ADDRESSING CHALLENGES IN PUBLIC SERVICE OPERATIONS MANAGEMENT: DATA-DRIVEN SOLUTIONS AND STRATEGIES

Yanhan (Savannah) Tang

A Dissertation Submitted to the Tepper School of Business, Carnegie Mellon University
In Partial Fulfilment of the Requirements for the
Degree of Doctor of Philosophy in Operations Management

Dissertation Committee:
Alan Scheller-Wolf (Chair)
Zhaohui (Zoey) Jiang
Andrew Li
Sridhar Tayur
Alan Montgomery
March 2024

© 2024 Yanhan (Savannah) Tang
All rights reserved

Abstract

This dissertation examines data-driven decision-making in crucial areas of public service operations management, with a specific focus on liver allocation and child welfare operations.

Chapter 1 provides an overview of the research background and describes the common challenges in public service resource allocation as well as context-specific operational complexities.

Chapter 2 introduces a decision support model for split liver transplantation (SLT) to enhance efficiency and fairness in liver allocation. Through a multi-queue fluid system model, optimal matching procedures are identified, demonstrating the potential benefits of increased SLT utilization.

Chapter 3 explores learning-informed algorithms for SLT resource allocation, utilizing a multi-armed bandit (MAB) model to balance exploration and exploitation in surgical team selection. Novel algorithms, L-UCB and FL-UCB, are developed and shown to exhibit superior performance in allocating organs while incorporating experience-based learning and fairness concerns.

Chapter 4 studies the impact of workload on the screening of child maltreatment reports, highlighting the need for load-aware risk protocols in human-AI collaborations.

Chapter 5 concludes the dissertation by outlining future research avenues and potential operational improvements in public service.

This thesis is dedicated to my family.

Acknowledgement

Throughout my PhD journey, I had the privilege of sharing experiences and building friendships with many extraordinary people. I am fortunate to have received support and advice along the way. I am grateful to those who helped me reach this point.

I would like to express my gratitude to Prof. Alan Scheller-Wolf, my advisor and committee chair. I could not have achieved several Ph.D. milestones without your responsiveness, time, and support. I am indebted to you for introducing me to applied operations research as an undergraduate student and helping me choose the academic path with confidence and excitement. You have also trained me to become an independent and resilient researcher, which empowers me to embrace challenges as opportunities for personal and professional growth.

I am thankful to Prof. Sridhar Tayur, my advisor and committee member, for broadening my horizons in research and entrepreneurship. You have inspired me to think creatively and seek solutions beyond the conventional. Thank you for guiding me toward compelling research avenues, resulting in the development of two research papers and two chapters for this dissertation. I am also immensely grateful for your advice and support during my toughest times. You have taught me not only research, presentation, and interview skills but also how to approach problems efficiently and holistically.

I am profoundly grateful to Prof. Alan Montgomery, who is not only my advisor and outside reader of this dissertation but also the head of Tepper's Ph.D. program. As an

advisor, you introduced me to the exciting world of applied machine learning research, helped me navigate the large datasets you provided, and gave me tremendous support. As the Ph.D. head of Tepper, you have dedicated yourself to improving the doctoral program and going above and beyond to help every student in need. I can not thank you enough for having students' best interests at heart and empowering them to pursue their passions. We are all extremely fortunate to have you in charge.

I would also like to thank Prof. Andrew Li, my coauthor and committee member. It is always a delight to chat with you, and your expertise is essential for bringing the third chapter to fruition. I greatly appreciate your valuable insight and unwavering support throughout the years. I am thankful to Prof. Zoey Jiang, my committee member and coauthor of the fourth chapter. Thank you so much for working closely with me on my first empirical research project and guiding me through extensive empirical OM and human-AI literature. Your thoughtful mentorship and persistent support have significantly accelerated my learning process and are fundamental to the impressive progress we have made so far.

I am also thankful to my wonderful coauthors and collaborators, Dr. Emily R. Perito (UCSF), Dr. John P. Roberts (UCSF), Dr. Lindsey Lacey (Allegheny County DHS) and Dr. Justine Galbraith (Allegheny County DHS). Your domain expertise and valuable feedback are crucial to the completion and success of our collaborative projects. I greatly appreciate your time and help!

I express my heartfelt gratitude to Lawrence Rapp and Laila Lee. Thank you so much for taking such good care of Tepper PhD students and making life much easier. I greatly appreciate your kindness and dedication.

I would also like to thank the faculty members with whom I had the opportunity to interact. Special thanks to Prof. Fei Fang and Prof. Alexandre Jacquillat for supporting me during the most challenging times. I am also grateful to Prof. Ann Melissa Campbell, Prof. Fatma Kılınç-Karzan, Prof. Mustafa Akan, Prof. Soo-Haeng Cho, Prof. Michael Hamilton, Prof. John Hooker, Prof. Tae Wan Kim, Prof. R Ravi, Prof. Kannan Srinivasan, and Prof. Bryan Wilder for their guidance and wisdom.

I am fortunate to have friends and fellow students by my side, and together, we share happy memories. Many thanks to my friends and colleagues: Neda Mirzaeian, Musa Çeldir, Neha, Nilsu Uzunlar, Tian Wang, Alex Lim, H. Satyam Verma, Zooey Meznarich, Shubham Akshat, Mehmet Aydemir, Franco Berbeglia, Siddharth Singh, Violet Chen, Yuyan Wang, Sagnik Das, Özgün Elçi, Kyra Gan, Su Jia, Anthony Karahalios, Melda Korkut, Tom Krumpolc, Thomas Lavastida, Lin An, Siyue Liu, Macarena Navarro, Vrishabh Patil, Daniel de Roux, Ryan Shi, Ziye Tang, Sebastian Vasquez, Jody Zhu, Rudy Zhu, Mik Zlatin, Serim Hwang, Samuel Levy, Julie Wang, Behnam Mohammadi, Flora Feng, Yuan Yuan, Duyu Chen, Zhaoqi Cheng, Jason Gates, Jaepil Lee, Martin Michelini, Shuoqi Sun, Qiaochu Wang, Lavender Yang, Titing Cui, Wenjie Hu, Yuxin Gong, Lucy Wang, Jaeyoung Kim, Jade Xiao, Jeremy Watts, Xiaoquan Gao, Qinyi Chen, Jingwei Zhang, Verna Tian, Yi Zhou.

I would like to extend special thanks to Violet Chen and Neda Mirzaeian for their invaluable help and support in my job search. My sincere appreciation goes to all current and past officers of the CMU INFORMS Chapter. Thank you very much for your contributions to our vibrant community, which gives me a strong sense of belonging.

Lastly, I extend my deepest gratitude to my family. I am indebted to Xiaomei and Qu for their unconditional love. To my partner, Tianming, thank you for standing by my side through thick and thin. I treasure the countless moments of joy and laughter we've shared together. Special thanks to my cherished feline companions, Kirby and Lucky, whose comforting presence brought warmth to even the coldest days.

List of Coathors

Chapter 2

Alan Scheller-Wolf, Carnegie Mellon University

Sridhar Tayur, Carnegie Mellon University

Emily R. Perito, University of California, San Francisco

John P. Roberts, University of California, San Francisco

Chapter 3

Andrew Li, Carnegie Mellon University

Alan Scheller-Wolf, Carnegie Mellon University

Sridhar Tayur, Carnegie Mellon University

Chapter 4

Zhaohui (Zoey) Jiang, Carnegie Mellon University

Alan Scheller-Wolf, Carnegie Mellon University

Lindsey Lacey, Allegheny County Department of Human Services

Justine Galbraith, Allegheny County Department of Human Services

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Data-driven decision-making in public service operations management has emerged as a vital approach that leverages data analysis and insights to inform strategic, tactical, and operational decisions within the public sector. By harnessing data from various sources such as government databases, surveys, and administrative records, public service organizations can gain valuable insights into the needs, service delivery performance, and resource allocation.

Several common challenges persist across various areas within public services. One significant challenge is the efficient allocation of limited resources amidst demand and budget constraints. Whether allocating organs for transplant procedures or distributing social welfare resources, public service organizations often face the dilemma of balancing competing needs and priorities within capacity and budget. Additionally, ensuring fairness and equity in resource allocation presents a constant challenge. Whether ensuring equitable access to healthcare services or distributing social assistance programs, public service organizations must navigate complex socio-economic factors to ensure fair and unbiased resource allocation.

Another common challenge in public services is improving service delivery processes to meet evolving expectations and regulatory requirements. This includes challenges such as reducing wait times for critical services (e.g., deceased-donor organs for transplants), improving the quality of care in healthcare facilities, or enhancing responsiveness and accuracy in child welfare services. Public service organizations must

continuously adapt their operations to address technological advancements, changing demographics, and legislative mandates while maintaining service quality and efficiency. Overall, addressing these challenges requires innovative approaches, strategic planning, and collaboration across multiple stakeholders to ensure effective and sustainable public service delivery.

This dissertation examines two critical areas within public services: liver allocation and child welfare operations. We explore strategies aimed at the efficient and equitable allocation of scarce resources, including deceased-donor livers and child welfare resources. Additionally, we delve into the dynamics of human and technology collaboration geared towards enhanced decision-making processes.

Chapter 2 studies a decision support model for split liver transplantation (SLT). SLT is a procedure that can save two lives using one liver, increasing the total benefit derived from the limited number of donated livers available. SLT may also improve equity, by giving transplant candidates who are physically smaller (including children) increased access to liver transplants. However, SLT is rarely used in the US. To help quantify the benefits of increased SLT utilization and provide decision support, we introduce a deceased-donor liver allocation model with both efficiency and fairness objectives. We formulate our model as a multi-queue fluid system, incorporating the specifics of donor-recipient size matching and patients' dynamically changing health conditions. Leveraging a novel decomposition result, we find the optimal matching procedure for the overloaded liver allocation system, enabling us to benchmark the performance of different allocation policies against the theoretical optimal. Numerical results, utilizing data from the Organ Procurement and Transplantation Network,

show that increased utilization of SLT can significantly increase total quality-adjusted life years, reduce patient deaths, and improve fairness among different patient groups.

Chapter 3 discusses medical learning in SLT and proposes learning-informed algorithms for resource allocation. Proficiency in many sophisticated tasks is attained through experience-based learning, in other words, learning by doing. For example, transplant centers' surgical teams need to practice difficult surgeries to master the skills required. Meanwhile, this experience-based learning may affect other stakeholders, such as patients eligible for transplant surgeries, and require resources, including scarce organs. To ensure that patients have excellent outcomes while expanding the base of qualified surgeons, the organ allocation authority needs to quickly identify and develop medical teams with high aptitudes. This entails striking a balance between exploring surgical combinations with initially unknown full potentials and exploiting existing knowledge based on observed outcomes. We formulate this problem as a multi-armed bandit (MAB) model, in which parametric learning curves are embedded in the reward functions to capture endogenous, experience-based learning. In addition, our model includes provisions ensuring that the choices of arms are subject to fairness constraints to guarantee equity. To solve our MAB problem, we develop the L-UCB and FL-UCB algorithms, variants of the upper confidence bound (UCB) algorithm that we prove attain the optimal O(log t) regret on problems enhanced with experience-based learning and fairness concerns. We demonstrate our model and algorithms on the SLT allocation problem, showing that our algorithms have superior numerical performance—arriving at better allocations faster—compared to standard bandit algorithms in a setting where experience-based learning and fairness concerns exist. From a methodological point of view, our proposed MAB model and algorithms

are generic and have broad application prospects.

Chapter 4 studies the workload effect and human-artificial intelligence (AI) teaming in the screening of child maltreatment reports. Child welfare organizations regularly receive a significant number of calls alleging child neglect or abuse. Due to limited resources available for investigations and services, it is crucial to accurately assess and screen these allegations before further investigation or intervention to maintain a sustainable workload. Furthermore, investigations initiated based on unsubstantiated allegations can lead to harmful consequences for the family involved. To aid these essential screening decisions and enhance overall efficiency, a Predictive Risk Model (PRM), essentially an AI tool, has been deployed by our research partner. However, the PRM is load-agnostic and cannot adapt to fluctuating workloads. We empirically investigate the impact of system load on the screening decisions made via the human-PRM collaboration. Our results indicate that the probability of screening-in allegations is inversely correlated with the system load. Moreover, we find that human workers appear to informally incorporate workload information in their screening decisions, tending to deviate more often from the PRM tool recommendation when the system load is either very high or low. We discuss strategies to enhance the collaboration between human workers and the PRM by adopting load-aware risk protocols. More broadly, our work contributes to the discussion on human-AI teaming in high-stakes decision-making.

Chapter 5 concludes the dissertation and discusses avenues for future research. Three primary directions are outlined: exploring effective operational strategy designs incorporating human and organizational learning, studying operational improvements

in child welfare organizations, and investigating sequential decision-making under uncertainty in the nonasymptotic regime. These promising research paths, informed by this dissertation's findings, offer valuable opportunities to enhance decision-making and enable operational improvements in public services.

Chapter 2

Split Liver Transplantation: An Analytical Decision Support Model

2.1 Introduction

Liver transplantation is the only effective treatment for patients with end-stage liver diseases (ESLD). In this paper we focus on the matching of deceased-donor livers and potential recipients, as more than 95% of US liver transplant surgeries use a deceased donor liver; the remaining 5% transplant living-donor livers, using a different matching procedure. In the US, the number of patients waiting for a liver transplant far exceeds the number of available donated livers: A total of 9528 liver transplants were performed in 2022, while 13179 ESLD candidates were added to the waiting lists. Moreover, 1043 candidates died while waiting in 2022. As of September 2023, there are 10081 candidates on the US liver transplant waitlists, and 615 waitlisted candidates have died before receiving a transplant. As liver shortages persist despite countless efforts to bridge this gap between supply and need, it is crucial to allocate the livers we have as efficiently and fairly as possible (OPTN & UNOS, 2022a).

Split liver transplantation (SLT) is a procedure that can save two lives using one donated liver: It is widely accepted that splitting qualified livers for suitable patients is appropriate, as SLT yields outcomes comparable to the traditional whole liver transplantation (WLT) in transplant centers (TCs) with adequate experience (OPTN & UNOS, 2016). There are two splitting methods: the *adult-child* split and the *adult-adult* split, according to to (OPTN & UNOS, 2016). SLT is possible because liver

cells have very high regenerative capability: When a partial liver is used for transplantation, it will grow to the proper size within months if the transplant is successful. Moreover, as transplant candidates of smaller sizes usually have longer expected waiting times, SLT could potentially increase the supply of smaller liver allografts and therefore provide more equitable access for smaller candidates. In fact, 92% of SLTs are performed for children in the US. Usually, fairness and efficiency are framed as trade-offs; SLT offers a unique opportunity to simultaneously improve both.

Sadly, despite SLT's potential to ameliorate the acute shortage of donated livers and improve equity, it is rarely used in the US. More than 10% of all deceased-donor livers in the US are of sufficiently high quality to be deemed medically splittable according to OPTN-specified criteria, yet less than 1.5% of livers are split (OPTN & UNOS, 2016). Barriers to increased SLT utilization include logistical difficulties, surgical expertise, geography, and the complexities of donor-recipient matching.

