eye Center. The Entry Process SRN includes nodes purely for emulating scheduling practices, which are the places ($morning, lunch, afternoon$) and associated transitions. The presence of a token in morning or afternoon places enables the arrival transition, which ultimately is disabled for the remainder of a simulated day. This figure connects with Figure 23 at the ($ready$) place; the SRN model is presented in two pieces for ease of reproduction.

Figure 22: Entry process for early eye center model.

The model in (R. B. Fricks et al., 2017) used the Anderson-Darling goodness-of-fit test (D'Agostino & Stephens, 1986) to select optimal transition distributions from common canonical distributions (Trivedi, 2002), and maximum likelihood estimation (Ascher & Feingold, 1984; Goodfellow, 2016; Manolakis et al., 2005) to parameterize the top indicated model. In every transition, non-exponential distributions were indicated for treatment step completion time.

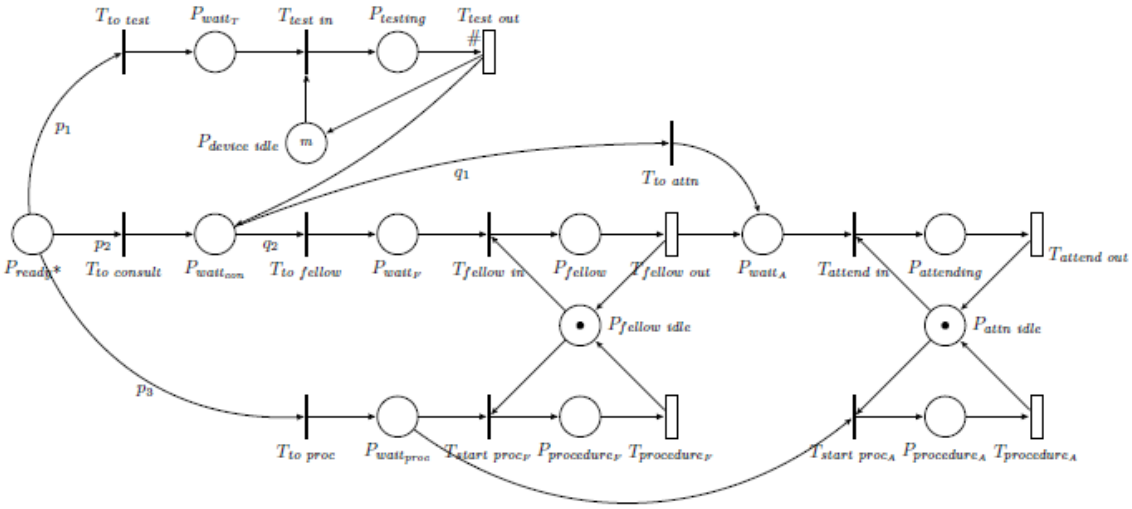


Figure 23: Subsequent steps in eye center treatment.

Finally, this model assumes an independent clinic with dedicated resources. The SRN depicted here connects to Figure 22 at the ready position to produce one complete model of clinic flow at glaucoma practices. Notably this representation explicitly includes visits with a trainee physician (fellow), and the possibility of a procedure during clinic days led by the attending physician, which were later deemed a rare occurrence. This and other modeling decisions throughout (see (R. B. Fricks et al., 2017) for comprehensive accounting) were unable to be validated in practice due to technical limitations. Specifically, the simulation implementation in the Petri net package SPNP does not readily report trace information for individual patients, rather it retains information on system features and reward function specifications such as token count distributions at different places (G. Ciardo et al., 1989; Hirel et al., 2000). To proceed with

the cross-validation outlined in figures 20-21, it became necessary to implement custom discrete event simulation to retain simulation traces.

5.2 Generalized Simulator Logic

Implementing a discrete event simulation program that retains timers for all entities in the system is less memory efficient than alternatives that process events without retention or that keep a limited log listing previous events. The custom program developed to simulate clinic activity is deliberately generous in memory allocations to facilitate debugging and experimentation with the simulator. The program follows general principles in discrete event simulation (Law & Kelton, 2000). In pseudocode, each iteration of the simulator must:

- Determine the next completed event, by finding the minimum of the remaining time on all active timers
- Advance the system clock to the next time point
- Update the system status variables such as a counter for the number of patients in the system, position of each patient, and returning personnel to an idle pool after a step has been completed
- Determine what queue, if any, to put the current patient in for the next round
- Update active timers

- Update all queues by assigning the top of the queue to the first idle personnel required, if the queue is not empty and there are idle personnel
- Record several custom metrics in the reported simulation trace

For a schedule with N patients requiring as many as M treatment steps, worst-case behavior requires NxM iterations of the main simulator loop. This worst-case behavior is reduced by the fourth step, which in practice advances patients past steps that may not be required. The complete, barebones implementation is relatively efficient for a simulator that retains so much potential debugging information and may be useful in testing other models. The core of the custom simulation program is included in Appendix C, as a MATLAB implementation.

5.3 Representations of Treatment Time

Provided a simulator capable of retaining individual patient traces, we now re-evaluate all modeling decisions which led to (R. B. Fricks et al., 2017), producing successive refinements. This begins with a black-box model of health care treatment steps, modeling the time until completion without considering which specific steps are undertaken during a visit. Limiting the granularity at which clinic flow is explicitly modeled by white-box methods avoids over-specifying a model which must ultimately present a simplified representation of clinical practice (Gunal & Pidd, 2010). This further distinguishes medical decisions from administrative decisions, as we do not propose changes within treatment steps.

Cross validation is used in selecting between several plausible probability distributions to model treatment step durations. Using measurements of step durations from the event status data, generalization performance is used to select the class of distribution and any associated hyperparameters from a finite set of distributions. Common alternatives such as exponential, uniform, or pareto distributions are omitted as in early evaluations they were significantly outperformed by the phase-type distributions. For each iteration in the k-fold scheme, one data subset (fold) is withheld as a validation set, and all other folds comprise one training set. In each iteration the phase type distribution is optimally parameterized with respect to the training set, then the mean absolute error (MAE) is computed against both training and validation sets. Parameters in each distribution are optimal fittings for that distribution for a given training set. In contrast hyperparameters such as the number of phases are selected based on the generalization performance. Phase type distributions are computed for discrete numbers of phases from 3 to 25 using BuTools 2 MATLAB implementation (Horvath & Telek, 2017). The algorithm implemented uses a subset of phase-type distributions described in(Thummler et al., 2005). In general phase-type distributions are absorbing continuous time Markov chains Using phase-type distributions to represent the time until a task is completed, the time until absorption is interpreted as the completion time and is given by

$$\pi(t) = 1 - \boldsymbol{\alpha} e^{At} \mathbf{1}$$

For a PH distribution with N phases, $\boldsymbol{\alpha}$ is a $1xN$ initial probability vector, A is a NxN infinitesimal generator matrix, and $\mathbf{1}$ is a column vector of ones. Parameters $\boldsymbol{\alpha}$ and A are optimized with respect to the training set, and N is treated as a hyperparameter. The model is evaluated for each fold as N is varied to select a value for N that minimizes the expected prediction error.

Figure 24 and Figure 25 plot the computed MAE as a function of the number of phases used in representing workup durations and imaging durations, respectively. Distributional models with varying number of phases are evaluated versus the training set used to parameterize the distribution and an independent validation set. K-folds cross-validation allows for k repetitions of the evaluation protocol and removes the emphasis from the choice of validation set. The ranges of error across folds are plotted as colored regions, with means indicated by phase number. The mean training errors (indicated by *) are the MAE averaged across all folds for a given number of phases and are within blue regions showing the range of all training set error results. Similarly, the validation error means, which approximate the prediction error, are denoted by (+). Red regions show the complete range of validation error results for all folds.

Workup durations for Figure 24 are computed by finding the elapsed time between the time stamps “Tech In” and “Tech Out.” From the glaucoma data subset after the test set is excluded, 3082 raw observations of patient visits yield 2623 valid workup duration measurements. The 3082 raw observations are separated into five

approximately equal folds using simple random sampling. While visually difficult to distinguish between plotted distributions, quantifying the error in k-fold cross-validation produces some interesting insights. As theorized, the PH distribution approximates the training data to a minimal error, converging at approximately 1.44% mean training error as the number of phases is increased. Similarly, the range of training error narrows as more phases are added. Herein lies the importance of evaluation against independent samples. While validation error does see significant improvement as phases are added, there is no significant improvement in mean validation error arguably after 10 phases have been used to represent this distribution. The range of validation error does not narrow as the training range does with increased phases, indicating no additional benefit to generalization performance from the continued addition of phases. Using the mean validation error as an estimate of prediction error, a numeric minimum of 2.07% is found at 15 phases used to represent workup durations.

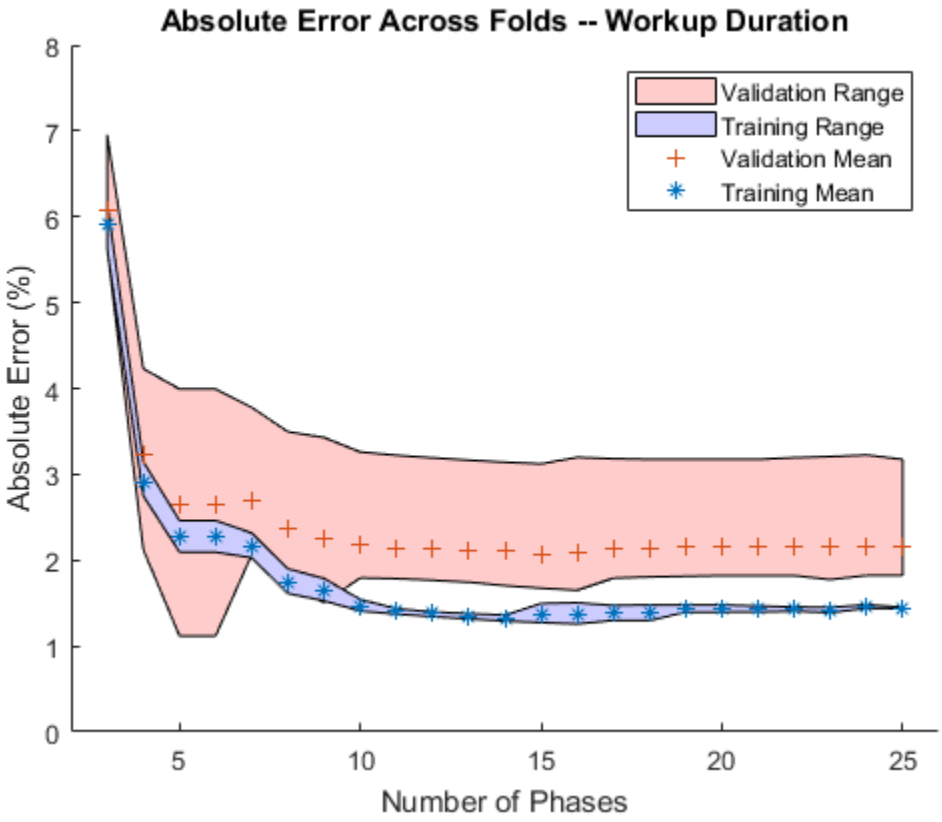


Figure 24: Mean absolute error for phase-type distributions representing workup durations.

Similar results are seen for training the imaging time durations. Figure 25 plots absolute error for phase-type (PH) distributions for imaging duration as a function of

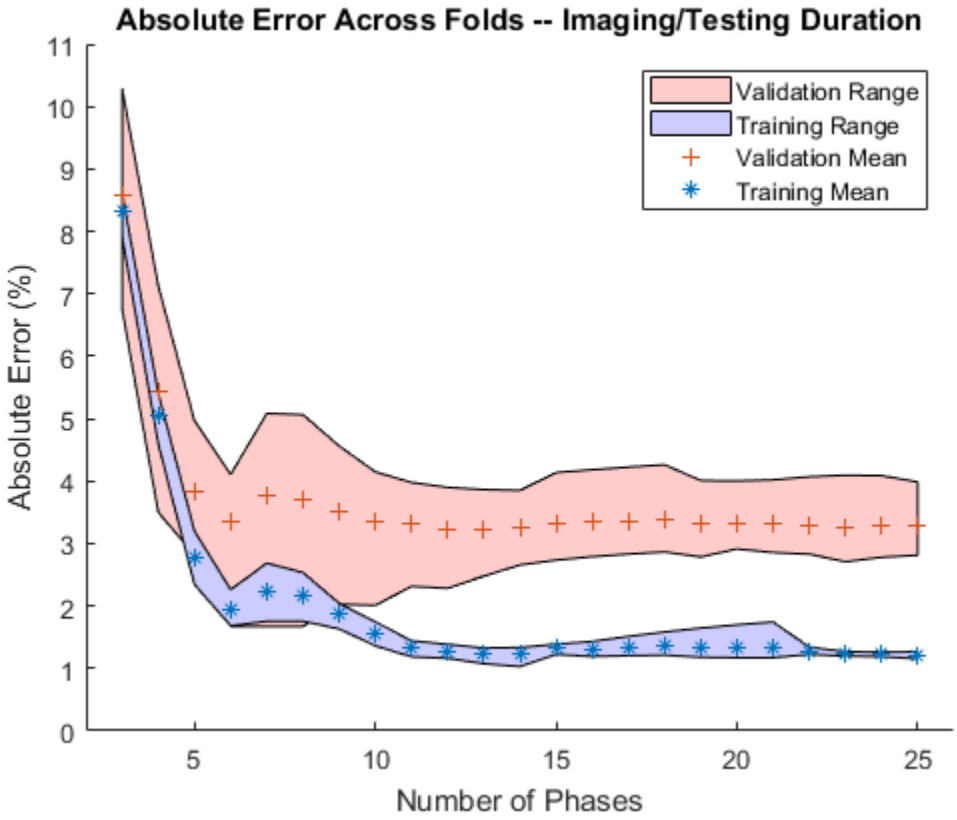


Figure 25: Absolute error for phase-type distributions with varying number of phases, used to model the duration of generic imaging procedures.

the number of phases. The elapsed time between "VF In" and "VF Out" is used as a proxy for imaging or testing duration. From the 3082 raw observations, 578 recorded durations of VF imaging are available in the k-fold subsets. For PH distribution results, the mean training errors (*) are shown within blue regions showing the k-fold range of individual training sample results. Similarly, the mean validation errors are denoted by (+), within red regions showing the range of absolute validation error results for all folds. Once again, the mean validation error decreases consistently as the number of phases is increased for this fitting method, stabilizing after approximately 10 phases

have been used to represent the transition. Adding more than 10 phases increases model complexity without appreciable decrease in expected prediction error. Greater variability in the range of error is observed in fitting the imaging durations. This may be partly attributable to the smaller sample size; imaging had approximately one fourth the available samples compared to workups, which are performed for every patient. The overall pattern is consistent; training error reduces to a minimum of 1.2% as more phases are added. A mean validation error minimum of 3.21% is found when 13 phases are used to represent imaging duration as a PH distribution.

Based on these results we select phase-type distributions with 15 phases for modeling workups, and 13 phases for modeling imaging procedures. Compared to the examined alternatives, these optimal phase numbers are estimated to minimize prediction error. Overall PH distributions exhibited consistent performance as indicated by a small range in error values, provided a sufficient number of phases are used in representation. Note that the durations of different treatment steps are not specific to each clinic. At DEC technicians within one specialty frequently examine patients for multiple clinics over the course of a week, even if one technician may spend most of a day evaluating patients for one clinic. There is insufficient evidence to indicate the duration of these steps varies based on any other patient or clinic categories.

In adapting cross-validation for this environment, results are consistent with well-understood principles of the use of cross-validation in many other disciplines

(Goodfellow, 2016; Hastie et al., 2009). Mild but demonstrable overfitting occurs when many phases are used to represent treatment step durations.

5.4 Modeling Concurrent Clinic Interdependence

An SRN (Figure 26) is now used to incorporate the treatment step duration with shared system resources, following the discussion in (R. B. Fricks et al., 2018). The goal is to model the time until workups are complete, in a manner that still provides insight into system dependencies and can be used for planning resource allocations. Consider the duration of a patient visit from check-in to workup completion as a random variable Y with probability density $f_y(t)$. Since the workup is the first step in a patient visit, Y is the sum of how long the workup procedure itself took, and any wait time until the workup began (once a technician begins a patient exam it is effectively not interrupted or restarted). Denoting the delay time X_d and the workup time X_w also as random variables, we assume that the delay duration is independent of the workup duration, and can write an expression for the probability density function of completion time Y as a convolution

$$Y = X_d + X_w$$

$$f_Y(t) = f_{X_d}(t) * f_{X_w}(t)$$

EHR logging provides several direct measurements of the workup time, X_w, as well as the completion time Y. The X_w data are used to parameterize a phase-type distribution from the class described in (Thummler et al., 2005), using the

implementation in BuTools 2 software package (Horvath & Telek, 2017). Phase-type distributions have been effective in fitting casual measurements for similar applications (Fackrell, 2008; Trivedi & Bobbio, 2017). Using phase-type distributions to represent the time until a task is completed.

For a PH distribution with N phases, $\boldsymbol{\alpha}$ is a $1xN$ initial probability vector, A is a NxN infinitesimal generator matrix, and $\mathbf{1}$ is a column vector of ones. Parameters $\boldsymbol{\alpha}$ and A are optimized with respect to the training set, and N is treated as a hyperparameter. We evaluate the model for each fold as N is varied to select a value for N that minimizes the expected prediction error. Y data is used in validation of the composite model, which estimates the convolution result from previously.

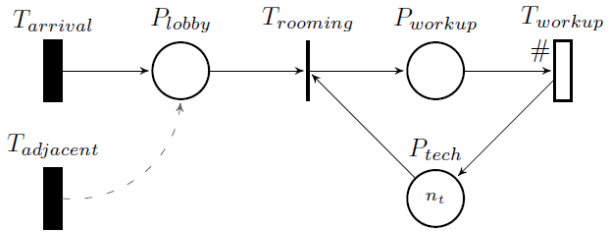


Figure 26: Iterations of the concurrent model demonstrated that delays in entry were solely due to staffing demand.

Table 1 displays results from the simulation, where three variations of the model (Fig. 26) are considered. The concurrent model has the adjacent arrivals enabled, which represent patients in clinics outside of the primary high-volume clinic studied. These results are contrasted to an independent clinic model (Ind. 4T/Ind. 3T) which disables adjacent arrivals and models dedicating technicians solely to one clinic. The 3T/4T

designations refer to the number of technicians assigned, to provide a range of similar patient-to-technician ratio (PTR) that provide upper and lower bounds to the optimal concurrent model PTR. Finally, error is reported as a percent, as the MRSE as computed can be interpreted as the average vertical distance between empirical data distribution and model. This distance, measured in percent, represents the average percentile deviation from a model estimate to a typical data set.

Table 1: Simulation System Parameters and Results for Entry Process Model

Metric	**Model**		
	Concurrent	*Ind. 4T*	*Ind. 3T*
Average Total Patients	89.08	56.08	56.08
Technicians Assigned	5	4	3
Patient-Tech Ratio	17.82	14.02	18.69
Average Training Error	2.38%	15.58%	12.50%
Average Validation Error	2.81%	15.48%	12.60%

The consistent reduction in error seen by explicitly modeling concurrency indicates a worthwhile addition of complexity to improve accuracy. The concurrent model minimizes error compared to the independent alternatives, by approximately 10-12%. There is little discernible change in training versus validation errors, further indicating a robust model that continues to perform as anticipated when new data is presented. Interpreted in this application, the model will accurately predict patient flow in subsequent clinic days. Plotting the predicted distributions along the measured data

from an arbitrary fold (Fig. 27), the concurrent model results are clearly distinguished from the alternatives. The concurrent model (blue line) is virtual indistinguishable from independent data for this checkpoint, although clearly the independent model choices flank the data without adequately representing obesrvations. While PTR values are similar, concurrency better captures the transient availability of technicians, where moment-to-moment a probabilistic number of technicians is available to any given clinic. Internally reviewing simulation traces further emphasizes this point. We conclude that representing explicit dependence improves accuracy by an appreciable margin and confirms the need for interconnected models.

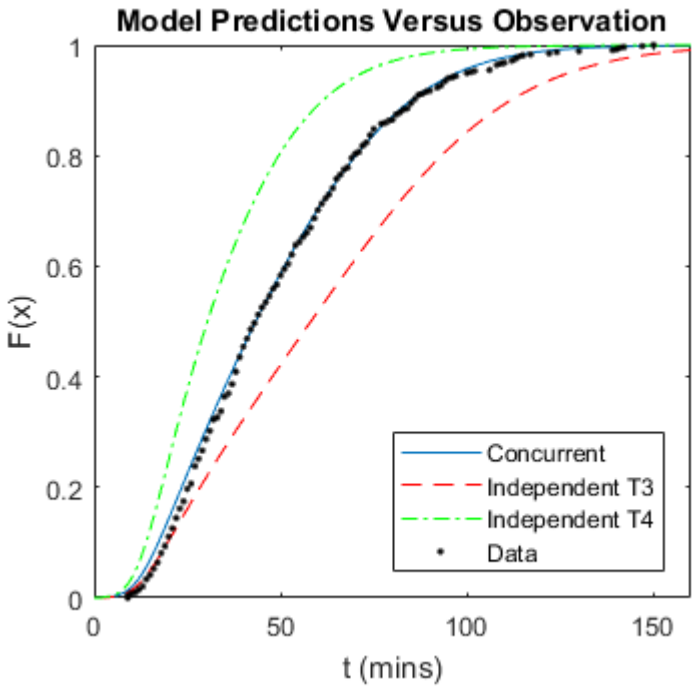


Figure 27: Workup prediction accuracy improves significantly when clinic concurrency is explicity modeled.

5.5 Evaluating Generalization and Reproducibility

Extending the concurrent clinic modeling procedure from workup to include the testing or imaging step, this section tests whether the proposed model is a suitable generic representation of glaucoma clinics.

The SRN model in Figure 28 combines individual transition distributions with resource requirements for a composite clinic model. The SRN represents glaucoma clinic practices at Duke Eye Center, with sufficient detail for planning decisions such as changing the number of personnel. Patients arrive and are checked in to a lobby waiting area. If a patient is in the lobby, the next available technician will pick up the lobby patient and begin a workup. Once a workup is completed, a decision to perform tests such as imaging is handled probabilistically in the SRN. In the model in Figure 28, the technician returns to an idle pool once a workup is complete, and a separate, next-available tester continues with the next treatment step. Finally, the patient is ready to see the physician. In addition to the demand on technicians and testers generated by glaucoma, demand for technicians and tests by adjacent clinics is represented in the nodes lower in the graph, similar to (R. B. Fricks et al., 2018). These demands are trace-driven and are associated to the arrival schedule. The trace-driven transitions are indicated as blackened timed transitions. Traditional timed transitions are marked by the corresponding phase matrix parameterization. While in principle the phase-type distributions used to model transition times here can be decomposed into a set of

exponential transitions (Trivedi & Bobbio, 2017), we choose to relax SRN conventions in favor of keeping the diagram compact and decipherable.

K-fold cross-validation is used at the Petri net level as well to ensure the model is adequately structured. In each fold evaluation the parameters $\boldsymbol{\alpha_1}$, A_1 are computed from workup durations in the training set, and similarly for $\boldsymbol{\alpha_2}$, A_2 from imaging durations. The number of technicians and imaging stations available (n_t, n_m) are optimized as hyperparameters, as the "effective" number of available staff or resources varies from a strict count of assigned technicians, who perform other duties. Finally, at this stage visit type information is incorporated in setting the probability that imaging is performed, and assigned a chance ranging from .15 for routine returns, to 0.6 for new patients, to 0.95 for patients flagged for imaging studies. These values accommodate for underreporting in records as well as a small degree of flexibility in assigning patients to available slots (i.e. a routine return patient scheduled under an imaging study slot). The Petri net is then simulated for 10000 iterations for each viable schedule to sample the model response for three simulated clinics designated SC A through SC C. Simulation is performed using the discrete event simulation implemented (Section 5.2) in MATLAB 2017a (Mathworks, Inc.), using BuTools 2 (Horvath & Telek, 2017) to generate random variates from the appropriate distributions. Results for each simulated patient are aggregated as an estimate of the distribution of elapsed times at checkpoints 1 and 2 in

Figure 20 and evaluated against training and test sets for several clinics.

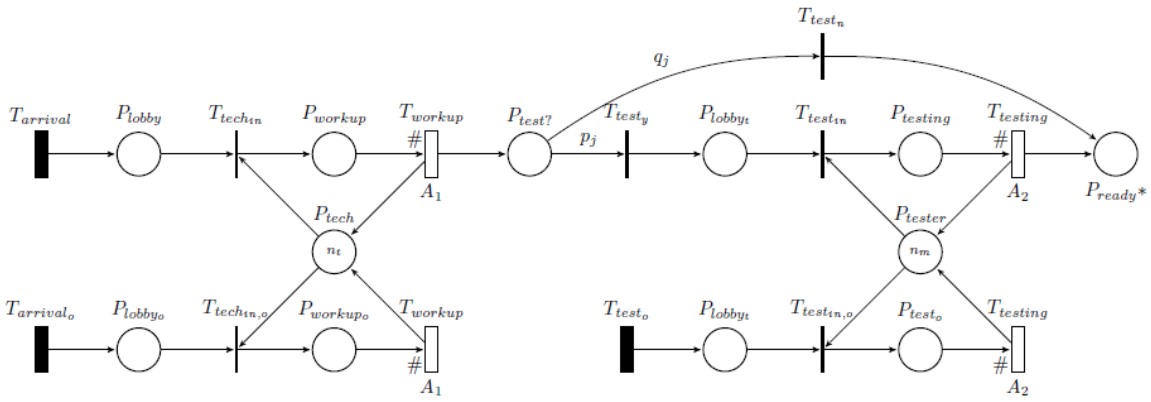


Figure 28: Model for glaucoma service incorporating workup and testing stages with the decision to test modeled as a probabilistic switch.

These results are tabulated in Table 2.

Table 2: Input parameters and output response of the glaucoma clinic stochastic reward net model (Range of values in parenthesis).

Description	Clinic		
	SC A	SC B	SC C
Number of Viable Schedules	9	12	19
Average Total Patient	25.56 (21-31) std: 3.61	56.08 (46-63) std: 5.05	49.68 (38-65) std: 7.27
Average Concurrent Patients for Workup	18.11 (8-30) std: 5.95	33.00 (24-39) std: 3.98	28.53 (9-42) std: 9.51
Average Concurrent Patients for Imaging/Testing	84.11 (66-106) std: 13.89	82.50 (41-112) std: 19.67	91.11 (26-127) std: 22:09
Average Technicians Assigned	2.44 (2-3) std: 0.53	5 (5) std: 0	4 (4) std: 0
Patient-Tech Ratio	18.33 (15.33-26) std: 3.60	17.82 (14-20) std: 1.57	19.55 (13.75-26) std: 3.14

Average Imaging Capacity Available	7	7	7
Patient-Imaging Ratio	15.64 (13.57-17.43) std: 1.57	19.80 (14.14-24.29) std: 2.90	20.11 (10-24.14) std: 3.33
Average Check1 Training Error	4.60% (4.26% - 4.74%)	2.38% (2.00% - 3.10%)	5.60% (5.27% - 6.24%)
Average Check1 Validation Error	4.83% (3.37% - 6.14%)	2.82% (1.23% - 5.12%)	5.83% (4.40% - 7.85%)
Average Check2 Training Error	4.23% (4.04% - 4.52%)	5.53% (5.25% - 5.80%)	6.73% (6.60% - 7.03%)
Average Check2 Validation Error	5.47% (4.97% - 6.04%)	6.64% (5.84% - 7.31%)	7.02% (4.59% - 10.32%)

The results indicate a suitable representation for glaucoma clinics. SC A required additional consideration to set the number of technicians on a schedule-by-schedule basis, as the clinic saw the widest relative range in total demand for technicians, including external demand. Overall technician assignments were selected to minimize error while maintaining a similar patient-to-tech ratio. Counts for the selected imaging and technician staffing numbers are consistent with actual practices. Note when the average training errors are consistently smaller than validation values, as is generally expected. This pattern does not necessarily hold within each fold, due to effects of random sampling. Averaging across folds provides a better estimate of the generalization error by minimizing the effect of sampling. While further optimization may yield some improvement, the error is suitably minimized and invariant between samples. The resulting model proceeded to be tested against the withheld sample.

Having estimated the expected prediction error, the model can now be evaluated against the sequestered test set portion of the initial sample (Figure 21). Where parameterization optimized the model with respect to the training sample, refinements during cross-validation optimized the sample with respect to the validation samples. Removing this test portion prior to model refinement retains one independent sample for evaluation. We now evaluate the model performance using the SRN, inputs, and number of phases determined in previous sections. The model is now parameterized using all folds as one sample, comprising 3082 raw observations, evaluated against a validation set of 770 raw observations. Table 3 displays the test results as well as the difference versus the expected prediction error at both check points, where expected prediction error is estimated by average validation error (Table 2). Positive valued differences indicate where the test error improved over the expected prediction error, whereas negative values indicate degraded model performance on the test set. Overall, the model performs better on a new sample than estimated, with improvements outweighing degraded predictions. In most cases however note the relatively small value of differences. These summary statistics again indicate an adequate representation of the glaucoma clinics. The model accurately predicts future clinic flow.

Table 3: Test results using the sequestered data, arranged by clinic and check point.

Description	Clinic		
	SC A	SC B	SC C
Check1 Test Error	3.92%	3.18%	5.97%
Difference vs. Prediction	0.91%	(-0.36%)	(-0.14%)
Check2 Test Error	3.08%	4.12%	7.08%
Difference vs. Prediction	2.39%	2.52%	(-0.06%)

Concretely, the SRN model can be modified to evaluate new staffing or equipment allocations, new patient schedules, or new clinic flow decisions. While distributional models or direct statistics often provide close approximations of the desired distribution, the SRN representation adds the ability to modify the system representation in simulation, while precisely describing modifications.

The preceding experiment presented a more rigorous approach to refining robust models. We can conclude that the model produced represents a very complex, interdependent clinic environment, and has been validated against operations data from several example clinics. Replicating the modeling procedure systematically for multiple clinics results in comparable error. While this approach benefited from a relatively large, high quality set of operations data from eye care clinics, our approach is outlined for application to other clinics, including outpatient clinics outside of ophthalmology. The result achieves a satisfactory mix of clear specification, high fidelity to data, and adjustable system representation for evaluating subsequent plans.

Finally, while aggregate statistics such as the mean absolute error are often adequate for model selection, examining absolute error on a measurement by measurement basis provides additional insight into model validity. Figures 29 and 30 provide graphical representation of (1) the comparison between data and model cumulative distribution functions, (2) the absolute error, computed at each point where a measurement occurs in the test set, and (3) the number of observations at each value. For brevity these results are plotted for SC B, the highest volume clinic, at both check points.

In Figure 29 the uppermost plot compares the duration between check in and completion of a workup, computed from test data, plotted against model results. Absolute errors at each measurement value are shown in the middle plot, with counts of each measurement in the lowermost plot. These plots provide context for the mean absolute error, which is weighted based on observation frequency. While fewer counts are available showing imaging, error values are minimized for the more common measurements (Figure 30).