The Organ Procurement and Transplantation Network (OPTN) oversees all organ procurement and allocation in the US. When a deceased donor liver becomes available, the organ is sequentially offered to appropriate ESLD patients on the waiting list according to a ranked list; patients may accept or reject the organ offer. These ranked lists incorporate information about the potential recipients' sizes, geographical locations, blood types, and health conditions measured by their Model of End-stage Liver Disease (MELD) scores for adults, or Pediatric End-stage Liver Disease (PELD) scores for children. MELD scores take integer values from 6 to 40; some critically sick children or adults may be listed as status 1A, and children who meet specific criteria may be listed as 1B instead of getting an integer score (Kamath & Kim, 2007). The

MELD and PELD scores are frequently updated and measure medical urgency based on lab tests: The higher the MELD/PELD score a patient has, the lower the expected survival rate of this patient in the next 90 days without a transplant. The US allocation policy prioritizes patients with the highest MELD/PELD scores, when other factors such as sizes, geographical locations, and blood types are compatible (OPTN & UNOS, 2022b). By default, livers are allocated as a whole to the sickest patients; only in exceptional cases where a highly ranked child or small adult requires a partial liver will a liver be split and used for SLTs, if medically safe.

In contrast to the policy in the United States, the UK adopts an "all-split" liver allocation policy, where all splittable-livers are split, except in exceptional circumstances. Procedure-wise, livers are evaluated and deemed to be split (or not split), and then recipients are chosen to get transplanted (NHS, 2022). By decoupling the liver-splitting decision and recipient choices, many logistical hurdles are alleviated. Our analysis indicates that based on current US data, splitting all splittable livers and then allocating the whole/partial livers according to our model performs nearly optimally in terms of maximizing total quality-adjusted life years (QALY) and minimizing the total number of patient deaths (TNPD). Moreover, under the current "sickest first" allocation priority, splitting all livers alone consistently improves the system performance under various transplantation objectives. These findings suggest that an adapted strategy analogous to the UK's "all-split" policy could work well in the current US liver allocation system.

We analytically model the deceased-donor liver allocation and matching problem incorporating the use of SLT for the first time; one of our main contributions is a

novel SLT decision support module/subroutine. We use a fluid model as a first-order approximation of the dynamic liver allocation problem. Our fluid model provides us with analytical tractability while yielding a faithful approximation, owing to the overloaded nature of liver transplant waitlists. We divide patients into groups with different static features (physical sizes) and dynamically changing health conditions (MELD/PELD scores). Our primary patient welfare objectives are maximizing patients' total QALY, and minimizing the number of patient deaths while waiting for a transplant (NPDWT).

As size matching and equity is an important concern related to who has access to donated livers—and as SLT involves physically splitting larger liver grafts into smaller parts—we explicitly quantify the effects of SLT on transplant access among patient size groups. We do so by incorporating explicit equity constraints among different groups' average probabilities of getting transplants (PGT). To properly address the equity concerns, we need to strike a delicate balance: We want to improve access for smaller patients—often children—but do not want excess utilization of SLT to lead to some patient groups facing unfairly low access to liver transplants - for example, overweight or critically sick candidates typically not qualified for SLT. Our fluid model/analysis helps provide insights regarding dynamic liver allocation in realistic settings, incorporating fairness concepts from philosophy, some of which have been applied to the SLT setting (T. W. Kim et al., 2021).

To summarize, we provide the first organ allocation model incorporating SLT. We show that incorporating our SLT decision support module can improve both utility and equity, compared to the current OPTN policy as well as other benchmark policies.

The policy implications and methodological improvement are academic contributions toward improving practice. We hope our work may inspire a more detailed clinical analysis of the barriers to, and benefits of, increased SLT utilization in the US.

Methodologically, we advance the accuracy and reduce the computation complexity of the organ allocation matching problem by finding an explicit solution to our fluid model through a novel fluid limit decomposition. We also prove that our solution is globally optimal in the interior of the fluid queue state space. To our knowledge, previous work using fluid queues to model organ allocation has relied on heuristic-based solutions that involve high complexity and large search spaces. Our decomposition method is not only efficient and exact in scenarios when queues are nonempty, but it also allows us to perform sensitivity analysis and shed insight into how different factors affect the optimal organ allocation policies and the objective values. For example, incorporating waitlisted candidates' endogenous, strategic accept/reject decisions in an optimal organ allocation problem has been considered intractable; our fluid decomposition method helps dissect the convoluted problem into simpler modules that are much more solvable. The methodological contribution of the proposed fluid decomposition goes beyond organ transplantation—it solves *transient* optimal dynamic control in multi-class queues formulated with fluid approximation before the *steady state* is reached in the interior of the fluid queue state space. Our fluid decomposition method sheds insights for other queueing control problems, for example, optimal scheduling of proactive service with dynamic patient health conditions (Hu et al., 2021), resource sharing among multiple queueing classes with customer abandonment (Larranaga et al., 2013), and hospital patient flow management (Dong & Perry, 2020).

The rest of the chapter is organized as follows. In section 2.2 we provide a literature review on related work from both the transplantation community and the management science/operations research (MS/OR) community; in section 2.3 we present our fluid model formulation, fairness metric and discuss the trade-offs between efficiency and fairness. In section 2.4, we introduce our fluid limit decomposition result and present structural properties of the optimal fluid control policy as well as new insights. In section 2.6 we present major numerical results and managerial insights for our problem. Finally, in section 4.6, we conclude with some discussions of the contributions of this chapter and future directions.

2.2 Literature Review

There are four streams of literature that are most relevant to our work: a) MS/OR work on organ transplantation; b) SLT papers in the transplantation community; c) fluid model literature; and d) literature on fairness.

OR/MS Literature on Organ Transplantation: OR/MS researchers have studied organ allocation over the past two decades. Some researchers use simulation models to evaluate and compare the effect of various organ allocation policies on the system's performance (e.g. Zenios et al., 2000b); the majority of the papers on the optimal allocation of organs—most commonly kidneys—focus on maximizing welfare across the entire system. More specifically, researchers typically solve the optimal matching problem that pairs organs and patients to maximize utilitarian objectives, e.g. total QALY, average 1/3/5-year patient survival probability, or quality of the prospective matches. For example, Su and Zenios, 2006 studied the impact of information asymmetry and mechanism design in a system where patients indicate their

preference for kidney offers upon joining the waiting list. Ata et al., 2021 used a fluid model to find the ranking policy that optimizes the efficiency-fairness tradeoff among all policies that take patients' strategic choices into account, in equilibrium. Kidney allocation is in some ways more straightforward than liver allocation, as in kidney transplantation, priority in getting a kidney offer is usually given to patients who have waited for the longest on the list within the same geographic region, similar to a first-come-first-serve (FCFS) rule. This is reasonable as if a kidney transplant is unavailable, patients can often survive on dialysis for several years.

However, as there is no therapy comparable to dialysis for ESLD patients, these patients' conditions may be quite volatile. Thus, the ranking of candidates in liver allocation is more dependent on patients' dynamically changing health conditions; while a lot of work has been done on kidney allocation, fewer papers have addressed the typically more complex allocation and decision-making for livers. Sandıkçı et al., 2008 formulated a Markov decision process (MDP) model to solve the optimal decisions for one patient in a stochastic environment, and to determine *the price of privacy*, i.e. the life days lost due to lack of information. Closer to our work, Akan et al., 2012 analytically modeled the whole liver transplantation (WLT) allocation system using a fluid model with utilitarian objectives (e.g. QALY and NPDWT) incorporating dynamically-changing MELD/PELD scores; however, they did not consider either SLT or fairness. Methodologically, Akan et al., 2012 relies on solving complex dual controls for the interior of the state space and heuristic solutions for the dynamic indexes and the boundary case. In contrast, we provide an explicit solution to the extended fluid optimization problem through our novel fluid limit decomposition result. We also prove the optimality of our proposed decomposition method in the interior

of the fluid queue state space. We show that the optimal policy greedily optimizes dynamic indexes that can be written in explicit form without solving the dual control in the interior. This enables us to accurately quantify the impact of, and offer insights for, the incorporation of SLT.

Table 2.1 summarizes the MS/OR literature applying fluid models to organ transplantation.

Paper/Feature	Organ	SLT	Explicit Solution	Structural Properties	Fairness	Patient Choice
Zenios et al., 2000b	Kidney			✓	✓	
Su and Zenios, 2006	Kidney			✓	✓	✓
Akan et al., 2012	Liver			✓		
Ata et al., 2021	Kidney			✓	✓	✓
Our Chapter	Liver	✓	✓	✓	✓	

Table 2.1: Comparison to relevant literature that used fluid models to model organ allocation policies. Although we do not explicitly consider patient choices in our main model, we discuss ways to incorporate them as an extension in Subsections 2.5.3 and A.2.3.

SLT Literature: SLT papers in the transplantation community are mostly comprised of retrospective reviews of TCs that have performed SLT, sharing their SLT experiences—both outcome statistics and medical techniques—in major transplantation journals (Emre & Umman, 2011). The rest include ethics discussions (Vulchev et al., 2004), statistical analysis using open data (Perito et al., 2019), clinical and medical research on SLT techniques and postoperative effects, and discussions of SLT policy making (OPTN & UNOS, 2016). Recent studies show that the outcomes of SLT can be as good as WLT in major TCs with sufficient experience in SLT (Perito et al., 2019).

Fluid models: The fluid approximation has been widely accepted as a standard method to model overloaded queueing systems. Fluid models have been used in modeling organ allocation, hospital patient flow, proactive service scheduling, and many other healthcare operations problems (Akan et al., 2012; Hu et al., 2021; Shi et al., 2016; Zenios et al., 2000b). By properly scaling time and space, the fluid limit gives a first-order approximation of a complex queueing system, with the goal of stripping away details to reveal essential features. The validity of fluid approximation is based on the *functional law of large numbers* (FLLN): The FLLN states that the fluid limits characterize the first-order dynamics of queueing systems under mild regularity conditions (Whitt, 2006). Fluid models have been shown to be appropriate stylized models in the context of organ transplantation (Akan et al., 2012; Zenios et al., 2000b).

Although the optimal control of fluid models in steady state has been well-studied, the exact optimal control in the transient state has not yet been solved. Hu et al., 2021 derived the optimal control for a two-queue fluid model with customer transitions and performed detailed transient analysis. Our work studies explicit allocation policies for the transient fluid-model optimal control problem with a finite number of server and customer classes and proves its optimality in the interior of the state space.

Fairness: Equity of access in organ allocation has received increasing attention within the transplantation and MS/OR community. For instance, OPTN introduced the acuity circles policy in 2019, which is designed for a more geographically equitable allocation of livers (Mogul et al., 2020). Similarly, Bertsimas et al., 2020 introduced a more accurate priority scoring system that allows for more boundaryless, broader, and smoother liver allocation. Specifically to SLT, T. W. Kim et al., 2021 addressed

the question of when splitting a large, medically-splittable liver is ethically desirable and why a sophisticated, dynamic allocation policy is crucial to the ethics of SLT.

There are several alternative definitions of fairness in organ transplantation, and thus far, there is no single standard accepted by both practitioners and academics (Bertsimas et al., 2013; Committee et al., 2009). In this chapter, we focus on a type of fairness analogous to the fairness concept proposed by Rawls, 1999; specifically, we enforce a lower bound on the likelihood of transplant for each group of patients. We use the *price of fairness* (POF) concept to understand the efficiency-fairness trade-off (Bertsimas et al., 2011b).

2.3 Model Formulation

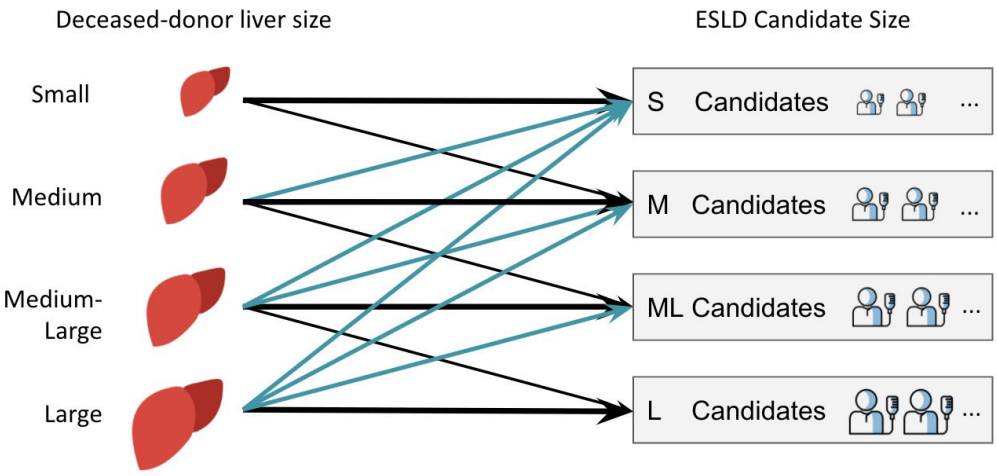


Figure 2.1: Thin black edges indicate a valid whole liver transplantation (WLT) liver-candidate size match, while teal edges indicate a plausible SLT liver-candidate size match. Thick black edges indicate a plausible size match for both WLT and SLT.

This section analytically models the deceased-donor liver allocation system (WLT and SLT) incorporating both efficiency and fairness concerns. We formulate a multi-queue fluid system with abandonment, including the specifics of donor-recipient size

matching and dynamically changing MELD or PELD scores. Our primary metrics are maximizing QALY and minimizing NPDWT, potentially subject to equity constraints that specify the minimal probabilities of getting transplants (PGT) by patient class. Our formulation and results directly extend to other well-known liver allocation objectives such as minimizing the total number of patient deaths (TNPD) and minimizing the number of patient deaths after transplant (NPDAT).

The fluid model and patient grouping in this section focuses exclusively on the size-matching and dynamic health conditions. We consider size as a static attribute in the primary model, grouping candidates into four patient classes: small (S), medium (M), medium-large (ML), and large (L). Patient health conditions are captured by patients' dynamic MELD or PELD scores. In this chapter, all boldfaced lower-case letters denote vectors; boldface upper case are matrices, and scalars are in regular/non-bold typeface. We consider a continuous time horizon $\mathcal{T} := [0, T]$. Patients of class (i, j), with sizes $i \in \mathcal{I} := \{S, M, ML, L\}$ (for convenience, $I := |\mathcal{I}|$, is an alternative notation for $\mathcal{I}$'s size/cardinality) and MELD/PELD scores $j \in \mathcal{J} : \{1, 2, \ldots, J\}$ where $J \in \mathbb{N}^*$ arrive at rate $\lambda_{ij}(t) \in \mathbb{R}_+$ at time $t \in \mathcal{T}$; each $\lambda_{ij}(t)$ is calibrated to OPTN data. $\boldsymbol{\lambda}(t) \in \mathbb{R}_+^{|\mathcal{I}|\times|\mathcal{J}|}$ denotes the patient arrival rate vector at $t \in \mathcal{T}$, and $\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{I}|\times|\mathcal{J}|} \times \mathcal{T}$. With slight abuse of notation, in this chapter, the subscript ij indicates the corresponding static class-i and dynamic class-j patient group's index in a vector; the double subscript $ij, i'j'$ is the corresponding index of a matrix or vector of a *patient pair* consisting of a patient of class ij and another patient of class $i'j'$.