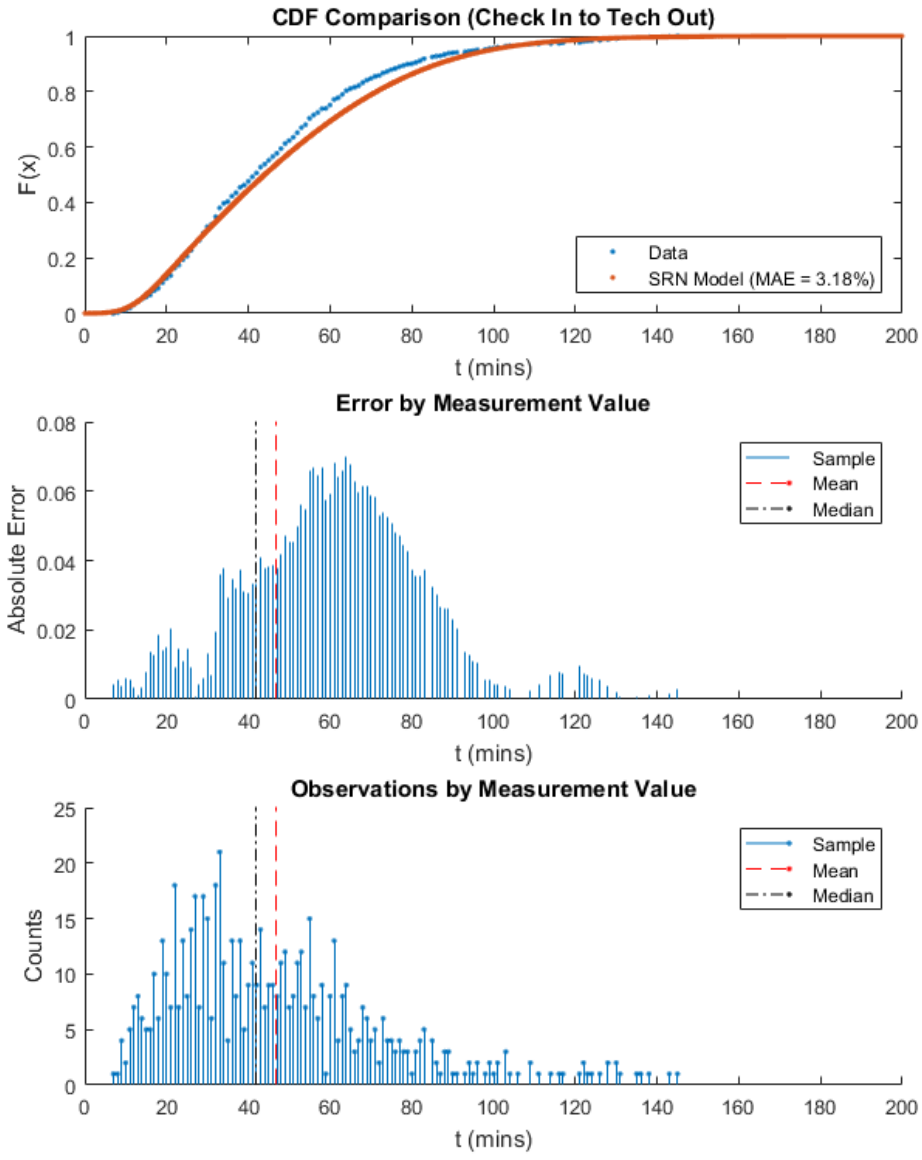


Figure 29: Test set results for SC B, at check point 1.

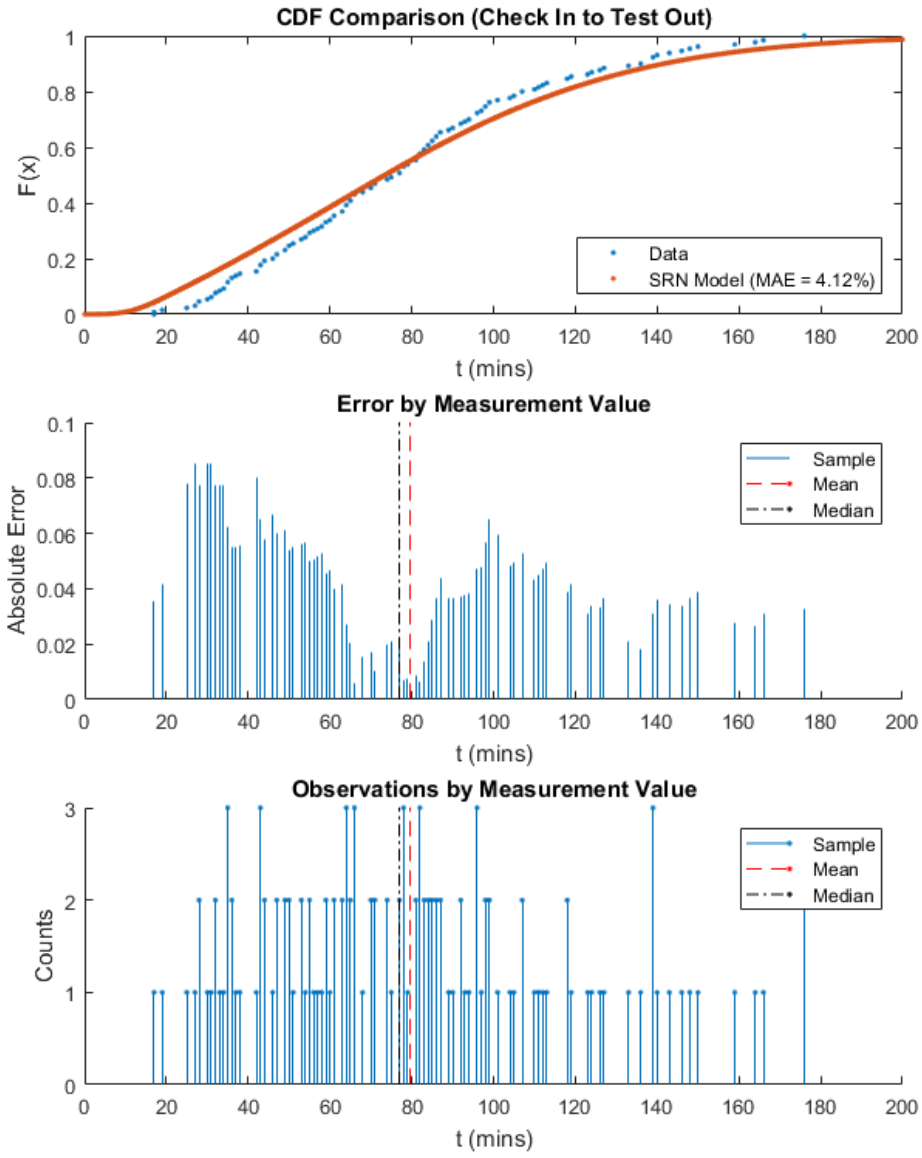


Figure 30: Test set results for SC B, at check point 2 show similar results.

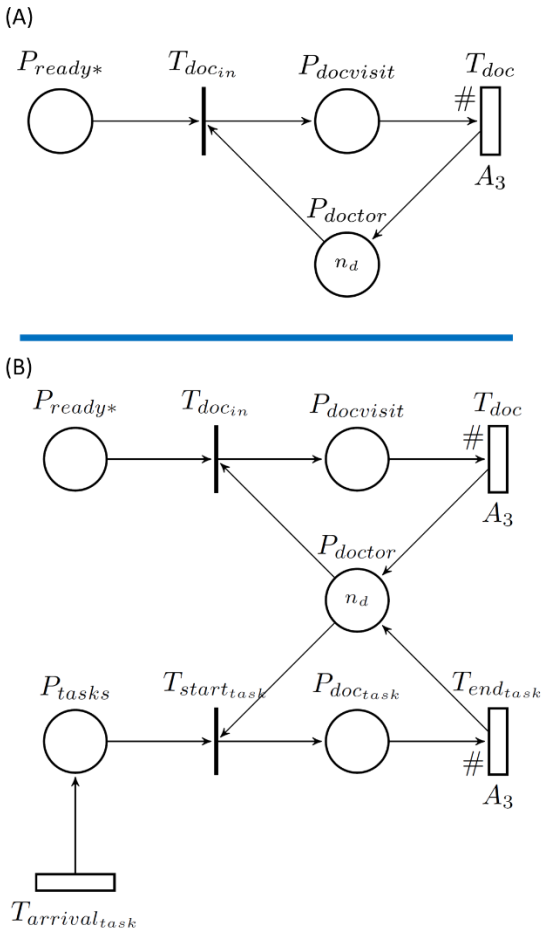
5.6 Accounting for Incongruencies Between Models and Practice

A key assumption in stochastic modeling methods, from queues to stochastic reward nets, is work conservation. Work conservation is a property of servers in a

performance model. A work conserving server does not idle when a task is available (Gianfranco Ciardo et al., 1993; Kleinrock, 1975; Trivedi, 2002). This assumption is problematic when the system represents human employees in health care. In modeling the workup, imaging, and testing, the model is tuned by adjusting the patient-to-technician ratio relative to the demand on those personnel (R. B. Fricks et al., 2018). In the case of the physician, who is unique in the clinic, the number of available doctors must be fixed at 1. Applying these constraints directly the model from Figure 28 can be extended by appending the Petri net in Figure 31A. The resulting mismatch between this approach and the recorded visit durations for 4084 patients in the extended test set is visible in Figure 32, and leads to significant mean absolute error of 15.8%.

Given that the mismatch is left-shifted, the model response is predicting shorter clinic visits than observed in practice. One explanation is that the work conserving assumption is incongruent with doctor's practices with patients. In observation, it is clear to see other tasks may interrupt treating patients. A doctor at DEC may file orders, respond to communique, perform charting activities, check on research, or even take short pauses, just to name a few example tasks. One strategy to represent the doctor unavailability seen in practice is to insert tasks at random intervals. A simple implementation is shown in the stochastic reward net in Figure 31B. In this net, tasks arrival according to a Poisson process, and the doctor attends to the next available patient or task on a first come, first served basis. For simplicity sake the doctor

examination time distribution is used for the task time distribution, which leaves one parameter, the arrival rate, for tuning purposes. A mean time between tasks of 20 minutes gives the result in Figure 32.



(A) treats the physician like a standard work-conserving server, that does not perform other tasks when a patient is available.

(B) inserts tasks according to a Poisson process with mean inter-task arrival time of 20 minutes. Tasks join a combined queue with patients, and the next patient or task is attended to on a first come, first served basis. For simplicity, tasks take as long as a patient visit with the same doctor, i.e. task completion time is drawn from the same distribution as patient-doctor visit time, a 13-phase PH distribution.

Figure 31: Two approaches to modeling doctor activity in a clinic.

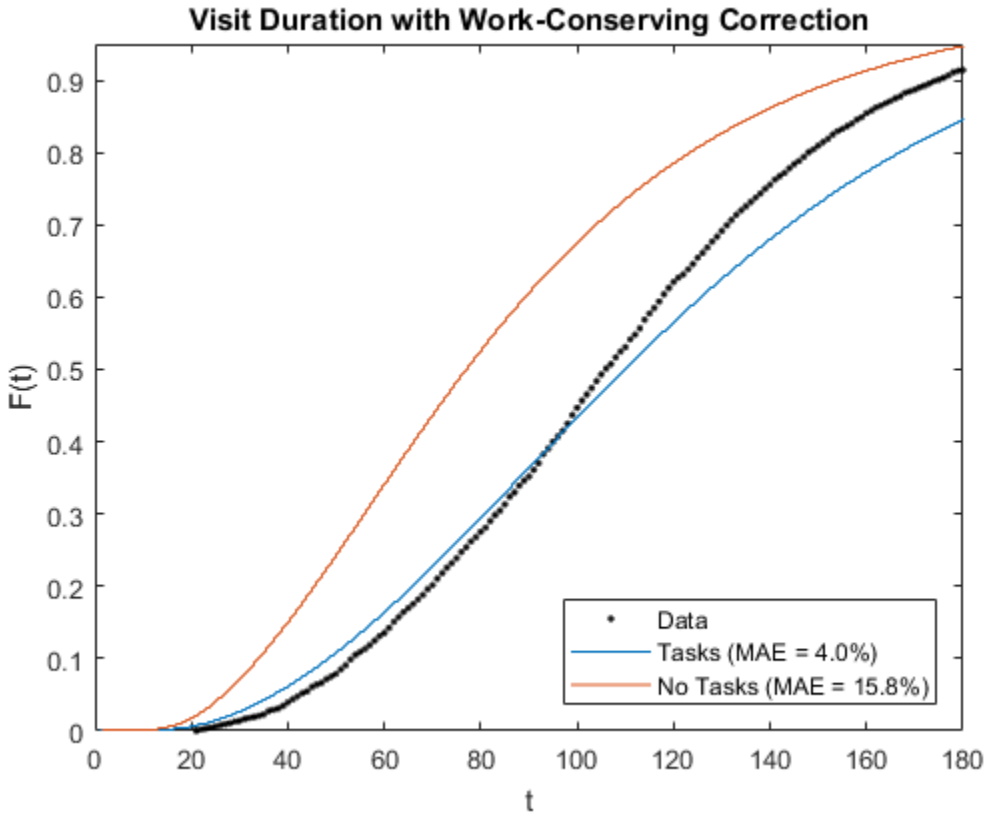
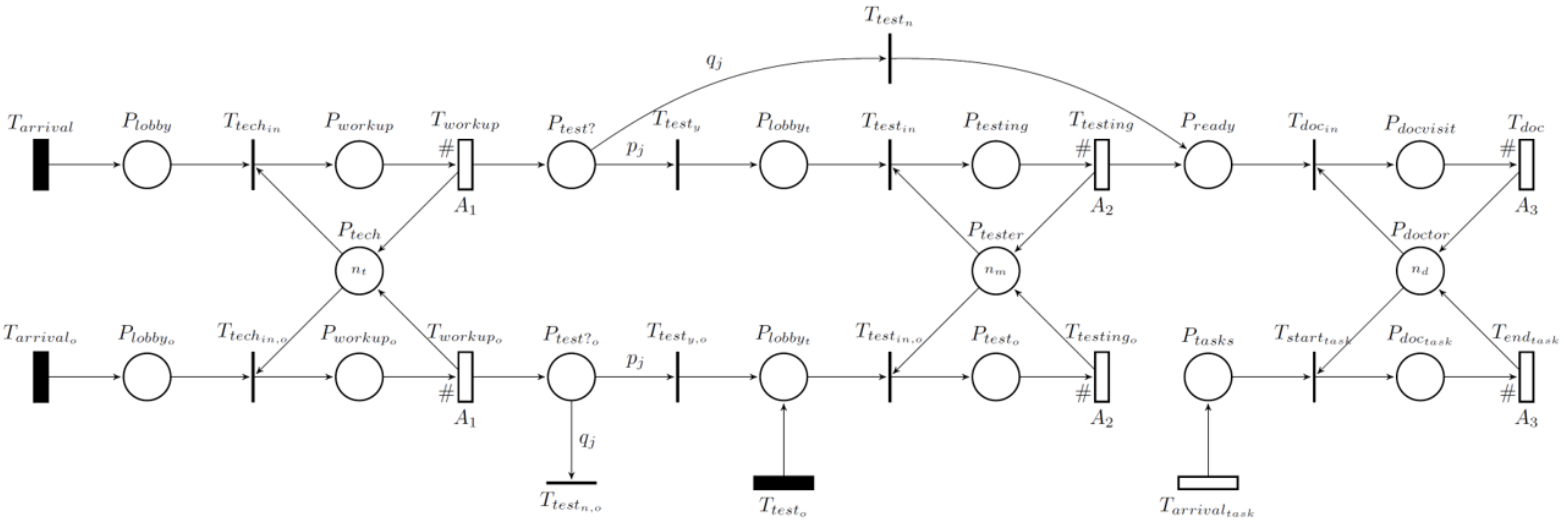


Figure 32: Comparison of model performance with and without inserted tasks.

The inclusion of tasks is necessary to accurately model physician business and overall clinic flow and visit duration. Inserting tasks to occupy the physician at random intervals provides effective correction for the work conserving property inherent in stochastic models such as queues and petri nets. In the absence of these tasks, simulated doctors are more available than their real counterparts, resulting in quicker service time than expected. Tasks are inserted according to a Poisson process, where the mean inter-task time is 20 minutes, and for simplicity tasks take as long to attend to as a patient visit. This technique can be retroactively applied to other treatment steps as needed.

5.7 Summarizing Clinic Modeling

The end result of the preceding iterative development is shown in Figure 33. Although further optimizations to model accuracy may be possible, the selected model exhibits consistently low error that does not vary significantly between independent samples. The model is tested using the cross-validation data pattern in Figure 21 on the complete data set in (R. Fricks et al., 2018), where 80% of the sample (18256 recorded visits to the glaucoma department) are used to parameterize the model. The model is tested against a 20% test set representing 4563 recorded patients visits. We achieve 3.94% test MAE comparing the model predictions for visit duration to recorded visit durations, while maintaining test errors at the intermediary check points below 4.5% (Figures 34-36). Invariant, low error is arguably superior to pursuing decreasing error at the small values observed in this study, given the known imperfections in our data. In previous attempts to independently assess the error of these measurements through in-person observation, and noted approximately a 3-5% discrepancy between in-person measurements and event logging (R. B. Fricks et al., 2018). The error exhibited by this model is comparable to the error between measurement methods. For operational decisions the reported error is effectively measurement noise. Similar white-box models for evaluating health care performance have not been produced or assessed for generalization.



This model, specified as a stochastic reward net (SRN), represents patient flow according to clinical practices (Chapter 3). Performance data collected over 18 months using event logging methods were verified through in-person observation for use in this model and anonymized contribution to an online research repository (Chapter 4). The model formulation fully uses SRN specifications, combined with discrete event simulation (Chapter 5). The Complete Glaucoma Clinic Model has been cross-validated for generalization performance, predicting independent test data with 4% total error. The white-box specification facilitates model-based evaluation of proposed quality improvement

Figure 33: Complete Glaucoma Clinic Model, based on outpatient practices at Duke Eye Center (DEC).