Patients may renege from the lists due to a) death; b) improving health conditions

(to lower MELD or PELD score) removing the need for a transplant; or c) worsening health conditions (to higher MELD or PELD scores) rendering them ineligible for a transplant. Let $d_{ij} \in \mathbb{R}_+$ be the death rate of queue ij per unit queue length and unit time, and $\mathbf{d} \in \mathbb{R}_{++}^{I \cdot J}$ be the death rate vector; $\alpha_{ij',ij}$ be the rate of patient transitioning from queue ij to queue ij', $j' \neq j$. Note that a patient' static class (i.e., size) is constant in time, only their dynamic class (i.e. health condition) may be time-varying, therefore, $\alpha_{ij,i'j'} = 0, \forall i \neq i'$. By definition of the transition probabilities, $\alpha_{ij,ij'} \geq 0$, $\forall i, j, j' \neq j$. Patients leave the waitlists, due to improved health condition in which a liver transplant is no longer needed, with probability $\beta \in \mathbb{R}^{I \cdot J}$; or conversely, patients are removed from the waitlists when they become too sick to receive a transplant, with probability $\gamma \in \mathbb{R}^{I \cdot J}$. Including transitions capturing reneging, the following holds: $\sum_{j' \neq j} \alpha_{ij',ij} + d_{ij} + \beta_{ij} + \gamma_{ij} = 1$; and $\alpha_{ij,ij} = -d_{ij} - \beta_{ij} - \gamma_{ij} - \sum_{j' \neq j} \alpha_{ij',ij}, \forall i, j$. The queueing transition matrix, $\mathbf{\Psi}$, satisfies: $(\mathbf{\Psi})_{ij,i'j'} = \mathbb{1}(i' = i)\alpha_{ij,i'j'}$.

On the liver supply side, we assume that livers of size $\ell \in \mathcal{L} := \{S, M, ML, L\}$ (and *splittable* livers of size ℓ) arrive to the system at rate $\boldsymbol{\mu}^\ell(t)$ (and $\bar{\boldsymbol{\mu}}^\ell(t)$) at $t \in \mathcal{T}$, where the $\boldsymbol{\mu}^\ell(t), \bar{\boldsymbol{\mu}}^\ell(t)$ are again calibrated to data. We denote $\boldsymbol{\mu}, \bar{\boldsymbol{\mu}} \in \mathcal{L} \times \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|} \times \mathcal{T}$. We incorporate liver quality as an attribute as our extension in the Appendix. Figure 2.1 shows a schematic of valid liver allocations, by size, in our model.

We denote the fluid queue length at time $t \in \mathcal{T}$ as $\mathbf{x}(t) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|}$, and $\mathbf{x} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|} \times \mathcal{T}$. The initial fluid queue length is greater than $\mathbf{0}$, i.e., $\mathbf{x}(0) > 0$. For each $\ell \in \mathcal{L}$, let $(\mathbf{u}^\ell, \mathbf{s}^\ell) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|} \times \mathcal{T} \times \mathbb{R}^{|\mathcal{I}|^2 \times |\mathcal{J}|^2} \times \mathcal{T}$ be the decision variable, i.e., the liver allocation, where $\mathbf{u}^\ell$ denotes the allocation rate of liver type ℓ for WLT and $\mathbf{s}^\ell$ for SLT. For convenience, define $\mathbf{U} = [\mathbf{u}^1 \quad \mathbf{u}^2 \cdots \quad \mathbf{u}^{|\mathcal{L}|}]^\top \in |\mathcal{L}| \times \mathbb{R}^{|\mathcal{I}| \cdot |\mathcal{J}|} \times \mathcal{T}$ and $\mathbf{S} =$

$[\mathbf{s}^1 \quad \mathbf{s}^2 \cdots \quad \mathbf{s}^{|\mathcal{L}|}]^\top \in |\mathcal{L}| \times \mathbb{R}^{|\mathcal{I}|^2 \times |\mathcal{J}|^2} \times \mathcal{T}$. Thus, $(\mathbf{U}, \mathbf{S})$ would be the allocation policy during the planning horizon $\mathcal{T}$. Naturally, $\mathbf{u}^\ell(t) \in \mathbb{R}^{|\mathcal{I}| \cdot |\mathcal{J}|}, \mathbf{s}^\ell(t) \in \mathbb{R}^{|\mathcal{I}|^2 \cdot |\mathcal{J}|^2}$ denote the decision rules of allocating type ℓ livers for WLT and SLT uses at time $t \in \mathcal{T}$; and $\mathbf{U}(t) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}| \times |\mathcal{L}|}, \mathbf{S}(t) \in \mathbb{R}^{|\mathcal{I}|^2 \times |\mathcal{J}|^2 \times |\mathcal{L}|}$ are the decision rules of all-liver-type allocation for WLT and SLT uses at time $t \in \mathcal{T}$, respectively. Similarly, for each $\ell \in \mathcal{L}$, let $\mathbf{P}^\ell \in [0,1]^{|\mathcal{I}||\mathcal{J}| \times |\mathcal{I}||\mathcal{J}|}, \bar{\mathbf{P}}^\ell \in [0,1]^{|\mathcal{I}||\mathcal{J}| \times |\mathcal{I}|^2 |\mathcal{J}|^2}$ be the expected WLT and SLT liver offer acceptance probability matrices, respectively: P^ℓ_{ij} denotes the probability of a type ℓ liver eventually being accepted by a type-ij patient during its match run; while $\bar{P}^\ell_{ij,i'j'}$ denotes the probability of a type ℓ liver eventually being accepted by a type-ij patient and a type-$i'j'$ patient during its match run. In other words, $\bar{P}^\ell_{ij,i'j'} = \bar{P}^\ell_{ij,(ij,i'j')} = \bar{P}^\ell_{i'j',(ij,i'j')}$. In a liver's match run, the whole or partial liver is offered sequentially to a list of candidates; if it is not accepted by any one on the list during the *cold ischemia time* (which lasts for at most $12 \sim 18$ hours), the organ expires and has to be discarded. In this chapter, we assume that a liver is offered sequentially to patients of the same type. For example, medium-sized adults with MELD scores greater or equal to 35 within 500 nautical miles of the UCSF transplant center. Note that for a whole or partial liver to be successfully transplanted, first the liver has to be accepted by a candidate, and then the transplant surgery has to be successful. In Subsection A.2.6, we show that retransplantation can be easily included in our framework.

2.3.1 The Base Model: Optimize over a Single Utility Objective with Hard Fairness Constraints

In the base case, we consider a single utility objective (e.g. NPDWT, QALY, etc.) and hard fairness constraints, i.e., a predefined proportion of arrivals of each class must

be offered a transplant before leaving the wait lists. The fluid optimization problem with the sole objective of minimizing NPDWT subject to the equity constraint (2.6) is:

$$\min_{(\mathbf{U},\mathbf{S})} \quad \text{OBJ}^{\text{NPDWT}} := \int_0^T \mathbf{d}^\top \mathbf{x}(t)dt \tag{2.1}$$

$$s.t. \quad \mathbf{x}(t) \geq 0, \quad \forall t \in [0,T) \tag{2.2}$$

$$\mathbf{u}^\ell(t), \mathbf{s}^\ell(t) \geq 0 \quad \forall \ell, t \in [0,T] \tag{2.3}$$

$$\mathbf{1}_{I\cdot J}\mathbf{u}^\ell(t) + \mathbf{1}_{I^2\times J^2}\mathbf{s}^\ell(t) \leq \boldsymbol{\mu}^\ell(t) \quad \forall \ell, t \in [0,T] \tag{2.4}$$

$$\mathbf{1}_{I^2\cdot J^2}\mathbf{s}^\ell(t) \leq \bar{\boldsymbol{\mu}}^\ell(t) \quad \forall \ell, t \in [0,T] \tag{2.5}$$

$$\sum_\ell \mathbf{u}^\ell(t) + \sum_\ell \mathbf{Z}\mathbf{s}^\ell(t) \geq \boldsymbol{\Theta}\boldsymbol{\lambda}(t) \quad \forall t \tag{2.6}$$

$$\dot{\mathbf{x}}^+(t) = \boldsymbol{\lambda}(t) - \sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) - \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t) + \boldsymbol{\Psi}\mathbf{x}(t) \quad \forall t \in [0,T) \tag{2.7}$$

In our fluid optimization problem (2.1) $\sim$ (2.7): The objective (2.1) minimizes the cumulative patient deaths across all patient groups before transplants. The inequality constraint (2.2) mandates that the fluid limits should be non-negative, which is equivalent to require $\dot{x}_{ij}(t) \geq 0$, if $x_{ij}(t) = 0$, for any i, j and t. Note that due to the properties of the patient transition matrix $\boldsymbol{\Psi}$, the constraint (2.2) is imposed on the allocation policy $(\mathbf{U}, \mathbf{S})$, not on the patient queue length process $\mathbf{x}(t)$. To see why this is the case, recall that $\boldsymbol{\Psi}$ is the queueing transition matrix. All $\boldsymbol{\Psi}$'s diagonal elements are negative (describing the rate of patients leaving the queue due to deaths, improved health, or transitioning to other queues); i.e., $\Psi_{ij,ij} < 0$. Still, non-diagonal variables are non-negative (describing the rate of patients transitioning from other queues to this queue), i.e., $\Psi_{ij,i'j'} \geq 0$ for $i = i', j \neq j'$ and $\Psi_{ij,i'j'} = 0$ if $i \neq i'$. When the queue for patient type ij (i.e., static class i and dy-

namic class j) becomes empty, i.e., $\mathbf{x}_{ij}(t) = 0$, we have $(\mathbf{\Psi x}(t))_{ij} = \Psi_{ij,ij} \cdot x_{ij}(t) + \sum_{(i',j')\neq(i,j)} \Psi_{ij,i'j'} \cdot x_{i'j'}(t) = \Psi_{ij,ij} \cdot 0 + \sum_{(i',j')\neq(i,j)} \Psi_{ij,i'j'} \cdot x_{i'j'}(t) \geq 0$. If our control variables $(u^\ell_{ij}(t), \mathbf{s}^\ell_{ij,:}(t)) = (0, \mathbf{0})$ for ℓ, then when $x_{ij}(t) = 0$, we have $\dot{x}^+_{ij}(t) = \lambda_{ij}(t) - \sum_\ell \mathbf{P}^\ell_{ij} u^\ell_{ij}(t) - \sum_{i',j'} \left(\bar{\mathbf{P}}^\ell_{ij,i'j'} s^\ell_{ij,i'j'}(t) + \bar{\mathbf{P}}^\ell_{i'j',ij} s^\ell_{i'j',ij}(t)\right) \geq \lambda_{ij}(t) + (\mathbf{\Psi x}(t))_{ij} = \lambda_{ij}(t) - 0 - 0 + (\mathbf{\Psi x}(t))_{ij} \geq 0$. In other words, the queue lengths $\mathbf{x}(t)$ have a natural tendency to stay at or go above $\mathbf{0}$ unless our allocation policy $(\mathbf{U}, \mathbf{S})$ imposes a downward force. The constraint $\mathbf{x} \geq 0$ is essentially restricting the feasible set of control variables $(\mathbf{U}, \mathbf{S})$.

The inequality constraint (2.3) requires that the amount of allocated whole or partial livers are non-negative as well. Equation (2.4) restrains the total amount of whole and split liver to not exceed the amount of available livers, while (2.5) says that the amount of assigned split livers should not be more than the total amount of medically-splittable livers. (2.6) requires that Θ_{ij} of all arrivals at queue ij at time t gets a transplant offer, where $\mathbf{Z} \in \mathbb{R}^{IJ\times IJ\cdot IJ}$ satisfies $(\mathbf{Z})_{ij,(i_1j_1,i_2j_2)} = \mathbb{1}(i = i_1, j = j_1) + \mathbb{1}(i = i_2, j = j_2)$. (Multiplying $\mathbf{Z}$ is necessary due to the differences in dimensionality between $\mathbf{u}^\ell$ and $\mathbf{s}^\ell$.) Note that $\mathbf{\Theta}$ is a predefined diagonal matrix. Thus, fairness is modeled as hard class-specific constraints (because Θ_{ij}'s can be different for each patient group ij). Alternative forms of fairness constraints can be used to replace (2.6); we discuss fairness soft constraints in Section 2.3.2, and cumulative maxmin probabilistic fairness constraints in Subsections A.2.1. Finally, (2.7) captures the evolution of the fluid process. Note that zero values in P^ℓ_{ij} and $\bar{P}^\ell_{ij,i'j'}$ imply "bad matchings" and thus size constraints are incorporated.

Besides NPDWT, quality-adjust life years (QALY) is another important utilitarian

measure of interest. Define q_{ij} as the expected QALY per unit queue length per unit time for a waiting patient of class ij, and let $H_{ij}^{\ell}/\bar{H}_{ij}^{\ell}$ denote the expected additional QALY a successful WLT/SLT earns for recipient(s). For convenience, $\mathbf{q} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$, $\mathbf{H}^{\ell} \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|}$, and $\bar{\mathbf{H}}^{\ell} \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|}$ are the matrix forms of the scalar notations above. The QALY-maximizing version of the fluid control problem can be written as follows:

$$\max_{(\mathbf{U},\mathbf{S})} \quad \text{OBJ}^{\text{QALY}} := \int_0^T \left\{ \mathbf{q}^{\top}\mathbf{x}(t) + \sum_{\ell} \mathbf{H}^{\ell}\mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) + \mathbf{H}^{\ell}\bar{\mathbf{P}}^{\ell}\mathbf{s}^{\ell}(t) \right\} dt \tag{2.8}$$

$$s.t. \quad (2.2) \sim (2.7) \tag{2.9}$$

Specifically, in (2.8) we accumulate total waiting patient QALY (i.e., $\int_0^T \mathbf{q}^{\top}\mathbf{x}(t)dt$) and total transplanted patient additional QALY (i.e. $\sum \int_0^T \mathbf{H}^{\ell}\mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) + \bar{\mathbf{H}}^{\ell}\bar{\mathbf{P}}^{\ell}\mathbf{s}^{\ell}(t)$). Let $\boldsymbol{\zeta} \in [0,1]^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{L}|}$, $\bar{\boldsymbol{\zeta}} \in [0,1]^{|\mathcal{I}|^2\times|\mathcal{J}|^2\times|\mathcal{L}|}$ denote the post transplant death vector for WLT and SLT, respectively. We can also write out explicit expressions for the TNPD and NPDAT objectives:

$$\min_{(\mathbf{U},\mathbf{S})} \quad \text{OBJ}^{\text{TNPD}} := \int_0^T \left(\mathbf{d}^{\top}\mathbf{x}(t)dt + \sum_{\ell} \boldsymbol{\zeta}^{\ell}\mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) + \bar{\boldsymbol{\zeta}}^{\ell}\bar{\mathbf{P}}^{\ell}\mathbf{s}^{\ell}(t) \right) dt \tag{2.10}$$

$$\min_{(\mathbf{U},\mathbf{S})} \quad \text{OBJ}^{\text{NPDAT}} := \sum_{\ell} \int_0^T \left(\boldsymbol{\zeta}^{\ell}\mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) + \bar{\boldsymbol{\zeta}}^{\ell}\bar{\mathbf{P}}^{\ell}\mathbf{s}^{\ell}(t) \right) dt \tag{2.11}$$

The objective NPDAT (2.11) captures the post-transplant deaths explicitly, while QALY (2.8) includes post-transplant adjusted life years and NPDWT (2.1) only considers pre-transplant deaths. TNPD (2.10) includes both pre- and post-transplant deaths. In practice, NPDWT and QALY are mostly commonly used transplant objectives (Akan et al., 2012). While we demonstrate our proposed methods mainly using NPDWT and QALY in Section 2.4, all insights and techniques directly generalize to other transplantation objectives (see Section A.1 in Appendix) and all objectives

mentioned above are tested in our numerical experiments (see Section 2.6).

2.3.2 Optimizing over a Single Utility Objective with Soft Fairness Constraints

For some situations, applying hard constraints may render the optimization problem infeasible. In such cases, we may formulate fairness as soft constraints. Specifically, we rewrite (2.6) as $\sum_\ell \mathbf{u}^\ell(t)+\sum_\ell \mathbf{s}^\ell(t) \geq \boldsymbol{\Theta}\boldsymbol{\lambda}-\boldsymbol{\xi}(t) \quad \forall i,j,t$. In scalar form it becomes

$$\sum_\ell u_{ij}^\ell(t) + \sum_\ell \sum_{i',j'} s_{ij,i'j'}^\ell(t) + s_{\ell,i'j',ij}(t) \geq \Theta_{ij}\lambda_{ij}(t) - \xi_{ij}(t) \quad \forall i,j,t \tag{2.12}$$

where $\xi_{ij}(t) \geq 0$ is the maximum allowable "fairness deficit" for queue ij; $\boldsymbol{\xi}$ is the fairness deficit in vector form. The larger the deficit at time t, the easier (2.12) is to be satisfied. At the same time, we replace the objective (2.1) with

$$\min_{(\mathbf{U},\mathbf{S}),\xi\in\mathcal{F}} \quad \text{OBJ}^{\text{NPDWT}} + \left(\mathbf{w}^{\text{NPDWT}}\right)^\top \int_{t=0}^{T} \boldsymbol{\xi}(t) \tag{2.13}$$

$$s.t. \quad (2.2) \sim (2.5), (2.7), (2.12) \tag{2.14}$$

where $\mathcal{F} \subseteq \mathbb{R}_+^{|\mathcal{I}|\times|\mathcal{J}|}$ is the set of all feasible fairness deficit vectors, $\boldsymbol{\xi} \in \mathcal{F}$ is the vector of permissible "fairness deficits" for all patient groups, and $\mathbf{w}^{\text{NPDWT}} \in \mathbb{R}_+^{|\mathcal{I}|\times|\mathcal{J}|}$ is the weight for the fairness objective, or in other words, the predefined penalty rate for deviating from the fairness constraints in the objective function (2.13). We can replace minimizing NPDWT with other optimization objectives, e.g. maximizing QALY. However, when we are maximizing a utility function, the corresponding weight for the fairness objective should be non-positive.

2.3.3 A Multi-Objective Optimization Framework

At times we may wish to optimize over multiple utility objectives; in such cases, the multi-objective fluid optimization problem can be solved to give a solution that balances these different and potentially conflicting objectives. For example:

$$\max_{(\mathbf{U},\mathbf{S}),\xi\in\mathcal{F}} \quad \text{OBJ}^{\text{Multi}} := -\text{OBJ}^{\text{NPDWT}} + \eta\text{OBJ}^{\text{QALY}} - \left(\mathbf{w}^{\text{NPDWT}} - \eta\mathbf{w}^{\text{QALY}}\right)^{\top} \int_{t=0}^{T} \boldsymbol{\xi}(t)dt \tag{2.15}$$

$$s.t. \quad (2.2) \sim (2.5), (2.7), (2.12) \tag{2.16}$$

Above, $\eta \in \mathbb{R}_+$ is the weight for the QALY objective.

2.4 Fluid Limit Decomposition and Exact Solutions to the Fluid Models in the Interior Case

This section presents one of the main results of this work: our fluid limit decomposition. This decomposition gives the explicit optimal solution to the fluid optimization problem (2.1) $\sim$ (2.7) (we call this the "fluid optimal," or "optimal split, optimal allocation" policy) in the interior of the state space: We only need to solve a standard LP at each time $t \in [0, T]$, e.g., we find the solution that greedily maximizes the index $U_{\ell,t}^{NPDWT}$ in (2.20) for each ℓ at t (see Proposition 1 below). Our solution confirms that the fluid solution is a "dynamic index policy"— a key finding in Akan et al., 2012. However, instead of solving the complex dual fluid optimization problem to approximate the optimal solution to the primal problem as Akan et al., 2012 did, our solution is analytical and exact, giving the explicit dynamic index that

each decision is trying to optimize, e.g., $U_{\ell,t}^{NPDWT}$ in (2.20). We show that our solution through decomposition is optimal in the interior of the state space, i.e., when $\mathbf{x}(t) > 0$, $\forall t \in \mathcal{T} \setminus \{T\}$.

2.4.1 Fluid Limit Decomposition for the Base Model

We first note that a queue ij's fluid limit x_{ij} satisfies one of the following conditions: $x_{ij}(t) = 0$ or $x_{ij}(t) > 0$ at any time $t \in \mathcal{T}$, because of the non-negativity constraint (2.2).

The interior case: Consider a scenario where no nonzero fluid limit hits zero in any interval $[t, t+\delta]$, i.e., $x_{ij}(t) > 0$, $\forall t \in \mathcal{T}$, $\forall \delta \in [0, T-t]$, and $\forall i \in \mathcal{I}, j \in \mathcal{J}$. Dropping (2.2) in the fluid optimization problem (2.1) $\sim$ (2.7), and recognizing that the differential equation (2.7) is a first-order, non-homogeneous, constant parameter differential equation, we can explicitly write $\mathbf{x}(t)$ as a function of $(\mathbf{u}, \mathbf{s})$ and the initial condition $\mathbf{x}(0)$:

$$\mathbf{x}(\tau) = \exp[\tau\mathbf{\Psi}]\mathbf{x}(0) + \int_0^\tau \exp[(\tau - t)\mathbf{\Psi}]F(\mathbf{U}(t), \mathbf{S}(t))dt \quad \forall \tau \in [0, T], \tag{2.17}$$

where the matrix exponential function is defined as $e^{\Psi} := \sum_{k=0}^{\infty} \frac{1}{k!}\Psi^k$. We also define:

$$F(\mathbf{U}(\tau), \mathbf{S}(\tau)) := \lambda(t) - \sum_{\ell} \left(\mathbf{P}^\ell \mathbf{u}^\ell(\tau) + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(\tau)\right). \tag{2.18}$$

Although we focused on a specific functional form of $F(\cdot)$, i.e., (2.18), in Section 2.4, our decomposition technique applies to any general-form, integrable function $F(\cdot)$ (please refer to Sections 2.5.3 and A.2.3 for detailed discussions). Written in scalar

form, (2.18) becomes

$$F_{ij}(\mathbf{U}(\tau),\mathbf{S}(\tau)) := \lambda_{ij}(t) - \sum_{\ell}\left(P^{\ell}_{ij}u^{\ell}_{ij}(\tau) + \sum_{i',j'}\left(\bar{P}^{\ell}_{ij,i'j'}s^{\ell}_{ij,i'j'}(\tau) + \bar{P}^{\ell}_{i'j',ij}s^{\ell}_{i'j',ij}(\tau)\right)\right). \tag{2.19}$$

Please see Section A.3 for discussions on the interior case and sufficient conditions for our original fluid optimization problem (2.1) $\sim$ (2.7) to stay in the interior of the state space.

Proposition 1. In the interior case, the optimal decision at any $t \in \mathcal{T}$ in the optimal solution to the fluid optimization problem (2.1) $\sim$ (2.7) is equivalent to the solution to the following LP:

$$\max_{(\mathbf{U}(t),\mathbf{S}(t))} \mathscr{U}_t^{\text{NPDWT}} := \mathbf{d}^{\top}\left\{\int_t^T \exp[(\tau-t)\mathbf{\Psi}]d\tau\right\}\left(\sum_{\ell}\mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) + \bar{\mathbf{P}}^{\ell}\mathbf{s}^{\ell}(t)\right) \tag{2.20}$$

$$s.t.\ (2.3) \sim (2.6), (2.26) \tag{2.21}$$

The optimal allocation policy with the pure objective to minimize queueing deaths solves (2.20) $\sim$ (2.21) for each $t \in \mathcal{T}$ and $\ell \in \mathcal{L}$. $\mathscr{U}_t^{\text{NPDWT}}(\mathbf{U}(t),\mathbf{S}(t))$ represents the utility of the current decision $(\mathbf{U}(t),\mathbf{S}(t))$ on reducing patient deaths from now on to the end of horizon.

In other words, we allocate livers as much as possible (subject to constraints) to the patient groups with the largest index(es) specified by the objective function.

Proof. (a) Plugging in the explicit expressions for the fluid limit vector $\mathbf{x}$ (2.17) $\sim$

(2.18):

$$\min_{(\mathbf{U},\mathbf{S})} \int_{\tau=0}^{T} \mathbf{d}^\top \exp[\tau\boldsymbol{\Psi}]\mathbf{x}(0)d\tau + \int_{\tau=0}^{T} \mathbf{d}^\top \int_0^\tau \exp[(\tau-t)\boldsymbol{\Psi}] \left(\lambda(t) - \sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) - \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t)\right) dtd\tau, \tag{2.22}$$

this objective written in the vector form is equivalent to

$$\begin{aligned}
&\min_{(\mathbf{U},\mathbf{S})} \int_{\tau=0}^{T} \sum_{i,j} d_{ij} \left(\exp[\tau\boldsymbol{\Psi}]\right)_{ij,i'j'} x_{i'j'}(0)d\tau \qquad (2.23)\\
&+ \int_{\tau=0}^{T} \sum_{i,j} d_{ij} \sum_{i',j'} \int_0^\tau \left(\exp[(\tau-t)\boldsymbol{\Psi}]\right)_{ij,i'j'} \lambda_{i'j'}(t)dtd\tau\\
&- \int_{\tau=0}^{T} \sum_{i,j} d_{ij} \sum_{i',j'} \int_0^\tau \left(\exp[(\tau-t)\boldsymbol{\Psi}]\right)_{ij,i'j'} \left(\sum_\ell P^\ell_{i'j'} u^\ell_{i'j'}(t)\right.\\
&\qquad\qquad \left. + \sum_{i'',j''} \bar{P}^\ell_{i'j',i''j''} s^\ell_{i'j',i''j''}(t) + \bar{P}^\ell_{i''j'',i'j'} s^\ell_{i''j'',i'j'}(t)\right) dtd\tau.
\end{aligned}$$

Dropping the constants, it can be further simplified to

$$\begin{aligned}
\min_{(\mathbf{U},\mathbf{S})} \int_{\tau=0}^{T} \sum_{i,j} d_{ij} \sum_{i',j'} \int_0^\tau \left(\exp[(\tau-t)\boldsymbol{\Psi}]\right)_{ij,i'j'} \left(-\sum_\ell P^\ell_{i'j'} u^\ell_{i'j'}(t)\right.\\
\left. - \sum_{i'',j''} \bar{P}^\ell_{i'j',i''j''} s^\ell_{i'j',i''j''}(t) - \bar{P}^\ell_{i''j'',i'j'} s^\ell_{i''j'',i'j'}(t)\right) dtd\tau.
\end{aligned}$$

Recall that in the interior case patient buffers are always non-empty. Thus, decisions decompose, and the expression above can be further decomposed into

$$\begin{aligned}
\max_{(\mathbf{U},\mathbf{S})} \int_{\tau=0}^{T} \sum_{i,j} d_{ij} \sum_{i',j'} \int_0^\tau \left(\exp[(\tau-t)\boldsymbol{\Psi}]\right)_{ij,i'j'} \left(\sum_\ell P^\ell_{i'j'} u^\ell_{i'j'}(t)\right.\\
\left. + \sum_{i'',j''} \bar{P}^\ell_{i'j',i''j''} s^\ell_{i'j',i''j''}(t) + \bar{P}^\ell_{i''j'',i'j'} s^\ell_{i''j'',i'j'}(t)\right) dtd\tau.
\end{aligned}$$

Now, we use a transformation to obtain a simple, myopic objective. More specifically, we look at the decision rule at each $t \in [0,T]$: The term $g_{\ell,ij,i'j'}(\tau,t) := (\exp[(\tau-t)\Psi])_{ij,i'j'}\left(P^{\ell}_{i'j'}u^{\ell}_{i'j'}(t)+\bar{P}^{\ell}_{i'j'}s^{\ell}_{i'j'}(t)\right)$ appears for all $T \geq \tau \geq t$, but not for $0 \leq \tau \leq t$; therefore,

$$\int_{\tau=0}^{T}\int_{t=0}^{\tau} g_{\ell,ij,i'j'}(\tau,t)dtd\tau = \int_{t=0}^{T}\int_{\tau=t}^{T} g_{\ell,ij,i'j'}(\tau,t)d\tau dt. \tag{2.24}$$

Note that the RHS of (2.24) is an integration over decisions at t for all $t \in \mathcal{T}$, and the expected influences of all decisions are fully extracted (looking forward), thus we do not see carry-over effects (we only have u and s variables in the expression); whereas in the original objective function (2.1), x appears and thus we couldn't decompose decision rules for each time t. Furthermore, because of the way we characterize capacity constraints ((2.4) $\sim$ (2.5)) in the fluid approximation, decisions at time t to optimize over $\int_{\tau=t}^{T} g_{\ell,ij,i'j'}(\tau,t)d\tau$ are independent. This is reasonable because in general livers deteriorate quickly, and they are allocated as soon as they become available. In other words, we extract the exact expected "impact" or influence of our decision at time t explicitly, and we can then solve the fluid optimal policy by directly optimizing over this impact at each time t:

$$\begin{aligned}\max_{(\mathbf{U}(t),\mathbf{S}(t))} \sum_{\ell}\sum_{i,j} d_{ij}\sum_{i'j'}\left\{\int_{t}^{T}(\exp[(\tau-t)\mathbf{\Psi}])_{ij,i'j'}\,d\tau\right\}\Bigg(\sum_{\ell} P^{\ell}_{i'j'}u^{\ell}_{i'j'}(t) \\ + \sum_{i'',j''}\bar{P}^{\ell}_{i'j',i''j''}s^{\ell}_{i'j',i''j''}(t)+\bar{P}^{\ell}_{i''j'',i'j'}s^{\ell}_{i''j'',i'j'}(t)\Bigg)\end{aligned} \tag{2.25}$$

Writing (2.25) in vector form we arrive at (2.20). □ □

Note that in the decomposed optimization problem, we only have simple constraints

(i.e. (2.3) $\sim$ (2.6), capacity and non-negativity constraints) at time t on u and s; crucially, there is no differential equation for fluid dynamics (i.e., (2.7) is not a constraint of the decomposed optimization problem, as we have incorporated the fluid evolution into the objective).