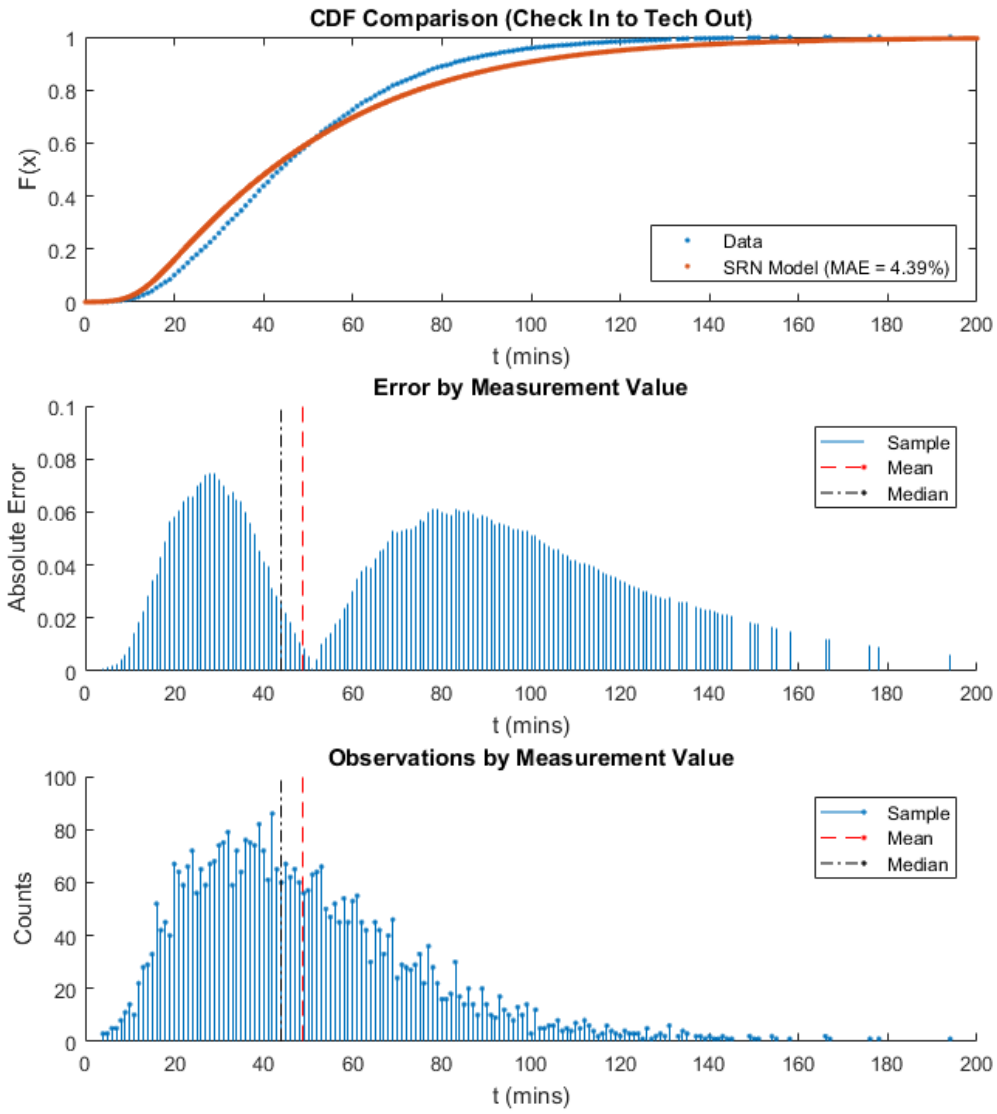


Figure 34: Comparison of check point 1 measurements to model predictions.

Check point 1 represents real world measurements of the time between check in and workup completion as registered by the tech out event. Viewing the by-measurement error detail we see this model matches central tendencies of

measurements at checkpoint 1 while providing an even balance of under and overestimation of clinic time.

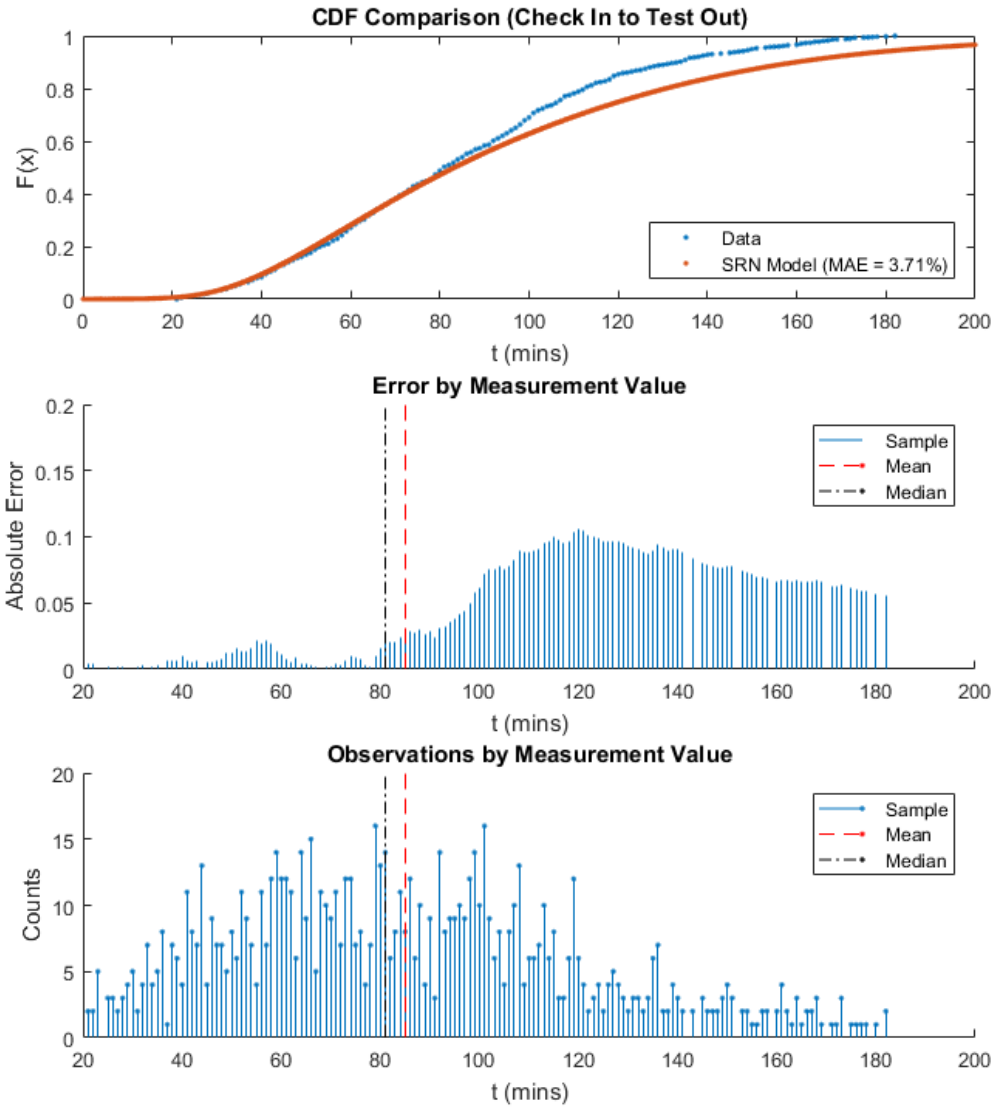


Figure 35: Comparison of check point 2 measurements to model predictions.

Comparison of check point 2 measurements to model predictions show similar overall pattern to check point 1, but with reduced mean absolute error. Notably the majority of visits, which complete required testing by the 100 minute mark, are predicted by the model with 5% or less error. Larger error is seen for longer check in to testing out durations, as the model predicts a conservatively long tail for more extreme values.

In an end-to-end comparison, the model achieves an average error of less than 4%, within the observed mismatch between in-person observation and event logging (Figure 36). The Complete model accurately predicts visit glaucoma practices at DEC in a systematic manner that was replicated at other DEC glaucoma practices.

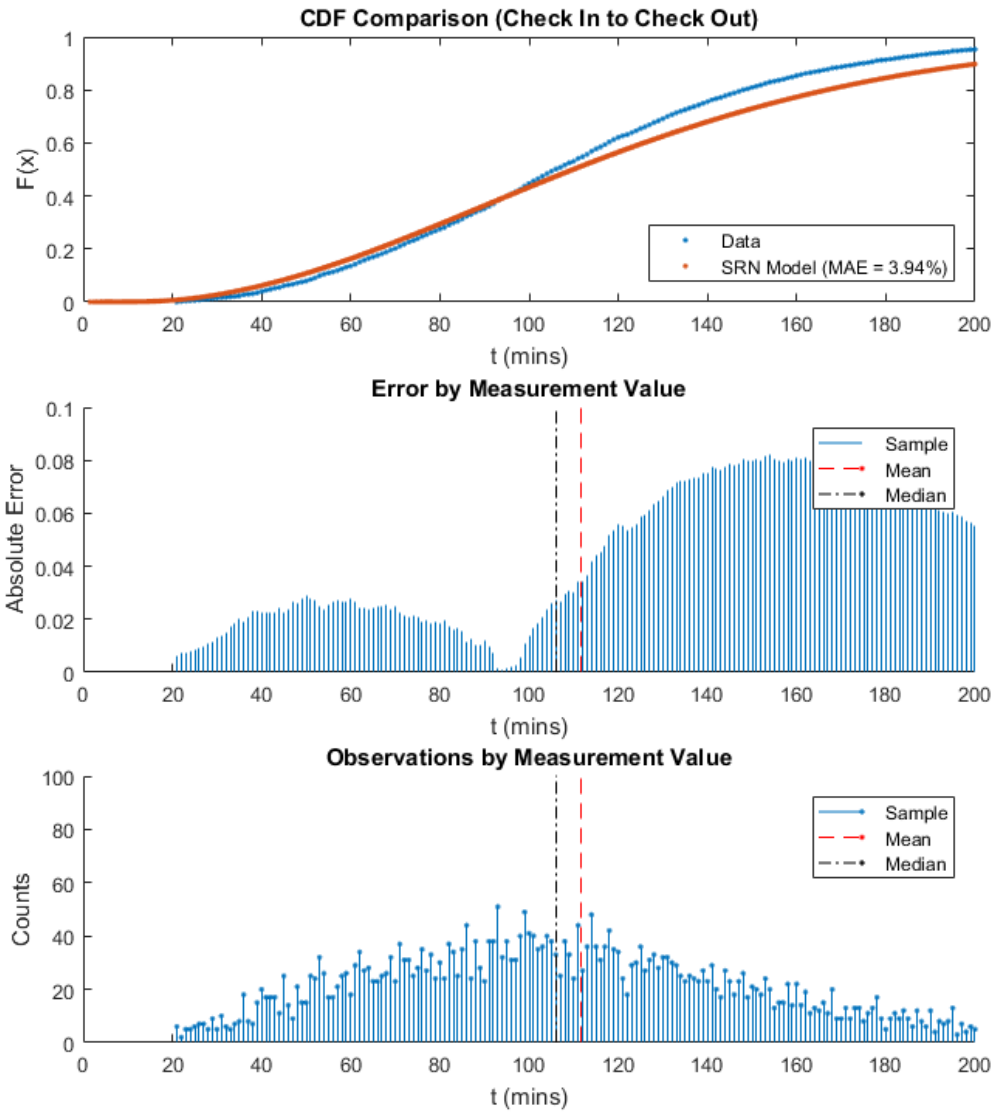


Figure 36: End-to-end comparison of simulation outputs versus recorded visit durations at a DEC glaucoma clinic over an 18-month interval.

6. Optimizing Clinic Performance

The resulting white-box Complete Glaucoma Clinic Model provides insight into the effectiveness of quality improvement decisions. However, quality of care is subjective, with extensive discussions on what constitutes care quality depending on the role in a health care system ((Brill, 2015; Kortbeek, 2012; Reid et al., 2005) present a few perspectives). Several competing notions exist in health care, such as utilization vs. idle resources (R. B. Fricks & Trivedi, 2016), or doctor versus patient wait times (Welch & Bailey, 1952).

The Complete Glaucoma model makes full use of stochastic reward nets (SRN) for modeling patient flow through clinics. An SRN model makes experiments with clinic schedule possible, to find scheduling rules that reduce wait times for patients and idle time for staff. This section applies heuristics to optimize wait times in a clinic, constraining the number of allowed appointments and staff to the real-time extracted schedules.

6.1 Dimensionality of Scheduling Input Space

To optimize clinic performance, the input space can be significantly reduced by observing a few practical limitations of a live patient schedule. Stated one way, we wish to generate n appointment slots at time t_n during the finite operating hours at the clinic. Restricting these appointment slots to the nearest 5 minute mark reduces the input space immediately from a continuous interval to a finite discrete set, while also limiting the

results to more realistic answers (it may be reasonable to tell a patient to arrive at 1:35 PM, but not 1:33 AM 52.8745 seconds). If a clinic operates from 7 AM to 5 PM, 10 hours are available to assign x appointments at 121 unique times. Without repetition, the simple combinatorial math shows that $2.26x10^{28}$ combinations are possible:

$$ {}_wC_x \equiv \binom{w}{x} \equiv \frac{w!}{x!\,(w-x)!} $$

Each permutation of schedule requires an execution of the SRN model. Simulating the results with every available trace-driven schedule for 5000 independent replications can be accelerated using parallelization, exploiting the independence of each simulative replication. Using 4 processing cores in this manner, Figure 35 plots execution times for 41 schedule permutations.

Testing each scheduling rule involves modifying and simulating the 61 original schedules for 5000 iterations per schedule. Execution employs the MATLAB parallel computing toolbox on a local pool of 4 workers. Discrete event simulation is naturally parallel work as several independent replications must be executed. The performance achieved here (Figure 37) on common hardware (Intel i7 4790k Processor, 32 GB DDR3 @ 2400 MHz) is reasonable for the application.

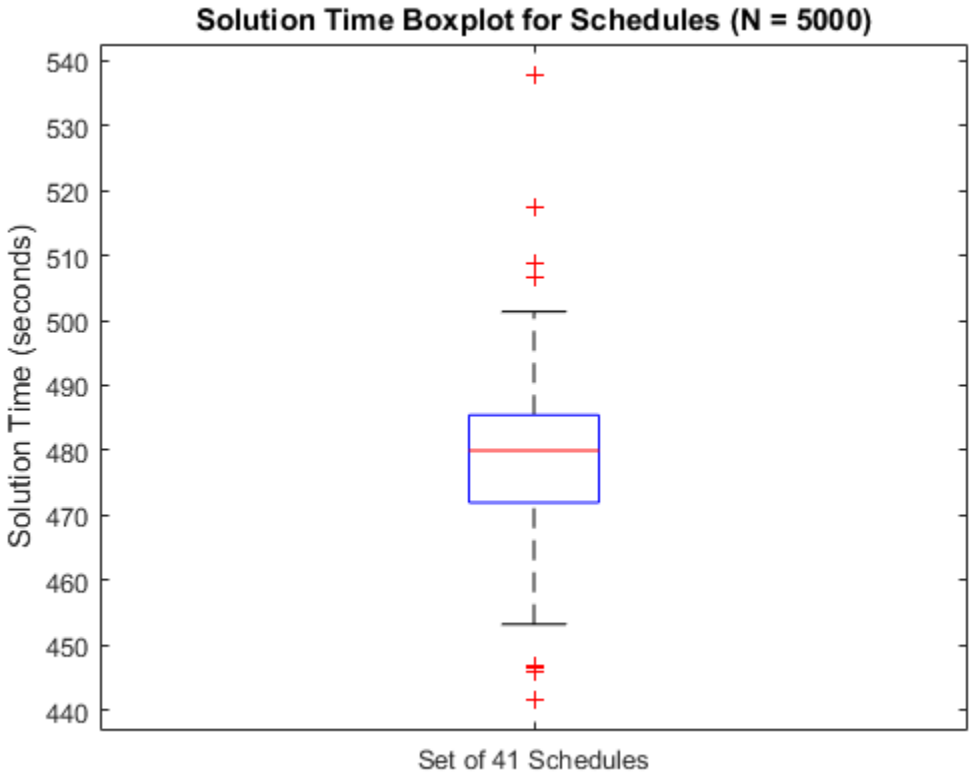


Figure 37: Execution time variability for 41 heuristic schedule generating methods.

At approximately eight minutes per schedule, solution time is non-trivial for the number of possible schedule permutations *before repeated time slots are considered*. Repeated time slots, or conversely batch arrivals, are a common feature in real clinics. Given the enormous range of possible schedules even when restricted to one class of patient at unique five-minute intervals, a heuristic approach is indicated.

The heuristic approach here is defined using the following terminology, based on current template schedules. These terms are illustrated in Figure 36 on a schedule example. Using a fixed lunch start time and duration, 40 schedule adjustments were

performed on existing schedules by specifying the batch interval and batch size. First batches, which occur at the beginning of clinic service and immediately after the lunch duration, can be specified to either (1) match the number of available technicians minus an offset value, or (2) remain a specified value.

1. Batch Size : the number of arrivals per appointment slot
2. Batch Interval : the period between repeating intervals
3. Lunch Start : a timepoint where arrivals are halted to allow for a lunch break
4. Lunch Duration : the duration of the halt in arrivals
5. First Batch Size : a specified batch size for the first arrival at start of day or end of lunch

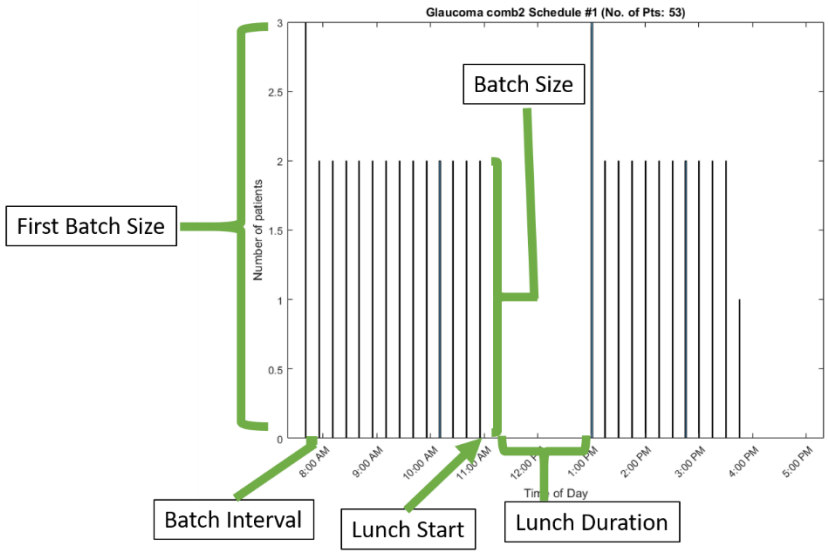


Figure 38: Schedule diagram, illustrating terminology for templated schedules.

6.2 Trade-Off Between Patient Wait Time and Staff Idle Time

By varying these features except for lunch start and duration, 40 scheduling rules were used to adjust existing schedules. These scheduling rules test batch sizes ranging from 1-3, intervals ranging from 5-30 minutes, and a variety of first-batch sizes. The

tradeoff between patient wait times and personnel idle time for these 40 scheduling rules is shown in Figure 37. Each point represents 61 schedules extracted from the event logging data, where each schedule is adjusted by a heuristic and simulated 5000 times. In this situation alone, we can define two competing notions of quality care; the time a patient spends waiting, versus the time between readying patients to see a physician.

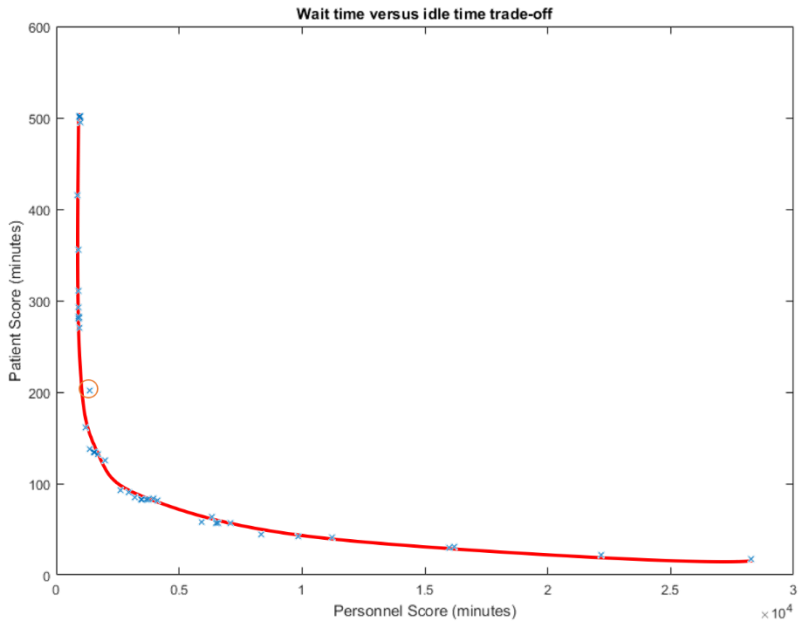


Figure 39: Patient wait time to personnel idle time tradeoff curve, with real schedule performance indicated by the orange circle.

Wait time versus total idle personnel time follows an apparent trend, indicated qualitatively in red. The heuristic schedules generated occupy many points along the optimal tradeoff points at the base of this curve. These schedules minimize patient wait times as well as personnel idle times compared to most alternatives, including the original schedules. However, simulating the white-box SRN model provides much

richer system information which can be used to further distinguish the performance of these schedules, as will be now examined.

6.3 Enhanced Quality Metrics

The clinic model estimates clinic performance under the input (X) schedule and staffing conditions. The model outputs are various clinic performance variables, where visit durations at various steps were the validation focus. However, the status of personnel, distribution of idle times, and patient wait time due to staff unavailability are also available from each execution. Fifteen possible metrics (M1-M15) related to health care quality are now defined, in patient-centric and personnel-centric categories.

1. Patient-Centric
    a. Median Wait time (M1), first quartile (M2), third quartile (M3)
        i. Minutes waited
    b. Wait targets by stage (M4-6) – "Wait no more than 20 minutes for workup"
        i. Percent of patients meeting this goal
    c. Wait target total (M7) – "Wait no more than 60 minutes total"
        i. Percent of patients meeting this goal
    d. Wait fraction total (M8-9) – "Wait no more than 70% of visit"
        i. Percent of patients meeting this goal
2. Personnel-Centric
    a. Median idle time
        i. Technician (M10)
        ii. Imaging or testing (M11)
        iii. Doctor (M12)

iv. Measured as minutes idle, normalized by number of personnel available

b. Overtime
   i. Technician (M13)
   ii. Imaging or testing (M14)
   iii. Measured as average labor minutes past 5PM

c. Doctor comfort target – "wait one time 30 or more minutes, but otherwise wait no more than 30 minutes between patients" (M15)
   i. Probability of meeting this goal

An enhanced patient cost and personnel cost can be formulated from these metrics to further distinguish schedules, using the transformation $g = 1 - log(x)$, where $\log(x)$ is the natural logarithm of x and the transformation inversely remaps the X value from (0-1) to (∞ – 4). The Q term therefore penalizes wait times that do not meet quality metrics above.

$$Patient\ Cost = Q \cdot (W_{med} + W_{Q1} + W_{Q3})$$

$$Q = \sum_{i=1}^{5} (1 - \log(q_i))$$

$$q_1 = Tech\ goal\ fraction$$

$$q_2 = IMVF\ goal\ fraction$$

$$q_3 = Doctor\ goal\ fraction$$

$$q_4 = total\ goal\ fraction$$

$$q_5 = mean\ value\ added\ fraction$$

Similarly for personnel metrics,

$$Personnel\ Cost = (1 - \log(q_6)) \cdot (I_{t,med} + I_{t,OT} + I_{i,med} + I_{i,OT} + I_{d,med})$$

$$q_6 = Comfort\ goal\ fraction$$

The resulting schedule scores can be plotted in logarithmic space with patient costs as the Y axis and personnel costs as the X axis (Figure 38). The cost functions in combination with the defined transformation more clearly separates scheduling alternatives. The original schedule is indicated as 'Copy,' with notable alternatives indicated. Two cost functions are defined in the form of personnel cost and patient cost which factor enhanced metrics into consideration, to further distinguish schedule performance in this plot. Higher cost indicates worse performance from the perspective of either patient or personnel, i.e. increased patient cost can be due to increased wait times.

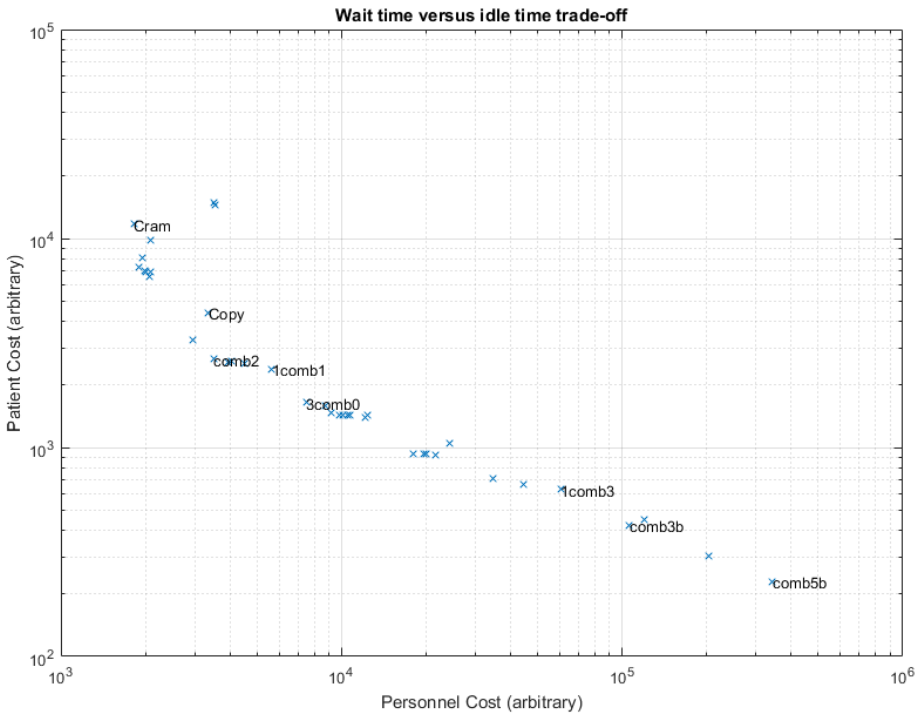


Figure 40: Scheduling rules plotted by cost function on logarithmic axes.

6.4 Clustering through Unsupervised Learning of Quality Metrics

Clustering the schedule performance may delineate competing notions of health care quality. Figure 29 shows that some schedules provide clear improvement over another alternative, such as the original schedule ('Copy') and the template pattern 'comb2' (diagrammed in Figure 27). It is useful now to determine if schedules correspond to any underlying archetype in terms of their scoring, to suggest alternative schedules with similar performance. K-means clustering is a form of unsupervised learning for identifying similarity in multi-dimensional data (Hastie et al., 2009). It is possible to apply k-means clustering with k = 3 to group score results as shown in Figure

30. These three groups ideally correspond to patient-centric, personnel-centric, and balanced performance schedules. However, as some heuristics such as 'comb5b' are worst-case behaviors, the k-means algorithm struggles with sub-classification in this application for the schedules with more realistic performance. As an example, adding additional clustering groups will subdivide Group 3 in Figure 30 as the dissimilarity between those schedules is greater than dissimilarity between all other schedules. The solution is to recursively apply clustering within the largest group, Group 2, as shown in Figure 38.

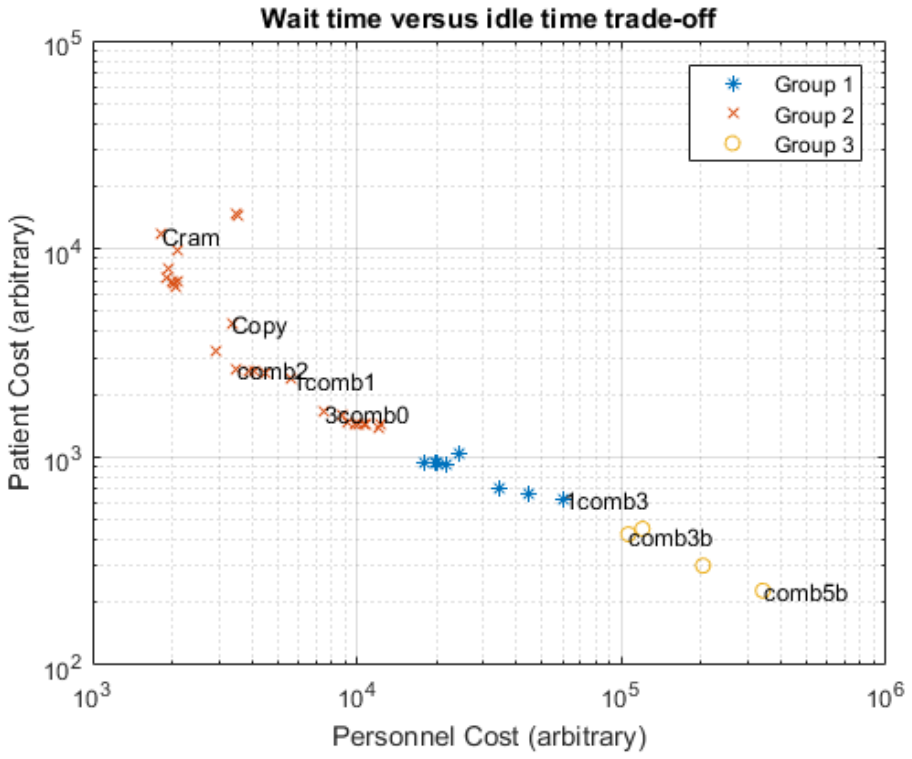


Figure 41: Preliminary clustering using k-means algorithm with k = 3.

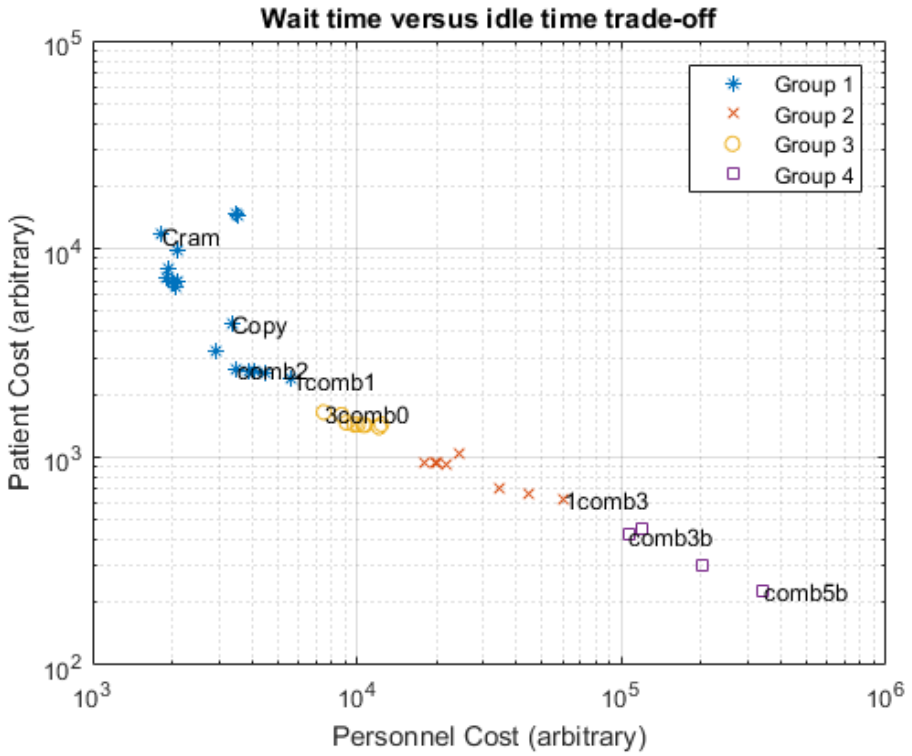


Figure 42: Functional grouped determined through iterative re-application of k-means clustering

The clustering differentiation now more properly distinguishes patient-centric schedules (group 2), personnel-centric schedules (group 1), balanced schedules (group 3), and extreme schedules (group 4). These clusters allow us to evaluate performance of one schedule in relation to its similar-purpose, similar-performing peers.

6.5 Schedule Selection Using Performance Clusters

Similar to health care models, a vast menagerie of optimization methods have been developed with tradeoffs in applicability, computation time, and ability to guarantee an optimal solution (Ahmadi-Javid et al., 2017; Cayirli & Veral, 2003). Considering the functional requirements of any analysis in this domain is a useable

clinic schedule allows for subjective curtailing of the variety of inputs to a sparse set. These inputs are far easier to interpret for the end personnel who must implement these schedules, and follow rules such as '2 patients every 15 minutes' rather than a complex representation of arrival as a stochastic process.