Although the liver transplant waitlists are overloaded, for theoretical completeness, we also discuss the *boundary* case:

The boundary case: Consider the more general case where $x_{ij}(t) \geq 0$, $\forall t \in \mathcal{T}$ and $i \in \mathcal{I}, j \in \mathcal{J}$. When $x_{ij}(t) = 0$, the right hand side (RHS) of (2.7) has to be non-negative, $t \in [0, T)$; otherwise, (2.2) is violated at time $t + \delta$, where $\delta \to 0^+$. More specifically, the following inequality must hold when $x_{ij} = 0$ for a particular ij patient group:

$$\lambda_{ij}(t) - \sum_{\ell} \left(P^{\ell}_{ij} u^{\ell}_{ij}(t) + \sum_{i',j'} \left(\bar{P}^{\ell}_{ij,i'j'} s^{\ell}_{ij,i'j'}(t) + \bar{P}^{\ell}_{i'j',ij} s^{\ell}_{i'j',ij}(t) \right) \right) + \sum_{i',j'} \Psi_{ij,i'j'} x_{i'j'}(t) \geq 0 \quad \forall i, j, t \text{ s.t. } x_{ij}(t) = 0 \tag{2.26}$$

Note that we can recapture (2.2) by enforcing (2.26) on zero-valued fluid limit(s) at each time t, if any. Our next decomposition result is closely related to the above explicit expression for $\mathbf{x}$ and (2.26):

In the boundary case: For each queue ij and $t \in [0, T)$ such that $x_{ij}(t) = 0$, inequality constraint (2.26) guarantees that the first-order derivative of $x_{ij}(t)$ be non-negative; in other words, $\lim_{\delta \to 0^+} \frac{x_{ij}(t+\delta) - x_{ij}(t)}{\delta} \geq 0$. Thus, $\lim_{\delta \to 0^+} x_{ij}(x + \delta) \geq x_{ij}(t) \geq 0$. Therefore, inequality constraint (2.26) is a sufficient condition for (2.2). Moreover,

because for each $t \in [0, T)$, (2.26) consists of only $\mathbf{U}(t)$, $\mathbf{S}(t)$, and $\mathbf{x}(t)$, and does not involve any $t' \in (t, T]$, we can directly add it as a constraint in our decomposed LP for t. With (2.26) in the constraints, the solutions to our decomposed optimization problem give decision rules subject to (2.2) and other constraints. With these decision rules at t's, we have a fluid-model policy that is compact and explicit. However, since (2.26) are linking constraints; our greedy policies may not always be optimal.

A similar formulation and the same decomposition techniques can be applied to the QALY objective version of the problem. Recall that $\mathbf{q} \in \mathbb{R}^{|\mathcal{I}|\cdot|\mathcal{J}|}$ denotes the vector of expected QALYs of patients on the waitlists, and $\mathbf{H}^\ell \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|\times|\mathcal{I}||\mathcal{J}|}, \bar{\mathbf{H}}^\ell \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|\times|\mathcal{I}|^2|\mathcal{J}|^2}$ are the matrices of expected QALYs of patients transplanted with type-ℓ livers in WLT and SLT, respectively.

Proposition 2. The fluid optimization problem (2.8) $\sim$ (2.9) can be decomposed into the following optimization problem:

$$\max_{\mathbf{U}(t),\mathbf{S}(t)} \mathscr{U}_t^{\text{QALY}} := -\sum_\ell \mathbf{q}^\top \left\{ \int_t^T \exp[(\tau - t)\mathbf{\Psi}] d\tau \right\} \left(\mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t) \right) + \mathbf{H}^\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{H}}^\ell \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t) \tag{2.27}$$

$$s.t.\ (2.3) \sim (2.6), (2.26) \tag{2.28}$$

In the interior case, the optimal allocation policy with the objective to maximize QALY solves (2.27) $\sim$ (2.28).

The matrix $\mathbf{\Psi}$, when estimated from real-world data with limited precision, is non-singular with very high probability; and it is indeed non-singular in our estimation using OPTN data from 2009 - 2019 (see Section 2.6 for more information.) Specifically,

$\mathbf{\Psi}$ is singular with probability 0 if we have infinite precision; and in the unlikely case that $\mathbf{\Psi}$ is singular, we can add a noise matrix $\boldsymbol{\epsilon} \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|\times|\mathcal{I}||\mathcal{J}|} \to \mathbf{0}$ so that $(\mathbf{\Psi}+\boldsymbol{\epsilon})$ is non-singular and $\lim_{\boldsymbol{\epsilon}\to 0} \mathbf{\Psi}+\boldsymbol{\epsilon} = \mathbf{\Psi}$. Below we present the simplification for our decomposed optimization problems (2.20) $\sim$ (2.21) and (2.27) $\sim$ (2.28) when $\mathbf{\Psi}$ is non-singular; in Subsection A.5, we present the simplification results with singular $\mathbf{\Psi}$'s, for theoretical completeness.

Proposition 3. When $\mathbf{\Psi}$ is non-singular, $\mathscr{U}^{NPDWT}$ and $\mathscr{U}^{QALY}$ can be simplified to

$$\mathscr{U}_t^{\text{NPDWT}} = \mathbf{d}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \left(\sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t)\right) \tag{2.29}$$

$$\begin{aligned}\mathscr{U}_t^{\text{QALY}} = -\mathbf{q}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \left(\sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t)\right) \\ + \sum_\ell \mathbf{O}^\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{O}}^\ell \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t)\end{aligned} \tag{2.30}$$

Proof. Proof: When $\mathbf{\Psi}$ is non-singular, it is invertible, i.e., $\mathbf{\Psi}^{-1}$ exists. The following always holds (Van Loan, 1978):

$$\int_{\tau=t}^{T} e^{\tau\mathbf{\Psi}} d\tau = \left(\exp(T\mathbf{\Psi}) - \exp(t\mathbf{\Psi})\right) \mathbf{\Psi}^{-1} \tag{2.31}$$

Plugging (2.31) into (2.20) and (2.27) gives (2.29) and (2.30), respectively. □

2.4.2 Fluid Limit Decomposition with Fairness as Soft Constraints

Note that at any time $t \in \mathcal{T}$, all soft fairness constraints only constrain $(\mathbf{U}(t), \mathbf{S}(t))$ and penalize the objective; therefore, we can directly borrow the previous decomposition results and write the decomposed fluid optimization problem with the NPDWT/QALY

objective at each time t:

$$\max_{(\mathbf{U}(t),\mathbf{S}(t)),\xi(t)\in\mathcal{F}} \quad \mathscr{U}_t^{\text{NPDWT/QALY}} + \left(\mathbf{w}^{\text{NPDWT/QALY}}\right)^\top \xi(t) \tag{2.32}$$

$$s.t. \quad (2.2) \sim (2.5), (2.7), (2.12) \tag{2.33}$$

2.4.3 Fluid Limit Decomposition for the Multi-Objective Framework

Collecting all cases from our results above, we can explicitly write the decomposed decision rule optimization problem for the multi-objective fluid optimization framework (2.15) $\sim$ (2.16), as stated in Theorem 2.4.1.

Theorem 2.4.1. The exact solution to the fluid optimization problem (2.15) $\sim$ (2.16) is a greedy policy, with all future impacts of current actions summarized in our explicit dynamic indices; the optimal decision rule at time $t \in \mathcal{T}$ is the solution to the following linear optimization problem:

$$\max_{(\mathbf{U}(t),\mathbf{S}(t)),\xi(t)\in\mathcal{F}} \quad \mathscr{U}_t^{\text{Multi}} := \mathscr{U}_t^{\text{NPDWT}} + \eta\mathscr{U}_t^{\text{QALY}} + \left(\mathbf{w}^{\text{NPDWT}} + \eta\mathbf{w}^{\text{QALY}}\right)^\top \xi(t) \tag{2.34}$$

$$s.t. \quad (2.3) \sim (2.5), (2.12), (2.26). \tag{2.35}$$

2.4.4 Optimality in the Interior Case and Dynamic Index Monotonicity

We have solved for the optimal policy π^* directly by finding its optimal decision rules at each $t \in \mathcal{T}$ through standard LPs, using our fluid limit decomposition method described above. The solutions to the decomposed LPs, if they exist (i.e. if all the

LPs are feasible), are the exact, globally optimal decision rules; any optimal policy we obtain from specifying its decision rules at all $t \in \mathcal{T}$ is globally optimal. Because we give the explicit LPs to solve at each $t \in \mathcal{T}$, it is clear that there exists at least one optimal policy for each single- or multi-objective fluid optimization problem with *soft* fairness constraints with sufficiently large fairness deficit $\boldsymbol{\xi}$ (i.e. $\boldsymbol{\xi} \geq \boldsymbol{\Theta}\boldsymbol{\lambda}(t)$) and no fairness constraints (i.e. $\boldsymbol{\Theta} = \mathbf{0}$). When formulated with hard fairness constraints, there is a possibility that some decomposed LPs are infeasible, which directly implies that there is no feasible decision rules or policy attaining the prescribed level of fairness.

At each $t \in \mathcal{T}$, our optimal decision rule optimizes over the objective $\mathscr{U}_t^i, i \in \{\text{NPDWT, QALY, TNPD, NPDAT, Multi}\}$. The dynamic index $\mathscr{U}_t^i$ can be interpreted as the cost or benefit of a certain decision rule $(\mathbf{U}(t), \mathbf{S}(t))$. For example, $\mathscr{U}_t^{\text{NPDWT}}$ summarizes the expected aggregate reduction in patient deaths while waiting for transplants during $[t, T]$ as a result of choosing $(\mathbf{U}(t), \mathbf{S}(t))$ at this moment t. $\mathscr{U}_t^{\text{QALY}}$ encapsulates the expected sum of QALYs increases of all the waiting and transplanted patients during $[t, T]$ under the decision rule $(\mathbf{U}(t), \mathbf{S}(t))$.

For notational convenience, denote $\mathscr{U}_t^{\text{NPDWT}}$'s coefficient vector for WLT $D(t) := \mathbf{d}^\top \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1} \sum_\ell \mathbf{P}^\ell$, $\mathscr{U}_t^{\text{NPDWT}}$'s coefficient vector for SLT: $\bar{D}(t) := \mathbf{d}^\top \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1} \sum_\ell \bar{\mathbf{P}}^\ell$. Similarly, we denote $\mathscr{U}_t^{\text{QALY}}$'s coefficient vector for WLT $Q(t) := -\mathbf{q}^\top \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1} \sum_\ell \mathbf{P}^\ell$, and $\mathscr{U}_t^{\text{QALY}}$'s coefficient vector for SLT $\bar{Q}(t) := -\mathbf{q}^\top \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1} \sum_\ell \bar{\mathbf{P}}^\ell$. Proposition 4 below shows that the coefficient vectors are monotonic functions in t:

Proposition 4. The time-varying coefficient vectors in the fluid optimization prob-

lems and dynamic indexes are monotonic functions in t: (a) $D(t)$ and $\bar{D}(t)$ are non-increasing in t, and (b) $Q(t)$ and $\bar{Q}(t)$ are nondecreasing in t.

Because, $D(t) > \mathbf{0}$, $\bar{D}(t) > \mathbf{0}$, $Q(t) \leq \mathbf{0}$, and $\bar{Q}(t) \leq \mathbf{0}$, for any $t \in \mathcal{T}$ (see Subsection A.6.4 for a formal proof). The derivatives of these time-varying coefficient vectors are in the opposite direction of their sign. Proposition 4 essentially tells us that the absolute impacts of earlier allocations are greater than later allocations in the overall transplantation objective values, which equal the cumulative sum of dynamic indices plus other decision-invariant constants. This new finding highlights the special properties of the fluid control problem for liver allocation with patient health condition transitions. The proof for Proposition 4, involving matrix calculus, is deferred to Subsection A.6.4.

2.5 Structural Properties and Extensions

2.5.1 New Insight on SLT: Supply and Fairness

This section studies the marginal benefits of having more livers, the impact of fairness, and when splitting as many livers as possible is an optimal strategy, in the interior case. For the sake of brevity, all proofs for Section 2.4.5 are deferred to the Appendix Section A.6.

Definition 2.5.1. $f(\boldsymbol{\mu}^1, ..., \boldsymbol{\mu}^{|\mathcal{L}|}, \bar{\boldsymbol{\mu}}^1, \ldots, \bar{\boldsymbol{\mu}}^{|\mathcal{L}|}, \boldsymbol{\Theta\lambda}, \boldsymbol{\lambda})$ is the optimal objective function value in (2.15) as a function of the *right hand side* (RHS) vectors $\boldsymbol{\mu}^\ell, \bar{\boldsymbol{\mu}}^\ell, \boldsymbol{\Theta\lambda}, \boldsymbol{\lambda}, \forall \ell \in \mathcal{L}$.

Proposition 5. $f(\boldsymbol{\mu}^1, ..., \boldsymbol{\mu}^{|\mathcal{L}|}, \bar{\boldsymbol{\mu}}^1, \ldots, \bar{\boldsymbol{\mu}}^{|\mathcal{L}|}, \boldsymbol{\Theta\lambda}, \boldsymbol{\lambda})$ is jointly concave.

Proposition 5 implies that the marginal benefit of resources (i.e. the cumulative multi-objective utility improvement gained from an additional unit of resources) is monotonically non-increasing. The marginal benefit of an additional liver is greater when the available livers are scarce, and the marginal benefit is smaller when there is an abundance:

Corollary 2.5.1. The marginal benefit of an additional liver is monotonically non-increasing in $\boldsymbol{\mu}$ and $\bar{\boldsymbol{\mu}}$.

Figure 2.2 illustrates the concave, increasing objective function values in $\boldsymbol{\mu}$ and $\bar{\boldsymbol{\mu}}$. While it is straightforward that increasing $\boldsymbol{\mu}$ or $\bar{\boldsymbol{\mu}}$ relaxes the feasible set, and therefore improves the objective value in (2.15), the effect of increasing λ is less clear. Specifically, the optimal objective function value f may not be monotonic in $\boldsymbol{\lambda}$. Recall that f may be a weighted sum of two objectives: Increasing $\boldsymbol{\lambda}$ increases the expected aggregate QALY (as more patients imply more QALY) but may increase the expected NPDWT (which has a negative weight in f) at the same time. We do know, though, that the effect of λ is concave.

Finally, when $\boldsymbol{\lambda}$ is fixed, increasing $\boldsymbol{\Theta}$ reduces the feasible set, thus lowering the optimal objective function value. To study the impact of fairness constraints in more detail, we introduce the *price of fairness* concept. To assure that our price of fairness is well-defined, when there is no imposed fairness (i.e. $\boldsymbol{\Theta} = \mathbf{0}$), we may choose to add a constant to the f function so that $f(\boldsymbol{\mu}^1, ..., \boldsymbol{\mu}^{|\mathcal{L}|}, \bar{\boldsymbol{\mu}}^1, \ldots, \bar{\boldsymbol{\mu}}^{|\mathcal{L}|}, \mathbf{0}, \boldsymbol{\lambda}) > 0$.

Definition 2.5.2. The price of fairness, denoted as PoF, is the relative utility loss due to the imposed fairness constraint. With fairness parameter $\boldsymbol{\Theta}$,

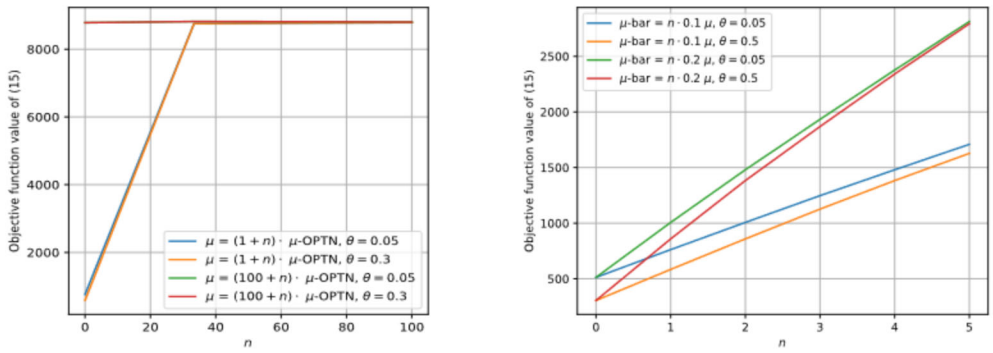


Figure 2.2: Sensitivity analysis of parameters μ and $\bar{\mu}$, based on OPTN data. Here we consider the fairness matrix to be of the following form: $\boldsymbol{\Theta} = \theta \mathbf{I}_{IJ,IJ}$, where $\theta \in [0, 1)$ is a scalar fairness level, and $\mathbf{I}_{IJ,IJ}$ is an identity matrix of dimension $IJ \times IJ$. The objective function value of (2.15) is a non-decreasing, concave function of μ and $\bar{\mu}$. On the left, different base liver supplies μ-OPTN and $100 \cdot \mu$-OPTN contribute to differences in intercepts, while fairness level θ has a relatively smaller impact on the objective function values. On the right, fairness level θ's determine the intercept; slopes are higher with larger θ and larger base. The objective functions are concave in the SLT proportion multiplier n.

$$\text{PoF}(\boldsymbol{\Theta}) = \frac{f(\boldsymbol{\mu}^1,\ldots,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,\ldots,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\mathbf{0},\boldsymbol{\lambda}) - f(\boldsymbol{\mu}^1,\ldots,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,\ldots,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\boldsymbol{\Theta}\boldsymbol{\lambda},\boldsymbol{\lambda})}{f(\boldsymbol{\mu}^1,\ldots,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,\ldots,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\mathbf{0},\boldsymbol{\lambda})}.$$

Corollary 2.5.2 describes the first- and second-order properties of PoF.

Corollary 2.5.2. With $\boldsymbol{\lambda}$ fixed, $\text{PoF}(\boldsymbol{\Theta})$ is monotonically non-decreasing and convex in $\boldsymbol{\Theta}$, over (2.15) $\sim$ (2.16)'s feasible range.

Corollary 2.5.2 shows that PoF's increase accelerates when $\boldsymbol{\Theta}$ is higher; this implies that it is crucial for policy makers to be prudent considering mandated fairness levels: Increases in fairness come at greater and greater costs with respect to efficiency. Figure 2.5b in Section 2.6 illustrates the monotonicity and convexity of PoF.

2.5.2 New Insight on When and When Not to Split

Proposition 6 presents conditions when the optimal policy in the interior case splits a medically-splittable liver. Corollary 2.5.3 describes a scenario in which splitting all splittable livers (subject to capacity and fairness constraints) is the optimal strategy. For convenience, denote e_l^m as an m-dimensional real vector where, except for the element corresponding to the lth element being 1, all other elements are 0. Consistent with our subscript conventions in this chapter, $e^{IJ}_{ij,i'j'}/e^{IJ*IJ}$ denotes a $I^2 \cdot J^2$-dimensional real vector where the element corresponding to the type-ij and type-$i'j'$ patient pair is 1 while all other elements are 0.

Proposition 6. For each splittable liver type ℓ, only split an incoming liver of this type if $\exists\ i, i' \in \mathcal{I}, j, j' \in \mathcal{J}$, s.t. (a) $\bar{D}^{\ell}_{ij,i'j'}(t)e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} D^{\ell}_{i''j''}(t)e^{IJ}_{i''j''}$, and b) $\bar{Q}^{\ell}_{ij,i'j'}(t)e^{IJ\times IJ}_{ij,i'j'} + (\bar{H}^{\ell}_{ij} + \bar{H}^{\ell}_{i'j'})\bar{P}^{\ell}_{ij,i'j'}e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} Q^{\ell}_{i''j''}(t)e^{IJ}_{i''j''} + H^{\ell}_{i''j''}P^{\ell}_{i''j''}e^{IJ}_{i''j''}$, the optimal policy (w.r.t (2.15), for any κ) tends to split as the incoming ℓ liver at t (subject to (2.5), (2.6), and (2.26)).

Proposition 6 states that splitting is optimal when there exists a patient type pair to whom, if the liver is split and transplanted into, will result in a higher positive impact on the waitlist system, than keeping the liver for the best WLT match. Our explicit dynamic indices summarize the impacts of SLT and WLT allocations.

Proposition 6 can also be used to analyze the optimal use of SLT splitting methods. There are currently two common liver-splitting methods: The “child-adult" split and the “adult-adult" split. It is possible that the “child-adult" split dominates the “adult-adult" split if we can calibrate corresponding parameters and see if they meet the

explicit conditions in Proposition 6.

Corollary 2.5.3. If for every liver type ℓ that is medically safe to be used for SLT, $\exists\ i, i' \in \mathcal{I}, j, j' \in \mathcal{J}$, s.t. (a) $\bar{D}^{\ell}_{ij,i'j'}(t)e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} D^{\ell}_{i''j''}(t)e^{IJ}_{i''j''}$, and b) $\bar{Q}^{\ell}_{ij,i'j'}(t)e^{IJ\times IJ}_{ij,i'j'} + (\bar{H}^{\ell}_{ij} + \bar{H}^{\ell}_{i'j'})\bar{P}^{\ell}_{ij,i'j'}e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} Q^{\ell}_{i''j''}(t)e^{IJ}_{i''j''} + H^{\ell}_{i''j''}P^{\ell}_{i''j''}e^{IJ}_{i''j''}$, the optimal policy (w.r.t (2.15), for any κ) splits all splittable liver at t (subject to (2.5), (2.6), and (2.26)), $\ell \in \mathcal{L}$.

Corollary 2.5.3 presents some fairly restrictive (sufficient) conditions that ensure all-split is optimal: The probability of acceptance and the gain in objective of splitting must dominate. Fortunately, given that there are two SLT patients versus a single WLT, these conditions may be satisfied, in practice.

Our formulation also yields some additional results: Appendix Section A.7 demonstrates that our fluid limit decomposition yields policy insights analogous to Propositions 1 and 2 in Akan et al., 2012. Corollary A.7.1 describes a sufficient condition where the "sickest-first" policy is optimal; and Corollary A.7.2 provides sufficient conditions under which giving priorities to certain static classes is optimal.

2.5.3 New Insight on Organ Allocation with Strategic Accept/Reject Decisions

In Section 2.3, we formulated the expected candidate acceptance and transplant success probabilities $\mathbf{P}$ (for WLT) and $\bar{\mathbf{P}}$ (for SLT) as known constants independent of our allocation policy, in line with previous fluid models used for organ allocation (Akan et al., 2012; Zenios et al., 2000b). But our decomposition technique also applies to cases where $\mathbf{P}$ and $\bar{\mathbf{P}}$ are functions of $(\mathbf{U}(t), \mathbf{S}(t))$, in other words, patients' accept/reject

decisions could be endogenous. Please refer to Section A.2.3 for detailed discussions on incorporating endogenous choices. In an example discussed in A.2.3, we solve the optimal control for an endogenous scenario where $\mathbf{P}$ and $\bar{\mathbf{P}}$ can be expressed as a linear function of $(\mathbf{U}(t), \mathbf{S}(t))$.

Optimal organ allocation with endogenous accept/reject choices has been viewed as a challenging problem: Existing analytical papers studying candidate's endogenous, strategic choices either assume the allocation policy to be exogenous (Alagoz et al., 2007a; Tunç et al., 2022) or use game-theoretic analysis to get a stable equilibrium for systems in which patients do not change classes (due to dynamic patient health conditions), and not based on queue length vectors as ours do (Ata et al., 2021; Su & Zenios, 2006). Akshat et al., 2023 builds a structural model which is simulation-based. S.-P. Kim et al., 2015 developed a module for the simulated allocation models that helps predict whether each potential recipient will accept an offered organ. The classifiers were trained using machine-learning methods (e.g., logistic regression, support vector machines, boosting, etc.) and evaluated using 2011 liver match-run data. Compared to existing work, our methodologies can be applied to optimize the organ allocation with (potentially) endogenous accept/reject choices in a transient/non-stationary, fully overloaded system using a fluid model.

2.6 Numerical Method and Results

We conduct two sets of experiments: First, we compare the objective values of our "fluid optimal" policy, with other policies within the fluid model framework. Second, we simulate the national liver allocation system and evaluate the performances of our fluid model solution against the benchmarks, moving from the fluid solution in a

deterministic environment to a simulation that captures second-order dynamics.

We base our experiments on data from the Standard Transplant Analysis and Research (STAR) files and the Potential Transplant Recipient (PTR) dataset provided by the United Network of Organ Sharing (UNOS), the Scientific Registry of Transplant Recipients (SRTR) data. Specifically, we use transplant and candidate data from January 2009 to September 2019 for the fluid model parameter estimation and the liver allocation simulation. We focus on the NPDWT and QALY objectives; we can easily extend the framework to include other objectives. Consistent with the literature, we estimate QALY using the Cox proportional hazards model (Akan et al., 2012; Cox & Oakes, 2018; Zenios et al., 2000b), while other parameters are estimated directly from STAR files and the PTR dataset.

We explore five policies. The first one is our "fluid optimal" policy, which is also called "optimal split, optimal allocation". The second one assumes all splittable livers are split as long as a patient meets the liver-split criteria and all whole and split livers are allocated optimally; we call this the "all-split, optimal allocation" policy; the "all-split, optimal allocation" policy is the solution to the fluid control problem with (2.5) replaced with $\mathbf{1}_{I^2.J^2}\mathbf{s}^\ell = \bar{\mu}^\ell, \forall \ell$. The third policy is the "no-split, optimal allocation" policy, where no livers are split, and all WLT matching is optimal; the "no-split, optimal allocation" policy is the solution to the fluid control problem with (2.5) replaced with $\mathbf{1}_{I^2.J^2}\mathbf{s}^\ell = 0, \forall \ell$. While there are many allocation policies that either split all livers or do not split any liver, we only compare with the "optimal" ones and (and later the "sickest first" ones) in the fluid model. By comparing with "optimal allocation" policies, we eliminate the influences from other variables (such

as suboptimal matching of livers and patients), highlighting the value of splitting livers optimally in a comparable and explicit manner. The fourth policy is the "all-split, sickest first" policy; in which all splittable livers are used for SLT and allocated to the sickest patient class. "All-split, sickest first" helps demonstrate the benefit of increased splitting without changing the current liver allocation priority rules in the US (i.e., "sickest first"). The fifth policy is "few split, sickest first," where the sickest patients get the highest priority, and livers are split only if the sickest patient requires an SLT. Currently, in the US, the liver allocation system is most similar to the "few split, sickest first" policy, and OPTN splits livers only in exceptional cases (only around 1% of all livers are actually split in the US according to OPTN and UNOS, 2016). Recipients of split livers are selected based on simple decision rules where sick children are prioritized.

While the Liver Simulated Allocation Model (LSAM) software created by the SRTR has been developed to support studies of alternative organ allocation policies, LSAM only supports threshold-based, open-loop policies whose thresholds/parameters are set at the beginning of the simulation. Our optimal fluid-model policy is a closed-loop policy that depends on the waitlist lengths, patient health transitions, etc. Thus, we simulate the liver allocation system without using LSAM, setting our parameters based on UNOS and SRTR data. In the simulation study, we use a discrete-time day-to-day simulation environment and run for one year. The allocation policy it uses, "optimal split, optimal allocation," is solved by decomposed LP: when the organ arrives, we input the simulated queue lengths $\mathbf{x}(t)$ into the decomposed LP and output the solution of the LP (decision rules) to the simulation model. Similarly, we use the fluid model with simulation input (e.g., actual queue lengths as $\mathbf{x}(t)$) to get the "all-

split, optimal allocation" and “no-split, optimal allocation" policies. Please refer to A.9 for a detailed description of the experimental settings and a link to access the code.

We first discuss the results of the fluid optimization with the NPDWT (2.1) and QALY objectives (2.8). Numerical results of the continuous, deterministic fluid-control problem (see Figures 2.3 and 2.4) show that our proposed matching solution achieves significantly better objective values compared against the “no-split, optimal allocation," ‘all-split, sickest first", and “few split, sickest first" policies. The performances of “sickest-first" policies are significantly worse than “optimal allocation" policies, yet “all-split, sickest first" has consistent advantages over the ‘few-split, sickest first" policy, an approximation of the current system.

Using the current UNOS/OPTN data, “all-split, optimal allocation" performs virtually as well as the “optimal split, optimal allocation" policy, confirming the effectiveness of the UK liver allocation strategy, where the decision to split all splittable livers is made first, before considering who we assign the liver/partial livers. We highlight that “optimal split, optimal allocation” outperforms “few-split, sickest first” when $\theta = 0$ by more than 4.5% for the NPDWT objcetive, and “no-split, optimal allocation” by 1.35% when $\theta = 0.8$, respectively. If sticking to “sickest first" priority rules, splitting all splittable livers can still bring consistent benefits.

With the expanded use of SLT in “optimal split, optimal allocation" and “all-split, optimal allocation" we bypass the efficiency-fairness tradeoff, as SLT increases both utility and equity, when compared to other policies.

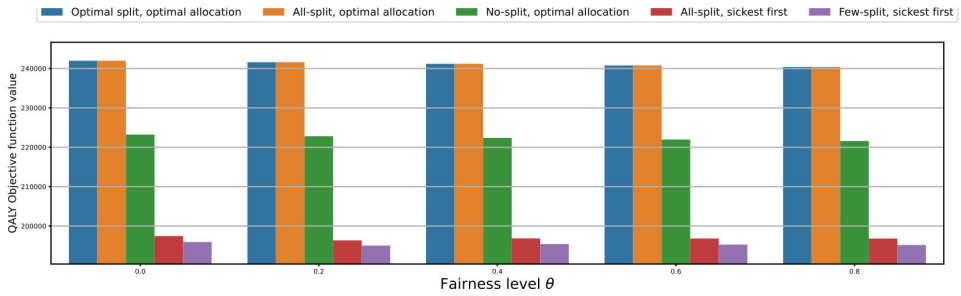


Figure 2.3: Comparisons of five policies under the maximizing QALY objective. The "optimal split, optimal allocation" is the solution of the fluid model; "all-split, optimal allocation" and "no-split optimal" are solutions to fluid models with additional constraints: all splittable livers are split, and no livers are split, respectively. "Few-split, sickest first" splits all splittable livers assume 10% of all donated livers are splittable, and allocate to the patient(s) with the highest MELD/PELD scores. The "few-split, sickest first" policy is a fair approximation of the current OPTN policy, allocating livers to the sickest patient(s) while splitting 10% of splittable livers. In this experiment, $(\boldsymbol{\Theta})_{i4,i4} = \theta$ for any $i \in \{0, 1, 2, 3\}$, meaning that the sickest patient group are guaranteed θ-probability of getting a liver at any time. $(\boldsymbol{\Theta})_{ij,ij} = 0.05\theta$, for $\forall i, j \neq 4$.

Figure 2.5a illustrates the net benefit of SLT, i.e., the objective values of (2.15) under the "optimal split, optimal allocation" subtracting those under the "no-split, optimal allocation"; positive slopes indicate that the net benefits of using SLT are increasing in θ, under different multi-objective weights (κ) and potential split liver capacities ($\bar{\boldsymbol{\mu}}$). Turning to Figure 2.5b, the price of fairness (PoF) under our "optimal split, optimal allocation" policy is moderate (less than 17.5% utility loss with a quite high hard fairness level of 0.6 across all patient groups), although the PoF curve is indeed convex (confirming our theoretical result in Section 2.5.1. These results shed insight on the potential utility and equity improvements with expanded SLT use and suggest the decision to split livers upon acquisition, and the choice of recipients can likely be decoupled in practice. In other words, the OPTN is recommended to decide to split a splittable liver first, start the procurement, and then select the recipients for the partial livers.