Though the 40 heuristics are a sparse sampling of all possible schedules, in evaluating the wait-time to idle-time tradeoff alone (Figure 3) we find that these simple rules sample this tradeoff reasonably well to infer a tradeoff trend and provide better tradeoffs than the baseline schedule.

The more complex metrics introduced in the cost functions further distinguishes these schedules by incorporating additional notions of quality. Schedules that provide similar wait times and personnel idle times may provide a more favorable balance of other quality factors, which have now been quantified. Plotting the cost values can be used to find superior performing alternatives.

However, determining what constitutes a 'superior' performance is still clinic-specific. For instance 'comb2' provides a superior choice to existing schedules, provided the goal is a personnel-centric system that reduces wait times at an acceptable personnel cost. The 'comb2' heuristic can be concisely described; it matches initial batches with the number of technicians available, then evenly divides 2 patient slots every 15 minutes proportionally among the clinics operating that day. This schedule template outperforms the existing template in nearly all enhanced metrics, while still maintaining

a quality of care profile similar to the existing schedule as indicated by the performance groups. Table 4 lists performance scores for each schedule.

Other clinics may balance quality further in the direction of personnel idle times, such as '3comb0,' or opt for a spare schedule such as '1comb3' that guarantees all patient guarantees are met even when individual visits are unusually long. 3comb0, which specifies a first batch size of 5, then 3 patients every 30 minutes, provides a more balanced move in the adjacent performance cluster (Table 4). The coarse-grain selection at performance group levels is subject to more economic and policy decisions, however within each performance group clear alternatives exist. Comb2 provides similar performance to the existing schedules labeled 'Copy,' with only slight increases in imaging and doctor idle time (M11, M12), and imaging over time (M14). 3Comb0 goes further in improving patient-centric metrics at the expense of staff, who see increases in idle and over time for reducing median wait times by half.

New methods may further sample the schedule space. The functional limits of clinic performance seem to indicate that new alternatives will fall within the inferred wait-to-idle tradeoff as has been frequently observed (Cayirli & Veral, 2003; Welch & Bailey, 1952). We further postulate that new alternatives will fit within the predicted performance groups. Improved system understanding, rather than input space sampling, may provide more insight into clinic performance.

Table 4: Comparison of scheduling heuristics by performance metrics.

Metric	Copy	Comb2	3Comb0
Wait Time, Median (minutes, M1)	64	42	39
Wait Time, First Quartile (minutes, M2)	31	17	15
Wait Time, Third Quartile (minutes, M3)	105	78	72
Wait target achieved, workup (percent, M4)	0.58	0.76	0.73
Wait target, imaging or testing (percent, M5)	0.79	0.83	0.86
Wait target, doctor (percent, M6)	0.56	0.64	0.68
Wait target, total (percent, M7)	0.54	0.7	0.74
Fraction of patients who wait no more than 30% of visit (percent, M8)	0.15	0.26	0.38
Portion of visit spent in treatment (percent, M9)	0.44	0.53	0.55
Technician Idle Time (minutes, M10)	369	307	410
Imaging Idle Time (minutes, M11)	827	860	1080
Doctor Idle Time (minutes, M12)	85	99	116
Technician Overtime (minutes, M13)	0	0	67
Imaging Overtime (minutes, M14)	68	98	297
Doctor has a break? (percent, M15)	0.88	0.87	0.79

7. Conclussions

At academic centers like Duke Eye Center, at times it may seem as though all options have been considered. At those points especially, predictive modeling is necessary to continue to make forward progress as opportunities for improvement become less obvious. Even as various quality improvement decisions have been conceived for this system, the capability for predicting those how those decisions will impact quality of care has thus far been nonexistent in this large practice. Changes must be tested live and measured in retrospect. The ability to evaluate those decisions prior to implementation allows administrators to predict effectiveness. Predicting an observed change allows an organization to more conclusively discriminate between legitimate improvements or fortunate clinic days. The capability to test arbitrary scenarios enables bolder, unforeseen clinic flow decisions.

For organizations to base decisions confidently on model results, modelers must consolidate standards for evaluating models for predictive performance. There is currently no standard approach to modeling, although many studies use similar methods in functionally identical practices (Gunal & Pidd, 2010). As a consequence, modelers may produce equal but different representations of the same system (Breiman, 2001), through the choice of how to represent the same information. In essence this is acceptable as different modeling approaches present different tradeoffs in the information presented, computational cost, and accuracy. However for consistent

advancement of modeling research and its adoption by organizations, each subjective choice must ultimately be objectively evaluated. Evaluation methods for predictive modeling in data-rich fields such as machine learning may serve as a template as data becomes more widely available with better access to improved clinic performance measurement.

The systematic, end-to-end approach employed here is detailed for replication in other health care practices. From the descriptions of clinical practices at Duke Eye Centers, to findings from in-person verification of the data, and the production of an anonymized repository, the hope is that others will attempt to replicate this approach or improve on the preceding results with the clinic performance measurements available as a basis for comparison. Replication, along with the exchange of data that protects individual patient data, provides an opportunity for objective refinement of health care models. The current deficit of data resources and comparative studies contributes to the sense of slow progress in model-based health care quality improvement.

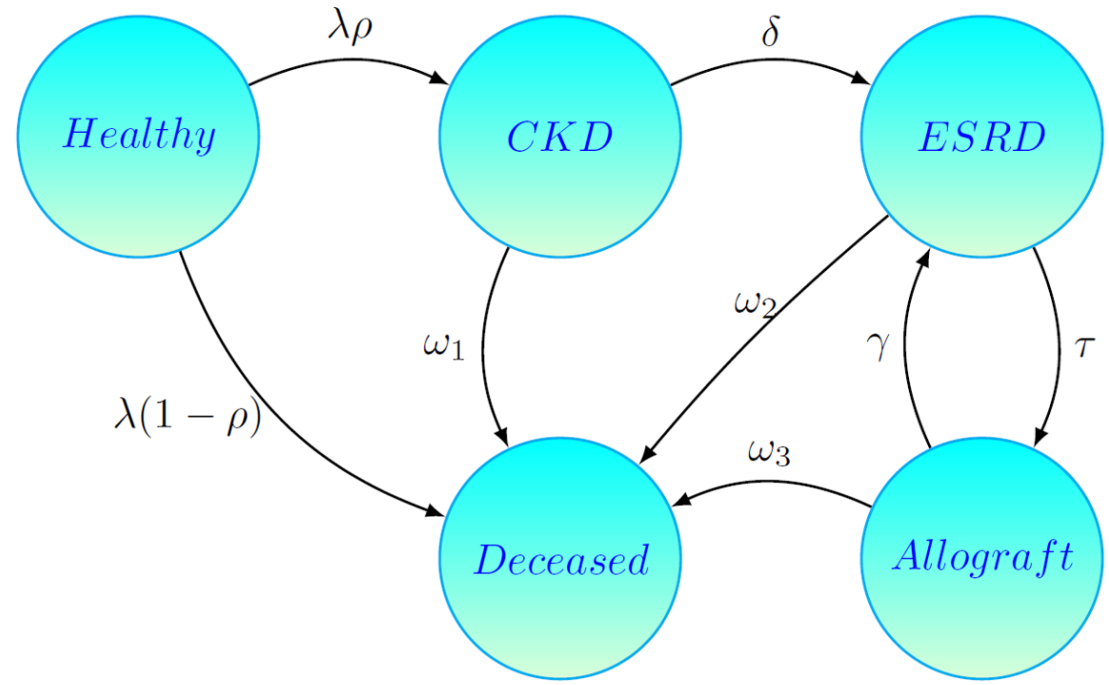
As the cost of health care continues to rise, every successful effort to improve health care efficiency helps lighten a burden that disproportionately affects the most vulnerable in society. It is heartening to know that research efforts, in this dissertation and elsewhere, continue the search for tools that help dedicated health care professionals in delivering high quality care as effectively as possible.

Appendix A: Analytic Modeling in Chronic Kidney Disease

One of the strengths and appeals of analytic models are the purported compact and efficient representation. Such models have a clear relationship between variables and may be represented relatively compactly in some cases. Here a hypothetical model of progression in chronic kidney disease is presented as an illustrative example of the pitfalls of seemingly innocuous models. This model was solved through numeric integration approximate methods when described in (R. B. Fricks et al., 2016), however it is feasible to seek closed-form analytic solutions for this model, which are presented here.

Suppose the following continuous time Markov chain (CTMC) model is proposed as a model for patient prognosis in chronic kidney disease (CKD). As in (R. B. Fricks et al., 2016), for short time frames it is reasonable to assume the transition rates are time homogeneous, i.e. they are constant as a function of time. The CTMC is illustrated in Figure 23, along with the infinitesimal generator matrix that, along with state probabilities, fully specifies the model. This model abstracts patient health into five states, requires seven transition rates to be estimated for a given patient, as well as a vector of initial probability of being in each health state. Presuming the patient is initially healthy, it is feasible to derive expressions for the transient probability of being

in each of the five health states at any arbitrary subsequent time. The precise derivation is now detailed.



$$\begin{bmatrix} -\lambda & \lambda\rho & 0 & 0 & \lambda(1-\rho) \\ 0 & -(\delta+\omega_1) & \delta & 0 & \omega_1 \\ 0 & 0 & -(\tau+\omega_2) & \tau & \omega_2 \\ 0 & 0 & \gamma & -(\gamma+\omega_3) & \omega_3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 43: Time-homogenous continuous time Markov chain model of chronic kidney disease progression where cadaveric transplantation is possible.

The model specified can be written as a series of differential equations

$$\frac{d\pi_H(t)}{dt} = -\lambda\pi_H(t)$$

$$\frac{d\pi_C(t)}{dt} = -(\delta + \omega_1)\pi_C(t) + \lambda\rho\pi_H(t)$$

$$\frac{d\pi_E(t)}{dt} = -(\tau + \omega_2)\pi_E(t) + \delta\pi_C(t) + \gamma\pi_A(t)$$

$$\frac{d\pi_A(t)}{dt} = -(\gamma + \omega_3)\pi_A(t) + \tau\pi_E(t)$$

$$\frac{d\pi_X(t)}{dt} = \lambda(1-\rho)\pi_H(t) + \omega_1\pi_C(t) + \omega_2\pi_E(t) + \omega_3\pi_A(t)$$

Using the Laplace transform converts differential equations into algebraic expressions of the form below, where the initially healthy assumption is introduced as a (-1) term in the first equation.

$$s\overline{\pi_H}(s) - 1 = -\lambda\overline{\pi_H}(s)$$

$$s\overline{\pi_C}(s) = -(\delta + \omega_1)\overline{\pi_C}(s) + \lambda\rho\overline{\pi_H}(s)$$

$$s\overline{\pi_E}(s) = -(\tau + \omega_2)\overline{\pi_E}(s) + \delta\overline{\pi_C}(s) + \gamma\overline{\pi_A}(s)$$

$$s\overline{\pi_A}(s) = -(\gamma + \omega_3)\overline{\pi_A}(s) + \tau\overline{\pi_E}(s)$$

$$s\overline{\pi_X}(s) = \lambda(1-\rho)\overline{\pi_H}(s) + \omega_1\overline{\pi_C}(s) + \omega_2\overline{\pi_E}(s) + \omega_3\overline{\pi_A}(s)$$

A transient expression for the healthy state can be found trivially by rearranging the first equation and performing one inverse Laplace transform.

$$\overline{\pi_H}(s) = \frac{1}{s+\lambda} \Leftrightarrow \pi_H(t) = e^{-\lambda t}$$

A transient expression for the CKD state can be similarly found

$$\overline{\pi_C}(s) = \frac{\lambda\rho}{s+\delta+\omega_1}\overline{\pi_H}(s) = \frac{\lambda\rho}{(s+\delta+\omega_1)(s+\lambda)} = \frac{\lambda\rho}{\delta+\omega_1-\lambda}\left(\frac{1}{s+\lambda} - \frac{1}{s+\delta+\omega_1}\right)$$

$$\overline{\pi_C}(s) \Leftrightarrow \pi_C(t) = \frac{\lambda\rho}{\delta+\omega_1-\lambda}\left[e^{-\lambda t} - e^{-(\delta+\omega_1)t}\right]$$

By the transplant (allograft) state, the expressions become substantially larger

$$(s+\tau+\omega_2)\overline{\pi_E}(s) = \delta\overline{\pi_C}(s) + \gamma\overline{\pi_A}(s)$$

$$\overline{\pi_E}(s) = \frac{\delta}{(s+\tau+\omega_2)}\overline{\pi_C}(s) + \frac{\gamma}{(s+\tau+\omega_2)}\overline{\pi_A}(s)$$

$$s\overline{\pi_A}(s) = -(\gamma+\omega_3)\overline{\pi_A}(s) + \frac{\delta\tau}{(s+\tau+\omega_2)}\overline{\pi_C}(s) + \frac{\gamma\tau}{(s+\tau+\omega_2)}\overline{\pi_A}(s)$$

$$\left(s+\gamma+\omega_3-\frac{\gamma\tau}{(s+\tau+\omega_2)}\right)\overline{\pi_A}(s) = \frac{\delta\tau}{(s+\tau+\omega_2)}\left[\frac{\lambda\rho}{\delta+\omega_1-\lambda}\left(\frac{1}{s+\lambda}-\frac{1}{s+\delta+\omega_1}\right)\right]$$

Rearranging and combining terms,

$$\overline{\pi_A}(s) = \left(s+\gamma+\omega_3-\frac{\gamma\tau}{(s+\tau+\omega_2)}\right)\left(\frac{\delta\tau}{(s+\tau+\omega_2)}\right)\left[\frac{\lambda\rho}{\delta+\omega_1-\lambda}\left(\frac{1}{s+\lambda}-\frac{1}{s+\delta+\omega_1}\right)\right]$$

$$\overline{\pi_A}(s) = \frac{\delta\rho\lambda\tau\big(\gamma(\omega_2+s)+(\omega_3+s)(\omega_2+\tau+s)\big)}{(\lambda+s)(\delta+\omega_1+s)(\omega_2+\tau+s)^2} = \frac{A}{\lambda+s}+\frac{B}{\delta+\omega_1+s}+\frac{Cs+D}{(\tau+\omega_2+s)^2}$$

From here, the placeholder terms A, B, C, D facilitate partial fraction expansion. Solving the new set of equations for A, B, C, D, they can be substituted in the time-domain equation below.

$$\begin{array}{ccc}
0 & = & A+B+C \\
\delta\rho\lambda\tau & = & A(2\tau+2\omega_2+\delta+\omega_1)+B(2\tau+2\omega_2+\lambda)+C(\lambda+\delta+\omega_1)+D \\
\delta\rho\lambda\tau(\gamma+\tau+\omega_2+\omega_3) & = & A\big((\tau+\omega_2)^2+2(\delta+\omega_1)(\tau+\omega_2)\big)+B\big((\tau+\omega_2)^2+\lambda 2(\tau+\omega_2)\big)+C\lambda(\delta+\omega_1)+D(\lambda+\delta+\omega_1) \\
\delta\rho\lambda\tau(\gamma\omega_2+\tau\omega_3+\omega_2\omega_3) & = & A\big((\delta+\omega_1)(\tau+\omega_2)^2\big)+B\lambda(\tau+\omega_2)^2+D\big(\lambda(\delta+\omega_1)\big)
\end{array}$$

$$\pi_A(t) = Ae^{-\lambda t}+Be^{-t(\delta+\omega_1)}-e^{-t(\tau+\omega_2)}(-C-Dt+C\tau t+C\omega_2 t)$$

Using a symbolic solver in Wolfram Mathematica 2014,

$$\left\{\left\{A \to \frac{\delta\,\lambda\,\rho\,\tau\,(\gamma\,(-\lambda+\omega 2)+(\lambda-\tau-\omega 2)\,(\lambda-\omega 3))}{(\delta-\lambda+\omega 1)\,(-\lambda+\tau+\omega 2)^2},\right.\right.$$

$$B \to \frac{\delta\,\lambda\,\rho\,\tau\,(\gamma\,(\delta+\omega 1-\omega 2)-(\delta-\tau+\omega 1-\omega 2)\,(\delta+\omega 1-\omega 3))}{(\delta-\lambda+\omega 1)\,(\delta-\tau+\omega 1-\omega 2)^2},$$

$$C \to -\frac{\delta\,\lambda\,\rho\,\tau\left(\gamma\left(\tau^2+\delta\,(-\lambda+\omega 2)+(\lambda-\omega 2)\,(-\omega 1+\omega 2)\right)+(\lambda-\tau-\omega 2)\,(\delta-\tau+\omega 1-\omega 2)\,(\tau+\omega 2-\omega 3)\right)}{(\delta-\tau+\omega 1-\omega 2)^2\,(-\lambda+\tau+\omega 2)^2},$$

$$D \to \frac{1}{(\delta-\tau+\omega 1-\omega 2)^2\,(-\lambda+\tau+\omega 2)^2}$$

$$\delta\,\lambda\,\rho\,\tau\left(\gamma\left(\delta\left(\tau^2+(\lambda-\omega 2)\,\omega 2\right)+\lambda\left(\tau^2+(\omega 1-\omega 2)\,\omega 2\right)-(\tau+\omega 2)\left(2\,\tau^2+(\omega 1-\omega 2)\,\omega 2+\tau\,(-\omega 1+\omega 2)\right)\right)+\right.$$

$$\left.\left.(\lambda-\tau-\omega 2)\,(\tau+\omega 2)\,(-\delta+\tau-\omega 1+\omega 2)\,(\tau+\omega 2-\omega 3)\right)\right\}\right\}$$

Figure 44: Mathematica Output for equation solver for A,B,C,D terms as functions of the model parameters

Equations for the remaining two states are similarly cumbersome but can be found by repeating the above process. These equations are only suitable for the assumption that a patient begins in the healthy state with probability 1. For arbitrary initial probability, the same process can be repeated and results for each transient state probability can be scaled by the initial probability and summed for a complete expression. Alternatively, this problem can be solved through numeric integration of the ordinary differential equations that describe the Markov chain as was performed in (R. B. Fricks et al., 2016), or trivial simulation. The computerized solutions are easier to adjust, should evidence warrant a different model structure. This kidney disease model illustrates the distinct disadvantages in analytic solution when even seven unique variables are present. The advantages in computation time are overshadowed by the difficulty in entering the solution equations alone.