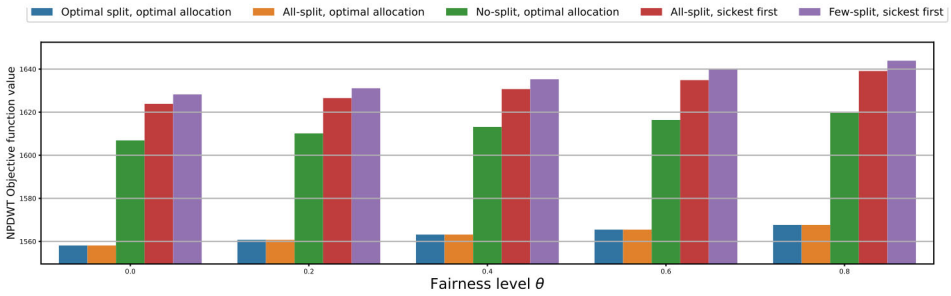
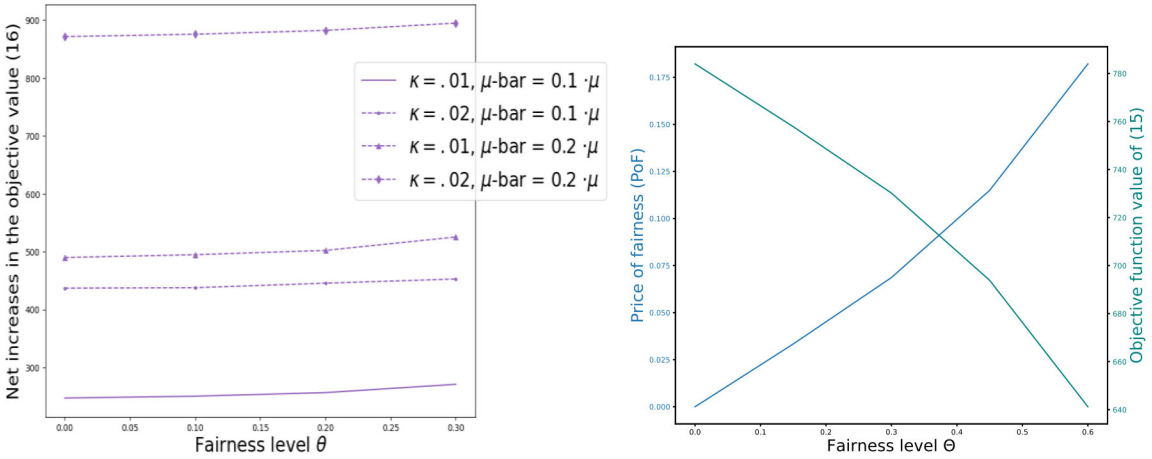


Figure 2.4: Comparisons of five policies under the minimizing NPDWT objective.



(a) The net benefit of SLT.

(b) Price of fairness

Figure 2.5: The net benefit of SLT increases as $\bar{\mu}$ increases, and is non-decreasing in θ; the price of fairness (PoF) is a monotonically increasing, convex function of fairness level θ, where $(\mathbf{\Theta})_{ij,ij} = \theta,\ \forall i, j$. $\mathbf{\Theta}$ is a matrix whose non-diagonal elements are 0.

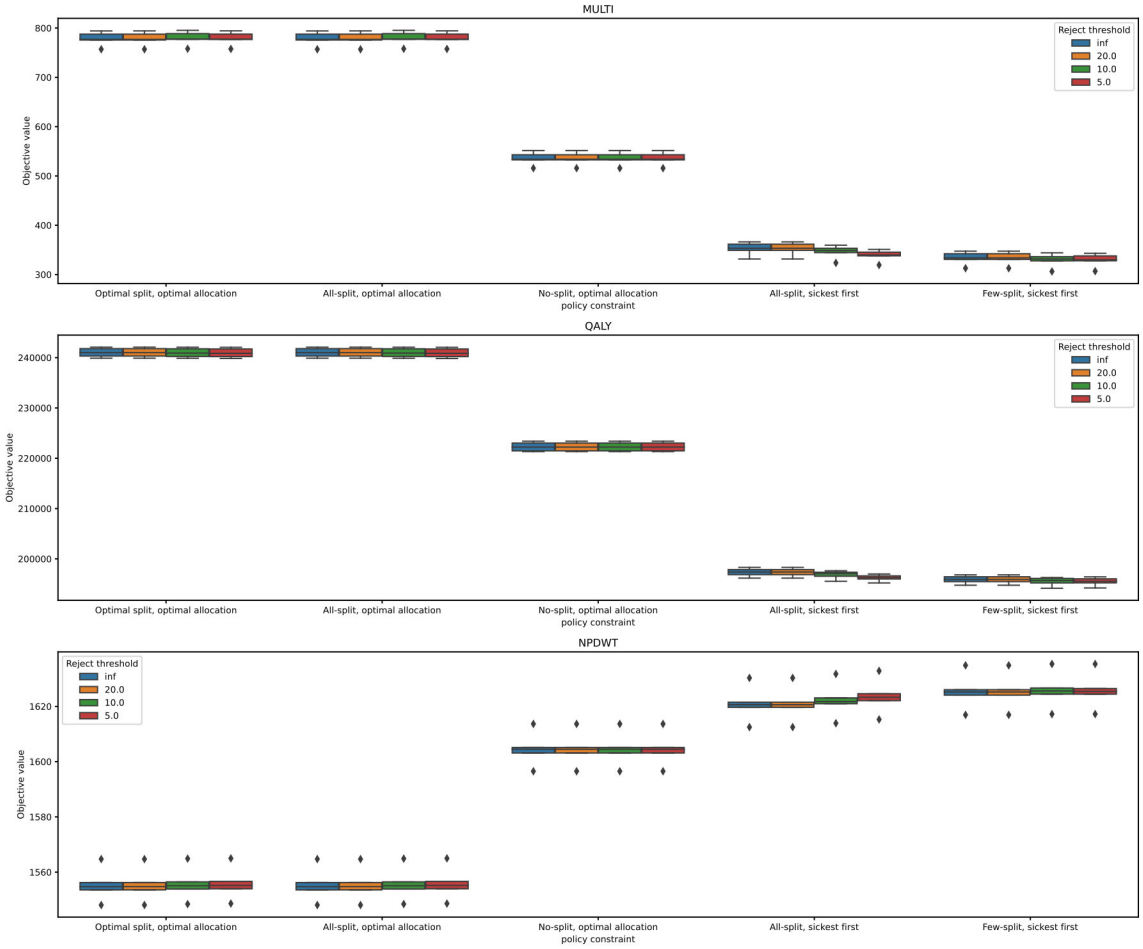


Figure 2.6: Simulation results based on OPTN data: The comparisons of five policies. For MULTI $\kappa = 0.01$ and QALY objectives, the most desirable policy maximizes the objective values; conversely, the best policies with respect to the NPDWT objective minimize the objective values. The fairness matrix $\mathbf{\Theta} = \mathbf{0}$. The *reject thresholds* capture patients' strategic accept/reject decisions: When the number of livers allocated to the patient class is greater or equal to the reject threshold, any SLT offer is rejected.

Figure 2.6 gives a summary of our simulation results and the performance comparisons among the four policies we used in the experiment ($\mathbf{\Theta} = \mathbf{0}$). The *reject thresholds* capture patients' strategic accept/reject decisions; in this simulation, $\mathbf{P}$ and $\bar{\mathbf{P}}$ are endogenously affected by the allocation policy. Specifically, we define the reject thresholds such that when $\sum_\ell P_{ij}^\ell u_{ij}^\ell(t)$ is greater or equal to the chosen reject threshold, any SLT offers are rejected. Figure 2.6 shows that while patient strategic behaviors negatively affect many policies in all objectives, the "all-split, sickest first" policy is most sensitive to reject threshold changes. Nevertheless, even incorporating endogeneity, wider use of SLT still improves the system objective function values, and our "optimal allocation" policies still outperform "sickest first" policies. Numerical results confirm that our optimal fluid-model policy outperforms all benchmarks (and is closely followed by "all-split, optimal allocation") in all objectives. In the simulation, more than 94% of all livers are transplanted to the sickest patients, and 10% of splittable livers are actually split in "few-split, sickest first." "Few-split, sickest first" policies myopically offer livers to the sickest patients to prevent immediate deaths, while the "optimal split, optimal allocation" and "all-split, optimal allocation" achieve lower NPDWT, QALY, and MULTI objectives, thanks to their proactivity and increased use of SLT. "All-split, sickest first" also performs consistently better than "few-split, sickest first" in all objectives. This shows that offering livers to the sickest patients does not even minimize waitlist deaths.

The fact that "optimal split, optimal allocation" and "all-split, optimal allocation" policies both perform better than no-split and "sickest-first" (in terms of NPDWT, MULTI, and QALY) demonstrate the value of expanding the SLT use in the US. The current OPTN policy is approximated by the "few split, sickest first" policy above.

In Section A.9.2, we present numerical results for the multi-objective fluid model performance comparison and additional reject thresholds for the simulation study.

2.7 Concluding Remarks

We provide theoretical and experimental results indicating that the expanded utilization of SLT can both improve utility and fairness in liver transplantation. We provide a simple yet powerful solution methodology using decomposition techniques that could be easily implemented as a subroutine or module in any allocation algorithm, possibly leading to SLT policy modifications.

Our fluid model framework is built on the fluid models in Zenios et al., 2000b and Akan et al., 2012; we advanced fluid model analysis by deriving the exact decision rules of the optimal policy in the interior of the state space. Our decomposition results show that the fluid model-optimal policy's decision rules in the interior case are solutions to standard linear programs—this finding significantly reduces the complexity of solving the fluid control problem with ODEs in the constraints (i.e., we removed the ODEs by applying the decomposition technique). It improves solution quality (i.e., we find the exact solution instead of giving a heuristic). The exact decomposed optimal decision rules also imply and corroborate the structural properties found in Akan et al., 2012. We also provide new insights on the impact magnitude of earlier decisions versus later decisions on the values of the cumulative objective. It is worth noting that our explicit solutions illuminate the full potential and inherent properties of the fluid approximation and fluid model-based optimization of an overloaded queueing system.

Finally, our fluid limit decomposition makes it possible to encapsulate the first-order

queueing dynamics as a modular building block that can be added to other analytical frameworks and to incorporate patients' strategic accept/reject choices in fluid models.

We expect that the insights developed here will inspire further detailed numerical analysis (including cost-benefit analysis) - indeed, some are already underway - and foster discussion within the transplant community about further incorporating fairness into their allocation rules and encouraging the wider use of SLT. More broadly, we hope that our methodological contributions advance the ability of operations researchers to study other models of optimal control of multi-class overcrowded queuing systems.

Chapter 3

Multi-Armed Bandits with Endogenous Learning Curves: An Application to Split Liver Transplantation

3.1 Introduction

Experience-based learning is everywhere. For example, young surgeons need to learn difficult medical procedures by performing them, staff in a call center need to handle customer calls to improve their ability to resolve customer issues efficiently and courteously, new franchisees learn to operate smoothly over time. Such human and organizational learning, while necessary and important in the long term, may come with a short-term cost and affect other stakeholders. For example, an inexperienced surgeon may only be able to perform certain surgeries in their learning phase, and may have a lower success rate. As a result, certain patients requiring a more intricate surgery may not be eligible for surgery with the surgeon, and expected outcomes may be worse even if a patient is eligible. Callers to a new customer-service agent may face longer wait times before a call is answered and may find the call unproductive. Customers patronizing a new franchisee restaurant may feel its products or service to be below expectations, or wait times overlong.

At the same time inexperienced surgeons, new call centers, and new franchisees learn by doing/operating, their supervisors learn about their aptitude/potential and may adjust strategies or allocate resources to achieve larger objectives. For example, the organ allocation authority needs to quickly identify and nurture enough young, promis-

ing surgeons to learn sophisticated surgeries to treat patients nationwide into the future. Call center networks need to route customer calls and schedule personnel to avoid significantly increased delays and lower service quality while building their employee base; food franchises need to determine the marketing/operation support for a new franchisee while maintaining the overall quality of the franchise as it expands. Crucial to efficiently learning and identifying the potentials of these surgeons/agents/businesses is to evaluate the results of experience-based learning in its early stages. Further complicating the problem, organizations may also seek to incorporate other customer-centered metrics into their decision-making, such as fairness, variety, and market breadth. We develop a methodology to solve these sorts of dynamic learning problems, focusing on one of the three scenarios described above for illustration purposes: Allocating livers for expanded utilization of split liver transplantation (SLT) in the US. We describe this problem in greater detail below.

SLT is a procedure that potentially saves two lives using one donor liver, by splitting the donor liver into two partial livers and transplanting each of them into a size-appropriate recipient (Emre & Umman, 2011). This is in contrast to traditional liver transplantation, also known as whole liver transplantation (WLT), which transplants one whole liver into a single recipient. SLT thus provides a unique opportunity to save more lives with the existing pool of livers—despite years of effort to increase organ donation in the United States, there remains a grievous shortage: As of July 2023, there are 10215 candidates on the liver waiting list; and the median waiting time on the list before receiving a liver was 1026.7 days. In the year 2022 alone, there were 13179 new additions to the waiting list, while only 9528 liver transplants were performed. However, despite the acute liver shortage, less than 1.5% of medically

splittable livers are actually used for SLT in the US (Perito et al., 2019).

Besides increasing the total number of transplants, SLT can potentially reduce treatment disparities based on size among patients with end-stage liver diseases (ESLD), whose only chance to survive is liver transplantation. Generally speaking, ESLD patients of smaller physical size face longer wait times and overall lower access to transplants, because there are fewer size-appropriate donated livers. (In WLT liver-recipient size matching, a recipient may receive an organ of the relatively same or slightly smaller size.) In fact, in the US, SLT is currently used primarily to increase liver availability for pediatric patients; resulting outcomes of SLT for children are comparable to those of WLT (Hackl et al., 2018). Liver transplant allocation in the US is managed by a central planner—the Organ Procurement and Transplantation Network (OPTN). Their allocation rule can be essentially described as "sickest first": ESLD patients' health conditions evolve dynamically, and their position on the national waitlist, and consequent probability of getting a transplant, change over time (Akan et al., 2012; Emre & Umman, 2011).

Surgeons' aptitudes for performing different types of transplant surgeries are affected by the coordination and expertise of their whole medical team, or even the entire transplant center. In fact, it is not uncommon for the expertise levels of surgeons and medical teams on complicated surgeries to be significantly influenced by the transplant centers (TCs) they belong to. This is because new surgeons are usually matched to TCs based on overall surgical skills, and usually, surgeries are performed by medical teams which involve supporting staff that may assist multiple surgeons in a TC. Moreover, new surgeons likely have the chance to observe, assist, and learn from

experienced surgeons of their TC during actual SLT surgeries, and these hands-on experiences are crucial in achieving their full potential. For all these reasons, TCs report transplant outcomes on the aggregate center level, not on the individual surgeon level. Therefore, we focus on transplant center level proficiency: We divide transplant centers into several classes based on their features, such as surgical experience, performance history, geographical regions, etc. Henceforth, we use "surgeon," "transplant center," "center," and "medical teams" interchangeably.

Like many complicated and potentially risky medical procedures, gaining SLT expertise requires a learning process. This process is not only arduous but also may involve a lower initial transplant success rate, as the medical team acquires skills (Perito et al., 2019). To encourage more transplant teams to learn SLT, and reduce surgical risks, policymakers might consider accommodating their learning, for example by allocating them high-quality organs. To evaluate the benefits of helping TCs acquire the skills needed to perform SLTs while quickly identifying the most suitable medical teams to specialize in SLT and the best surgical combinations, we model organ allocation in a centralized transplantation network using a *multi-armed bandit* (MAB) model. We then develop novel variants of the *upper confidence bound* (UCB) algorithm to find allocation policies that balance the exploitation of existing knowledge and the exploration of surgical combinations that might have high aptitudes but whose full potentials are initially unknown.