Appendix B: Template for Observation

Observer	Rafael Fricks	
Observation Time	7:45 AM	10:40 AM
Observation Date	5/4/2016	
Provider	Henry Tseng	
Workup Tech	Marjorie Veihl	
Imaging Tech	Mike Kelly	
Visit Type	Return Visual Field	
Appointment Time	8:00 AM	
	IN (HH:MM AM)	**OUT (HH:MM AM)**
Workup		
Tech	8:33 AM	8:59 AM
Imaging		
OCT	9:12 AM	9:34 AM
Autofluoresence		
Color Fundus		
Slit Lamp Photos		
Binocular optic nerce photo		
Fluoroscence Angiogram		
ICG Angiography		
Humphrey Visual Field	9:40 AM	10:01 AM
Octopus (goldman) Visual Field		
Physician Contact		
Resident		
Fellow	10:15 AM	10:23 AM
Attending	10:23 AM	10:33 AM
Waits		
Dialation		
Imaging		
Testing		
Physician	10:01 AM	10:15 AM
Other		

Figure 45: Example observation template form used in data collection.

Appendix C: Core of Simulation Program

The core simulation function is included here with annotation and syntax highlighting. This implementation in the MATLAB scripting language returns the entire simulation trace at completion.

```
function [result, delays, trace] =
clinSim_complete(deltas,occurs,n_techs,n_vf,n_doc)
%%% SET UP SYSTEM STATE VARIABLES %%%
[n_slots, n_stages] = size(deltas);
% I want in trace
% 1 - time
% 2 - patient
% 3 - stage complete
% 4 - n_techs idle
% 5 - n_vf idle
% 6 - n_doc idle
trace = zeros(n_slots*n_stages,5);

% Time/Transitions %
current_time = 0;
t_enabled = ones(n_stages,1);

% Patients %
arrival_ind = 1;
n_in_sys = 0;
arrival_times = deltas(:,1);

% Staff %
% next_ind = 0;
n_done = 0;
% n_techs = 1;

w_tech = -1*ones(n_techs,1);
w_vf = -1*ones(n_vf,1);
w_doc = -1*ones(n_doc,1);

techQueue = [];
vfQueue = [];
docQueue = [];


tick = 1;

phases = ones(n_slots,1); % pointer for what phase of treatment a
patient is in
```

```
delays = zeros(n_slots,n_stages);

result = zeros(n_slots,n_stages);
% result(:,1) = deltas(:,1);

active = ones(n_slots,1);
%       // 0 for out of system,
%       // 1 for active
%       // 2 for waiting

%%% PROCESS EVENTS %%%
while(1) %until the simulation is complete
    % while(tick<50)
    if(n_done == n_slots) % end condition
        % deal with last patient of the day here
        % end locked
        break
    end

%     deltas(1:10,:) %%% FOR DEBUGGING %%%%
    %%% FIND THE NEXT ACTION
    curr_min = 50000;
    for i = 1:arrival_ind % through all patients that have arrived +
next arrival
        current_step = phases(i);

        if(current_step <= n_stages)
            if(t_enabled(current_step))
                if(active(i) == 1)
                    if(deltas(i,current_step) < curr_min)
                        curr_min = deltas(i,current_step);
                        next_ind = i;
                    end
                end
            end
        end
    end

    %%% ADVANCE THE SYSTEM TO NEXT TIME POINT
    if(curr_min ~= 0) % avoid excessive zero addition operations
        if(phases(next_ind)==1)
            % copy over the system time, to avoid excessive addition
            % avoiding numerical problems
            current_time = arrival_times(next_ind);
        else
            current_time = current_time + curr_min; % update the system
time
        end
    end
```

```
    deltas(next_ind,phases(next_ind)) = 0; % for debugging,
illustration
    result(next_ind,phases(next_ind)) = current_time;


    %%% UPDATE THE SYSTEM STATUS, based on current phase & type of
event
    if(phases(next_ind) == 1)  % an arrival //while busy
        active(arrival_ind) = 1; % activate this patient, probably
redundant now
        arrival_ind = arrival_ind + 1;
        n_in_sys = n_in_sys + 1;

        if(arrival_ind > n_slots)
            t_enabled(1) = 0; %turn off arrivals
            arrival_ind = n_slots;
        end
        %
        occurs(next_ind,1) = 0; %%% arrival has occurred
    elseif(phases(next_ind) == 2) %rooming is done
%         active(next_ind) = 2; % auto wait
%         techQueue = [techQueue next_ind]; % store to process below

        occurs(next_ind,2) = 0; %%% rooming has occurred

    elseif(phases(next_ind) == 3) % workup is done
        n_techs = n_techs + 1; %return tech

        pos = find(w_tech==next_ind);
        w_tech(pos) = -1; % reset that tech

        % repeat as with techs before; put them in VF queue
%         active(next_ind) = 2; % auto wait
%         vfQueue = [vfQueue next_ind]; % store to process below
        occurs(next_ind,3) = 0; %%% tech has occurred

    elseif(phases(next_ind) == 4) % vf is done
        n_vf = n_vf + 1;

        pos = find(w_vf==next_ind);
        w_vf(pos) = -1; % reset that vf

%         active(next_ind) = 2; % auto wait
%         docQueue = [docQueue next_ind]; % store to process below
        occurs(next_ind,4) = 0; %%% vf has occurred

    elseif(phases(next_ind) == 5) % Doc is done
        n_doc = n_doc + 1;

        pos = find(w_doc==next_ind);
```

```
        w_doc(pos) = -1; % reset that doc

        occurs(next_ind,5) = 0; %%% doc has occurred

%         n_in_sys = n_in_sys - 1; %patient leaves
%         active(next_ind) = 0;
%         n_done = n_done + 1;


    end

    %%% FIND WHAT QUEUE TO PUT THE PATIENT IN %%%
%     phases(next_ind) = phases(next_ind) + 1; % advance patient to
next phase
    previousOccurence = phases(next_ind);
    if(phases(next_ind)<=n_stages)
        whereTo = find(occurs(next_ind,:),1);

        if(isempty(whereTo)) %%% nowhere left to go
            n_in_sys = n_in_sys - 1; %patient leaves
            active(next_ind) = 0;
            phases(next_ind) = n_stages+1;
            n_done = n_done + 1;
        else %%% still stuff to do
             active(next_ind) = 2; % auto wait
             phases(next_ind) = whereTo;

             if(whereTo==3) %%% this can be less hard-coded later, if
needed
                 techQueue = [techQueue next_ind]; % store to process
below
             elseif(whereTo==4)
                 vfQueue = [vfQueue next_ind]; % store to process below
             elseif(whereTo==5)
                 docQueue = [docQueue next_ind]; % store to process
below
             end
        end

    end



    %%% UPDATE DELTAS
    start = find(active,1);
    for i = start:n_slots
        if(i ~= next_ind) %no progress on the NEXT step of the current
patient
            if(phases(i) <= n_stages) %if they arent done
                if(curr_min ~= 0) %avoid excessive zero subtraction
operations, prevent queued items from subtracting
```

```
                    %
                    if(active(i)==1) %if working, make progress
                        deltas(i,phases(i)) = deltas(i,phases(i)) -
curr_min;
                    elseif(active(i)==2) %if waiting, record delay
                        delays(i,phases(i)) = delays(i,phases(i)) +
curr_min;
                    end
                    %
                end
            end
        end
        %
    end

    %%% UPDATE THE QUEUE, for next step
    if(~isempty(techQueue))  %someone waiting?
        if(n_techs > 0) %tech is available

            pos = find(w_tech== -1,1); %find the available tech
            temp_i = techQueue(1);
            w_tech(pos) = temp_i; %front of the queue is with tech now
            active(temp_i) = 1; %reactivate this patient
            techQueue(1) = [];
            n_techs = n_techs - 1;
        end
    end

    if(~isempty(vfQueue))  %someone waiting?
        if(n_vf > 0) %VF is available

            pos = find(w_vf== -1,1); %find the available VF
            temp_i = vfQueue(1);
            w_vf(pos) = temp_i; %front of the queue is with VF now
            active(temp_i) = 1; %reactivate this patient
            vfQueue(1) = [];
            n_vf = n_vf - 1;
        end
    end

    if(~isempty(docQueue))  %someone waiting?
        if(n_doc > 0) %VF is available

            pos = find(w_doc== -1,1); %find the available VF
            temp_i = docQueue(1);
            w_doc(pos) = temp_i; %front of the queue is with VF now
            active(temp_i) = 1; %reactivate this patient
            docQueue(1) = [];
            n_doc = n_doc - 1;
        end
    end
%
```

```
%         end

    % I want in trace
    % 1 - time
    % 2 - patient
    % 3 - stage complete
    % 4 - n_techs idle
    % 5 - n_vf idle
    % 6 - n_doc idle
    trace(tick,1) = current_time;
    trace(tick,2) = next_ind;
    trace(tick,3) = previousOccurence;
    trace(tick,4) = n_techs;
    trace(tick,5) = n_vf;
    trace(tick,6) = n_doc;
    tick = tick + 1;

%     re-enable this and pass in a writer object for visualization
%     fig = figure(1);
%     imshow(deltas,[0 60])
%     truesize(fig,[600 60])
%     F = getframe(gcf);
%     writeVideo(writerObj,F)

%%% FOR DEBUGGING: THE SUMMARY %%%
%         happened = [next_ind previousOccurence]
end

trace = trace(1:tick-1,:); %%% correct for jump ahead
% toc
```

References

45 CFR 164.514 - Other requirements relating to uses and disclosures of protected health information. (2000, June 7, 2013). *Code of Federal Regulations.* Retrieved from https://www.law.cornell.edu/cfr/text/45/164.514

Adan, I., Bekkers, J., Dellaert, N., Vissers, J., & Yu, X. (2008). Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science, 12*(2), 129. doi:10.1007/s10729-008-9080-9

Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research, 258*(1), 3-34. doi:https://doi.org/10.1016/j.ejor.2016.06.064

Al-Mohy, A. H., & Higham, N. J. (2010). A New Scaling and Squaring Algorithm for the Matrix Exponential. *SIAM Journal on Matrix Analysis and Applications, 31*(3), 970-989. doi:10.1137/09074721x

Alfonso, E., Xie, X., Augusto, V., & Garraud, O. (2012). Modeling and simulation of blood collection systems. *Health Care Manag Sci, 15*(1), 63-78. doi:10.1007/s10729-011-9181-8

Ascher, H., & Feingold, H. (1984). *Repairable systems reliability : modeling, inference, misconceptions and their causes*. New York: M. Dekker.

Bertsekas, D. P. (2008). *Introduction to probability*. Belmont, Mass.: Athena Scientific.

Bertsimas, D. (1990). An Analytic Approach to a General Class of G/G/s Queueing Systems. *Operations Research, 38*(1), 139-155. doi:doi:10.1287/opre.38.1.139

Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (2006). *Queueing networks and Markov chains [electronic resource] : modeling and performance evaluation with computer science applications*. Hoboken, N.J.: Wiley-Interscience.

Brailsford, S. C. (2007). Tutorial: Advances and challenges in healthcare simulation modeling. *Proceedings of the 2007 Winter Simulation Conference, Vols 1-5*, 1415-1427.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci., 16*(3), 199-231. doi:10.1214/ss/1009213726

Brennan, P. F., Chiang, M. F., & Ohno-Machado, L. (2018). Biomedical informatics and data science: evolving fields with significant overlap. *Journal of the American Medical Informatics Association, 25*(1), 2-3. doi:10.1093/jamia/ocx146

Brill, S. (2015). *America's bitter pill : money, politics, backroom deals, and the fight to fix our broken healthcare system* (First edition. ed.). New York: Random House.

Cayirli, T., & Veral, E. (2003). Outpatient Scheduling in Health Care: A Review of the Literature. *Production and Operations Management, 12*(4), 519-549. doi:doi:10.1111/j.1937-5956.2003.tb00218.x

Choi, H., Kulkarni, V. G., & Trivedi, K. S. (1994). Markov regenerative stochastic Petri nets. *Performance Evaluation, 20*(1), 337-357. doi:http://dx.doi.org/10.1016/0166-5316(94)90021-3

CHOIR. (2018). Welcome to CHOIR: The Center for Healthcare Operations Improvement and Research at the University of Twente. Retrieved from https://www.utwente.nl/en/choir/

Christodoulou, G., & Taylor, G. J. (2001). Using a continuous time hidden Markov process, with covariates, to model bed occupancy of people aged over 65 years. *Health Care Management Science, 4*(1), 21-24. doi:10.1023/a:1009641430569

Ciardo, G., Blakemore, A., Chimento, P. F., Muppala, J. K., & Trivedi, K. S. (1993). Automated Generation and Analysis of Markov Reward Models Using Stochastic Reward Nets. In C. D. Meyer & R. J. Plemmons (Eds.), *Linear Algebra, Markov Chains, and Queueing Models* (pp. 145-191). New York, NY: Springer New York.

Ciardo, G., Muppala, J., & Trivedi, K. (1989, 11-13 Dec. 1989). *SPNP: stochastic Petri net package.* Paper presented at the Proceedings of the Third International Workshop on Petri Nets and Performance Models, PNPM89.

Ciardo, G., & Trivedi, K. S. (1993). A decomposition approach for stochastic reward net models. *Performance Evaluation, 18*(1), 37-59. doi:https://doi.org/10.1016/0166-5316(93)90026-Q

Çınlar, E., & Sollenberger, N. J. (2013). *Introduction to stochastic processes* (Dover edition. ed.). Mineola, New York: Dover Publications, Inc.

Cobelli, C., & Carson, E. R. (2008). *Introduction to modeling in physiology and medicine* (1st ed.). Amsterdam ; Boston: Academic Press.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. New York: M. Dekker.

Duke University Health System (DUHS). (2018). Duke Eye Center Information. Retrieved from https://www.dukehealth.org/locations/duke-eye-center

Durrett, R. (2012). *Essentials of stochastic processes* (2nd ed.). New York ; London: Springer.

Fackrell, M. (2008). Modelling healthcare systems with phase-type distributions. *Health Care Management Science, 12*(1), 11. doi:10.1007/s10729-008-9070-y

Fenton, N. E., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Boca Raton: Taylor & Francis.

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., . . . Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *J Public Health Med, 25*(4), 325-335.

Fricks, R., Veihl, M., Tseng, H., Trivedi, K., & Barr, R. (2018). Event Logging Performance Data (Publication no. doi/10.7910/DVN/HKWBP3). from Harvard Dataverse https://doi.org/10.7910/DVN/HKWBP3

Fricks, R. B., Bobbio, A., & Trivedi, K. S. (2016, 25-28 Jan. 2016). *Reliability models of chronic kidney disease.* Paper presented at the 2016 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. B., & Trivedi, K. S. (2016, 25-28 Jan. 2016). *Analysis methods for performance & availability in critical care medicine.* Paper presented at the 2016 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. B., & Trivedi, K. S. (2017, 23-26 Jan. 2017). *Automated life cycle processing for complex medical imaging devices.* Paper presented at the 2017 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. B., Tseng, H., Pajic, M., & Trivedi, K. S. (2017). Transient Performance & Availability Modeling in High Volume Outpatient Clinics. *2017 Annual Reliability and Maintainability Symposium (RAMS)*, 1-6.

Fricks, R. B., Tseng, H., Veihl, M., Trivedi, K. S., & Barr, R. C. (2018, 17-21 July 2018). *Robust Prediction of Treatment Times in Concurrent Patient Care.* Paper presented at the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Fricks, R. B., Tseng, H. H., Pajic, M., & Trivedi, K. S. (2017, 23-26 Jan. 2017). *Transient performance & availability modeling in high volume outpatient clinics.* Paper presented at the 2017 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. M., Puliafito, A., Mikl, #243, Telek, s., & Trivedi, K. S. (1998). Applications of non-Markovian stochastic Petri nets. *SIGMETRICS Perform. Eval. Rev., 26*(2), 15-27. doi:10.1145/288197.288204

Ghosh, R. (2012). *Scalable Stochastic Models for Cloud Services.* (Dissertation/Thesis). Retrieved from http://duke.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2AQNtIz0EUrE9D2gMFuqEM7TQDUkzTUNzMEbZliNQId1AmaffUxQzmpH1zwuwky8LggTVgLMTCl5gkzCMCuQlCA5gwRBt1goF9Bu5AUgkvykzMSQacfK4DuGcspVgA2CxWcc_JLUxRgGVOUwcLNNcTZQxdkX3wKMFMnp8aDTmAG8SE8YDcWzIsHVgXgBVKG8SAXG4kxsAA77KkSDAqJRgZGaaB7xYFtBxNT05TEpGRDgyRgwx7YNkozNzCSZDAk2XgpMvRIM3ABq3cjyICBDANrGjBBp8pCwlIOHLAAsQF8jQ

Ghosh, R., Longo, F., Frattini, F., Russo, S., & Trivedi, K. S. (2014). Scalable Analytics for IaaS Cloud Availability. *IEEE Transactions on Cloud Computing, 2*(1), 57-70. doi:10.1109/TCC.2014.2310737

Goodfellow, I. a. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Griffiths, J. D., Price-Lloyd, N., Smithies, M., & Williams, J. E. (2005). Modelling the Requirement for Supplementary Nurses in an Intensive Care Unit. *The Journal of the Operational Research Society, 56*(2), 126-133.

Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (2012). Retrieved from

https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

Gunal, M. M., & Pidd, M. (2010). Discrete event simulation for performance modelling in health care: a review of the literature. *J of Sim, 4*(1), 42-51.

Ha, J. F., & Longnecker, N. (2010). Doctor-Patient Communication: A Review. *The Ochsner Journal, 10*(1), 38-43.

Haas, P. J. (2002). *Stochastic Petri nets : modelling, stability, simulation*. New York: Springer.

Hastie, T. (2009). *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction Springer series in statistics*, (pp. xxii, 745 p.). Retrieved from http://getitatduke.library.duke.edu/?sid=sersol&SS_jc=TC0000145389&title=The%20Elements%20of%20Statistical%20Learning%3A%20Data%20Mining%2C%20Inference%2C%20and%20Prediction%2C%20Second%20Edition

Hirel, C., Tuffin, B., & Trivedi, K. S. (2000). *SPNP: Stochastic Petri Nets. Version 6.0*, Berlin, Heidelberg.

Horvath, G., & Telek, M. (2017). *BuTools 2: a Rich Toolbox for Markovian Performance Evaluation*. Paper presented at the proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools on 10th EAI International Conference on Performance Evaluation Methodologies and Tools, Taormina, Italy.

Hribar, M. R., Biermann, D., Read-Brown, S., Reznick, L., Lombardi, L., Parikh, M., . . . Chiang, M. F. (2016). Clinic Workflow Simulations using Secondary EHR Data. *AMIA Annual Symposium Proceedings, 2016*, 647-656.

Hribar, M. R., Read-Brown, S., Goldstein, I. H., Reznick, L. G., Lombardi, L., Parikh, M., . . . Chiang, M. F. (2018). Secondary use of electronic health record data for clinical workflow analysis. *Journal of the American Medical Informatics Association, 25*(1), 40-46. doi:10.1093/jamia/ocx098

Hribar, M. R., Read-Brown, S., Reznick, L., Lombardi, L., Parikh, M., Yackel, T. R., & Chiang, M. F. (2015). Secondary Use of EHR Timestamp data: Validation and Application for Workflow Optimization. *AMIA Annual Symposium Proceedings, 2015*, 1909-1917.

Hulshof, P. J., Boucherie, R. J., Essen, J. T., Hans, E. W., Hurink, J. L., Kortbeek, N., . . . Zonderland, M. E. (2011). ORchestra: an online reference database of OR/MS literature in health care. *Health Care Manag Sci, 14*(4), 383-384. doi:10.1007/s10729-011-9169-4

International Joint Commission on Allied Health Personnel in Ophthalmology (IJCAHPO). (2018). Certification & Recertification for Ophthalmic Professionals. Retrieved from http://www.jcahpo.org/certification-recertification/

Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Application of Discrete-Event Simulation in Health Care Clinics: A Survey. *The Journal of the Operational Research Society, 50*(2), 109-123. doi:10.2307/3010560

Keep Me Waiting: Medical Practice Wait Times and Patient Satisfaction. (2010). Retrieved from https://helpandtraining.pressganey.com/Documents_secure/Medical%20Practices/White%20Papers/Keep_Me_Waiting.pdf

Kendall, D. G. (1951). Some Problems in the Theory of Queues. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(2), 151-185.

Kennedy, J. E. (2018). Taking a Lean Approach in Health Care. Retrieved from http://dukeendowment.org/story/taking-a-lean-approach-in-health-care

Kleinrock, L. (1975). *Queueing systems*. New York,: Wiley.

Klenke, A. (2014). *Probability Theory [electronic resource] : A Comprehensive Course*. London: Springer London : Imprint: Springer.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. Cambridge, Mass.: MIT Press.

Kopach-Konrad, R., Lawley, M., Criswell, M., Hasan, I., Chakraborty, S., Pekny, J., & Doebbeling, B. N. (2007). Applying Systems Engineering Principles in Improving

Health Care Delivery. *J Gen Intern Med, 22*(Suppl 3), 431-437. doi:10.1007/s11606-007-0292-3

Kortbeek, N. (2012). *Quality-driven Efficiency in Healthcare*: University of Twente [Host].

Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd ed.). Boston: McGraw-Hill.

Lawal, A. K., Rotter, T., Kinsman, L., Sari, N., Harrison, L., Jeffery, C., . . . Flynn, R. (2014). Lean management in health care: definition, concepts, methodology and effects reported (systematic review protocol). *Systematic Reviews, 3*, 103-103. doi:10.1186/2046-4053-3-103

Leite, C. R. M., Martin, D. L., Sizilio, G. R. M. A., Santos, K. E. A. d., B. G. de Araujo, Valentim, R. A. d. M., . . . Guerreiro, A. M. G. (2010, Aug. 31 2010-Sept. 4 2010). *Modeling of medical care with stochastic Petri Nets.* Paper presented at the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology.

Liker, J. K. (2004). *The Toyota way : 14 management principles from the world's greatest manufacturer*. New York: McGraw-Hill.

Manolakis, D. G., Ingle, V. K., & Kogon, S. M. (2005). *Statistical and adaptive signal processing spectral estimation, signal modeling, adaptive filtering, and array processing Artech House signal processing library* (pp. 1 online resource (xviii, 796 p). Retrieved from http://getitatduke.library.duke.edu/?sid=sersol&SS_jc=TC0000251010&title=Statistical%20and%20adaptive%20signal%20processing%20%3A%20spectral%20estimation%2C%20signal%20modeling%2C%20adaptive%20filtering%2C%20and%20array%20processing

Marshall, A., Vasilakis, C., & El-Darzi, E. (2005). Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions. *Health Care Management Science, 8*(3), 213-220. doi:10.1007/s10729-005-2012-z

Masic, I., Miokovic, M., & Muhamedagic, B. (2008). Evidence Based Medicine – New Approaches and Challenges. *Acta Informatica Medica, 16*(4), 219-225. doi:10.5455/aim.2008.16.219-225

McManus, M. L., Long, M. C., Cooper, A., & Litvak, E. (2004). Queuing theory accurately models the need for critical care resources. *Anesthesiology, 100*(5), 1271-1276.

Michael, M., Schaffer, S. D., Egan, P. L., Little, B. B., & Pritchard, P. S. (2013). Improving wait times and patient satisfaction in primary care. *J Healthc Qual, 35*(2), 50-59; quiz 59-60. doi:10.1111/jhq.12004

Miller, R. G., Gong, G., & Muñoz, A. (1981). *Survival analysis*. New York: Wiley.

National Institutes of Health (NIH). (2015). Facts About Glaucoma. Retrieved from https://nei.nih.gov/health/glaucoma/glaucoma_facts

Nelson, W. (2003). *Recurrent events data analysis for product repairs, disease recurrences, and other applications*. Philadelphia: Society for Industrial and Applied Mathematics.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models : an algorithmic approach*. Baltimore: Johns Hopkins University Press.

OECD. (2017). *Health at a Glance 2017*.

Okamura, H., Dohi, T., & Trivedi, K. S. (2011). A refined EM algorithm for PH distributions. *Performance Evaluation, 68*(10), 938-954. doi:https://doi.org/10.1016/j.peva.2011.04.001

Pham, T. Q., Wang, J. J., Rochtchina, E., Maloof, A., & Mitchell, P. (2004). Systemic and ocular comorbidity of cataract surgical patients in a western Sydney public hospital. *Clin Exp Ophthalmol, 32*(4), 383-387. doi:10.1111/j.1442-9071.2004.00842.x

President's Council of Advisors on Science and Technology (PCAST). (2010). *Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward*. Retrieved from https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports

President's Council of Advisors on Science and Technology (PCAST). (2014). *Better Health Care and Lower Costs: Accelerating Improvement Through Systems Engineering*. Retrieved from https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports

Principles of Performance and Reliability Modeling and Evaluation [electronic resource] : Essays in Honor of Kishor Trivedi on his 70th Birthday. (2016). Cham: Springer International Publishing : Imprint: Springer.

Reid, P. P., Compton, W. D., Grossman, J. H., Fanjiang, G., National Academy of Engineering., Institute of Medicine (U.S.), & National Academies Press (U.S.). (2005). *Building a better delivery system : a new engineering/health care partnership*. Washington, D.C.: National Academies Press.

Rigdon, S. E., & Basu, A. P. (2000). *Statistical methods for the reliability of repairable systems*. New York: Wiley.

Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Boston: Brooks/Cole, Cengage Learning.

Ross, S. M. (1997). *Introduction to probability models* (6th ed.). San Diego, CA: Academic Press.

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association, 14*(1), 1-9. doi:10.1197/jamia.M2273

Sahner, R., Trivedi, K. S., & Puliafito, A. (1996). *Performance and reliability analysis of computer systems : an example-based approach using the SHARPE software package*. Boston: Kluwer Academic Publishers.

Silva, M. (2012). 50 years after the PhD thesis of Carl Adam Petri: A perspective. *IFAC Proceedings Volumes, 45*(29), 13-20. doi:https://doi.org/10.3182/20121003-3-MX-4033.00006

Sozu, T. (2015). *Sample Size Determination in Clinical Trials with Multiple Endpoints [electronic resource]*. Cham: Springer International Publishing : Imprint: Springer.

Takagi, H., Kanai, Y., & Misue, K. (2017). Queueing network model for obstetric patient flow in a hospital. *Health Care Management Science, 20*(3), 433-451. doi:10.1007/s10729-016-9363-5

Takagi, H., Misue, K., & Kanai, Y. (2014, 23-25 April 2014). *Queuing Network Model and Visualization for the Patient Flow in the Obstetric Unit of the University of Tsukuba Hospital.* Paper presented at the 2014 Annual SRII Global Conference.

Thummler, A., Buchholz, P., & Telek, M. (2005, 28 June-1 July 2005). *A novel approach for fitting probability distributions to real trace data with the EM algorithm.* Paper presented at the 2005 International Conference on Dependable Systems and Networks (DSN'05).

Trivedi, K. S. (2002). *Probability and statistics with reliability, queuing, and computer science applications* (2nd ed.). New York: Wiley.

Trivedi, K. S., & Bobbio, A. (2017). *Reliability and availability engineering : modeling, analysis, and applications*. New York, NY, USA: Cambridge University Press.

Trivedi, K. S., & Sahner, R. (2009). SHARPE at the age of twenty two. *SIGMETRICS Perform. Eval. Rev., 36*(4), 52-57. doi:10.1145/1530873.1530884

Vakili, S., Pandit, R., Singman, E. L., Appelbaum, J., & Boland, M. V. (2015). A comparison of commercial and custom-made electronic tracking systems to measure patient flow through an ambulatory clinic. *International Journal of Health Geographics, 14*(1), 32. doi:10.1186/s12942-015-0023-7

van Dijk, N. M., & Kortbeek, N. (2009). Erlang loss bounds for OT–ICU systems. *Queueing Systems, 63*(1), 253. doi:10.1007/s11134-009-9149-2

Welch, J. D., & Bailey, N. T. (1952). Appointment systems in hospital outpatient departments. *Lancet, 1*(6718), 1105-1108.

Willis, M. C. (2008). *Medical terminology : a programmed learning approach to the language of health care* (2nd ed.). Philadelphia: Lippincott Williams & Wilkins.

Womack, J. P., & Jones, D. T. (2003). *Lean thinking : banish waste and create wealth in your corporation* (1st Free Press ed.). New York: Free Press.

Biography

Rafael Fricks took interest in computation and medical sciences from an early age, leading to a lifelong pursuit for creative innovation in health care. He received a Bachelor of Science degree in Biomedical Engineering at the University of Texas in Austin in 2013, and a Master of Science degree in Biomedical Engineering from Duke University in 2017. Since 2012 he has been a member of the Institute of Electrical and Electronics Engineers (IEEE), a student member of the Biomedical Engineering Society since 2015, and a member of the American Medical Informatics Association (AMIA) since 2017. As a graduate student, he was a NSF Graduate Research Fellow and Dean's Graduate Fellow at Duke University researching applications of stochastic models and machine learning in health care, with a focus on model validation and clinic flow optimization. Since 2015 he has modeled clinic flow at Duke Eye Centers, a multi-specialty outpatient facility seeing as many as 350 patients per day. He is a recipient of the 2016 Thomas L. Fagan Award for best student paper by IEEE Reliability and Maintainability Symposium. From 2015 to 2017 he applied research concepts in modeling diagnostic imaging device dependability at Siemens Healthineers. He intends to continue researching predictive algorithms in medical imaging in a joint project with the Department of Veteran's Affairs and the Department of Radiology at Duke University.

The following items were published his graduate studies:

Fricks, R., Veihl, M., Tseng, H., Trivedi, K., & Barr, R. (2018). Event Logging Performance Data (Publication no. doi/10.7910/DVN/HKWBP3). from Harvard Dataverse https://doi.org/10.7910/DVN/HKWBP3

Fricks, R. B., Bobbio, A., & Trivedi, K. S. (2016, 25-28 Jan. 2016). *Reliability models of chronic kidney disease.* Paper presented at the 2016 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. B., & Trivedi, K. S. (2016, 25-28 Jan. 2016). *Analysis methods for performance & availability in critical care medicine.* Paper presented at the 2016 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. B., & Trivedi, K. S. (2017, 23-26 Jan. 2017). *Automated life cycle processing for complex medical imaging devices.* Paper presented at the 2017 Annual Reliability and Maintainability Symposium (RAMS).

Fricks, R. B., Tseng, H., Pajic, M., & Trivedi, K. S. (2017). Transient Performance & Availability Modeling in High Volume Outpatient Clinics. *2017 Annual Reliability and Maintainability Symposium (RAMS)*, 1-6.

Fricks, R. B., Tseng, H., Veihl, M., Trivedi, K. S., & Barr, R. C. (2018, 17-21 July 2018). *Robust Prediction of Treatment Times in Concurrent Patient Care.* Paper presented at the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Fricks, R. B., Tseng, H. H., Pajic, M., & Trivedi, K. S. (2017, 23-26 Jan. 2017). *Transient performance & availability modeling in high volume outpatient clinics.* Paper presented at the 2017 Annual Reliability and Maintainability Symposium (RAMS).