Within our MAB model we explicitly incorporate the following features:

- **Endogenous learning curves**: Transplant centers' SLT expertise increases as they accumulate experience. In our MAB formulation, arms' rewards are

parametric functions (with unknown scalar or vector parameters) of previous arm choices, capturing increasing proficiency with practice.

- **Fairness**: A UNOS public comment proposal (OPTN & UNOS, 2016) states that "...increased utilization of split liver transplantation could increase access to transplants," and "The Committee affirms that optimal allocation policies involving whole livers or split liver allografts should reflect a balance between the principles of equity and utility." We propose two fairness notions: best-K probabilistic fairness (BK-fairness) and arbitrary arm fairness (AA-fairness). These notions seek to expand the number of facilities equipped with SLT capabilities, and/or address equity concerns by imposing rules that diminish disparities in access to transplants.

The incorporation of these model features significantly complicates the MAB model; nevertheless, we propose the L-UCB and FL-UCB frameworks, which solve a broad class of MAB problems where endogenous, nonstationary reward curves and fairness constraints exist. We prove that our L-UCB and FL-UCB algorithms achieve the optimal $O(\log t)$ regret, where t denotes the number of transplants, under benign conditions.

We note that our problem could also be captured using a reinforcement learning (RL) model. We choose not to formulate a general RL model because of our problem's special structure: The more SLTs performed by a medical team, the more experienced the medical team becomes. By exploiting this structure, we can use an enhanced MAB with learning curves embedded in its non-stationary rewards to fully characterize the structured RL problem, while maintaining parsimony and tractability.

Our methodology could potentially be applied to help evaluate strategies to increase the proliferation of SLT and other medically-difficult procedures, for example how to effectively and fairly develop a base of skilled practitioners. Moreover, our model and algorithms can be applied to any resource allocation problem where learning exists, including the call center and franchisee examples mentioned previously.

This chapter is organized as follows: Section 2 discusses the literatures relevant to our work. Section 3 introduces the SLT learning problem and the MAB model formulation with learning curves embedded in the arm reward functions. Section 4 describes the L-UCB algorithm and analyzes its regret bound for the MAB models. Section 5 introduces our novel fairness notions, and describes our FL-UCB algorithm with its $O(\log t)$ regret bound. Section 3.6 discusses extensions to our MAB model and summarizes relevant findings. Section 3.7 presents the results of numerical experiments based on real-world SLT data. Section 3.8 summarizes the conclusions and contributions of this chapter and discusses the limitations and potential directions for future work. An appendix, containing more details about the extensions, simulations, and all proofs, can be found in the appendix.

3.2 Literature Review

This work is closely related to six streams of literature: a) exploration and exploitation trade-off; b) dynamic learning; c) organ transplantation; d) MAB with delayed feedback; e) experience-based learning; and f) fairness.

Exploration and exploitation trade-off. A classical model for the exploration-exploitation dilemma used in statistics, artificial intelligence (AI), and MS/OR is the

multi-armed bandit (MAB), first introduced by Thompson, 1933 for clinical trials. In this chapter, we formulate a *stochastic bandit* with parameterized, endogenously non-stationary reward functions. Researchers have also studied contextual bandits, adversarial bandits, and linear bandits extensively (Lattimore & Szepesvári, 2020).

In the vanilla stochastic MAB problem, arm rewards are stationary; however, to model endogenous experience-based learning and its resulting improved proficiency as experience accumulates, we embed a learning curve in each arm's reward function. Specifically, we consider parametric learning curve functions with unknown scalar or vector parameters. Nonstationary rewards, i.e., a reward distribution that can evolve over time, in MABs have primarily been studied when nonstationarity comes from the exogenous environment (Besbes et al., 2019; Cheung et al., 2020; Garivier & Moulines, 2011), making arm rewards independent of policy history. Cheung et al., 2020 also studied endogenous reward nonstationarity using a discrete-time Markov decision process (MDP), where both the discrete reward and discrete state-transition distributions depend (solely) on the current state and action. We consider an infinite-horizon, continuous-time formulation where nonstationarity can be fully characterized by a parametric learning curve. Our formulation of parametric nonstationary rewards is significantly different from existing work, and has advantages in terms of parsimony and extending the upper confidence bound algorithm class.

Dynamic learning. Dynamic learning problems in an endogenously or exogenous changing environment have been studied in different contexts, e.g., online search and consumer lending. In endogenously changing environments, den Boer and Keskin, 2022 studied a dynamic pricing problem where demand is influenced by the current

selling price and also by customers' hidden reference prices that may endogenously evolve over time. The seller needs to learn customers' true reference price through price exploration and balance the tradeoff between demand learning and earning. In exogenously changing environments, Keskin and Zeevi, 2017 studied a dynamic pricing problem where a seller faces an unknown demand model that can exogenously change subject to some finite variation "budget"; their variation metric allows for a broad spectrum of temporal behavior. Keskin and Li, 2021 considered heterogeneous customers and exogenous Markovian market transitions and analyzed a firm's optimal pricing policy and its structural properties. In this chapter, we study an MAB variant where the reward curves following parametric functions and the expected rewards of arms change endogenously as a function of historical pulls. This parsimoniously captures our motivating applications.

Our work is also relevant to recent work on feature-based rewards and high dimensionality in dynamic learning problems (see Section 3.6.2). Ban and Keskin, 2021 proved bounds for expected regret in a personalized demand model with customers' characteristics encoded as a potentially high-dimensional feature vector, where a seller learns the relationship between customer features and product demand through sales observation. Keskin et al., 2023 considers an electric utility company serving retail electricity customers over a discrete time horizon, where the company observes customers' consumption, high-dimensional customer characteristics, and exogenous factors, and dynamically adjusts price at the customer level. They jointly optimized spectral clustering and feature-based pricing and show their proposed policy achieves near-optimal performance.

Organ transplantation. While much work has been done on kidney allocation (Zenios et al., 2003), fewer papers have addressed the allocation problem for livers (Akan et al., 2012; Bertsimas et al., 2020), and those have only studied whole liver allocation. Akan et al., 2012 analytically modeled the liver allocation problem as a fluid model with utilitarian objectives incorporating patients' dynamically-changing MELD/PELD scores. Their work did not consider medical learning, the practice of SLT, or any fairness concerns. Bertsimas et al., 2020 proposed a novel continuous distribution model that balances efficiency and fairness in liver allocation. None of these papers considered SLT or experience-based learning. Our chapter concentrates on the selection of transplant centers, surgical techniques, and livers for specialized procedures in their initial phases of expanding uses, specifically, SLTs.

In the transplantation community, most SLT papers are retrospective reviews, in which transplant centers share their SLT experiences. Other topics covered include ethics (Vulchev et al., 2004), statistical analysis using open data (Perito et al., 2019), and policy guidelines (OPTN & UNOS, 2016). Recent studies show that the outcomes of SLT can be as good as WLT in big TCs, for example, the transplant center at the University of California, San Francisco (UCSF).

MAB with delayed feedback. In many healthcare applications, including organ transplantation, the outcomes may only be observed after some delay (Anderer et al., 2022; Kantidakis et al., 2020). For example, 90-day survival labels are only obtained 90 days after the surgery, and quality-adjusted life years (QALY) may not be fully observed until many years later, but some transplant objectives can be observed right away or within days after transplant, e.g., postoperative outcomes, including

graft function/dysfunction/failure and cellular rejection. There is a stream of research specifically discussing using such early-on intermediate indicators or surrogate outcomes for medical decision-making. For example, Anderer et al., 2022 study algorithms that use surrogate and true outcomes to improve decision-making within a late-phase clinical trial. We adopt a similar approach to extend our base algorithm to accommodate a delay in observing true rewards: We consider using estimated (expected) rewards (based on demographic features and clinical metrics) available immediately after surgery as temporary/surrogate outcomes. When true rewards are observed, the estimates are replaced with the true outcomes. We show in Section 3.6.1 that we obtain the same $O(\log t)$ regret for our problems under mild assumptions.

Delayed feedback arises in multi-armed bandit applications beyond healthcare, such as searching over fast-charging policies for electrochemical batteries to maximize battery lifetime (Grover et al., 2018; Joulani et al., 2013). Joulani et al., 2013 found that the delayed feedback inflates the regret in an additive fashion in stochastic MAB problems, and developed modifications of UCB algorithms. Grover et al., 2018 considered a setting where partial feedback is available (analogous to our surrogate outcomes) and proposed an extension where an agent can control a batch of arms. Chick et al., 2022 studied non-covariate bandits with delay with one arm compared with a standard and shed insights on a unified policy defining the experiment regions and stopping boundaries for sequential sampling, incorporating the size of the delay. Chick et al., 2022 considered multiple arms with delay and took an in-depth look at intriguing issues of randomization and finding useful prior distribution via empirical Bayes methods and pilot data. Alban et al., 2022 studied the sequential allocation of sample observations for personalized treatment strategies, motivated by the design of adaptive clinical

trials that learn the best treatment as a function of patient covariates. Compared with these works, our contributions differ in the following ways: We show that in an expanded class of MAB problems where the expected rewards are endogenous and nonstationary. If we have temporary estimates of the delayed rewards that satisfy mild conditions, we obtain the same optimal regret upper bound scale: $O(\log t)$.

Experience-based learning. For transplant surgeons, many procedures involve the same repetitive tasks; thus it is appropriate to use learning curves with a focus on increases in success rate to represent learning and improvement in performance over time. Several functional forms have been used in the literature to capture human learning, such as S-curves (Sigmoid curves), diminishing-returns curves, and increasing-returns curves. We primarily use the Sigmoid functional form for our SLT numerical study, which nicely captures the features of learning complicated surgeries, such as a slow learning rate at the beginning and stable long-term performance (Le Morvan & Stock, 2005; Pusic et al., 2015). In the call center literature, Arlotto et al., 2014 studied the hiring and retention of heterogeneous agents who learn over time and formulated it as an infinite-armed bandit with switching costs. Arlotto et al., 2014 also computationally investigated families of curves indexed with one random variable and presented the optimal curves. Discussions on learning in other applications are deferred to B.8.

Fairness. Concerns about the fairness of access to medical care and resources have existed for centuries. We study a fairness notion that is not based entirely on a meritocratic basis but on some protected features (e.g., patient physical size, age, geographical region, etc.). Similar to Schumann et al., 2019, we define our notion of

fairness probablistically. However, instead of equal or proportional group probability, we use max-min group probability, a notion adapted from Rawls, 2001, where the arms within the group (corresponding, for example, a specific group of patients) are guaranteed to be selected with no less than certain probabilities. To characterize the efficiency-fairness trade-off, we adapt the *price of fairness* definition from Bertsimas et al., 2011a, that is, the ratio of total reward loss to the optimal total rewards.

3.3 Problem Formulation and Model Setup

In this chapter we focus on the SLT problem where medical teams need to learn SLT by actually performing SLT surgeries on patients. Meanwhile, a central planner learns which combinations of surgical teams, liver types, and recipient types have the highest long-term rewards (i.e., 1-year graft survival). We aim to develop a novel bandit algorithm to accelerate learning the highest full-potential combinations under stochastic (and potentially delayed) observations.

We explicitly model each type of surgery as an "arm" in the vocabulary of bandit problems: Each arm, or surgery, incorporates information about the features of the transplant center(s), the liver(s) to be transplanted, and the patient(s) associated with the surgery.

3.3.1 SLT Learning Problem Formulation

Consider a discrete-time horizon $\mathcal{T} := \{1, 2, \ldots, T\}$. We group transplant centers into classes. Let $\mathcal{D}$ denote the set of transplant center classes comprising centers (with no, little, or some prior SLT experience) yet who are willing to learn and practice SLT. Throughout the planning horizon, each transplant center of class d is capable

of learning and performing SLTs, provided there is a medically appropriate patient and liver pair. We assume that there are significantly more patients than the number of transplant centers and livers. In other words, at least one or one pair of ESLD patients of each defined patient class is always available so that any transplant center can perform any surgery in each period. This assumption is reasonable because the liver waitlists are overloaded (patient arrival rates are greater than liver arrival rates) and a large transplant center typically consists of more than six surgeons and tens of supporting staff so at least one medical team is on duty at any time.

The set of liver types is $\mathcal{L}$, where a liver type is determined by its quality, the geographical location of the donor, and compatibility requirements; let $L = |\mathcal{L}|$. A fixed portion of all deceased-donor livers are eligible for splitting. To focus on the SLT problem, we consider only livers that are medically splittable, assuming non-splittable livers are assigned by another process. Moreover, we assume that all information on transplant centers' experiences prior to the planning horizon, which are publicly available (UNOS, 2020), are summarized in the shapes and structures (intercept, slopes, etc.) of their SLT learning curves. Patients who are medically compatible with a liver of type ℓ might have different health conditions, e.g., some might be critically sick while others are healthier; we denote the set of these patient classes as $\mathcal{P}^\ell$.

When a splittable liver is split, the two partial grafts can be allocated to two recipients in different transplant centers at different times. At each time stamp $t \in \mathcal{T}$ there is exactly one (partial) liver arrival: ℓ_t, which can be transplanted into one recipient. Let P_t be a potential recipient, i.e. $P_t \in \mathcal{P}$. The action space at time t is to choose an allocation, defined as an eligible center-recipient(s)-type pair $(d, P_t) \in \mathcal{D} \times \mathcal{P}$.

We define the arm set of the MAB problem $\mathcal{A}_t$. For presentation clarity, we focus on the case where livers are homogeneous (i.e., $\mathcal{L} = \{\ell\}$, and $\mathcal{A}_t = \mathcal{A} := \mathcal{D} \times \mathcal{P}$); the heterogeneous-liver case is a direct extension and is discussed in Sections B.5.2. Henceforth, we use the term "arm" and "surgery" interchangeably. Each arm is associated with a known accumulated experience level $s_{a,t}, a \in \mathcal{A}$, sometimes written as $s_a(t)$, for $t \in \{1, \ldots, T\}$, and with an unknown aptitude/full potential $\alpha_a \in \mathcal{U}$, i.e., the highest possible mastery level. The experience level indicates the efforts and experience of the surgeon or medical team, corresponding to a value on the x-axis of the learning curves. Depending on $s_{a,t}$, and the learning curve structure, we obtain a $\theta_a(s)$ value, denoting the current mastery or proficiency level of the arm.

Figure 3.1 illustrates three learning curves. The variable s on the x-axis represents learning efforts, or the number of attempts, while θ on the y-axis represents the proficiency or mastery of a specific arm. A higher proficiency level is associated with a higher survival probability or a better expected outcome of a surgery. All curves are Sigmoid curves (the "S"-curves), i.e., $\theta_i(s) := \frac{\alpha_i}{1+\exp(-s+\omega_i)}$, $i = 1, 2, 3$, with $\alpha_1 = 0.5$, $\omega_1 = 1$ (blue), and $\alpha_2 = 0.85$, $\omega_2 = -6$ (orange), and $\alpha_3 = 0.8$, $\omega_2 = -1$ (green). We assume that we know the structures and form of the arms' learning curves (e.g., a Sigmoid curve parameterizing a Bernoulli variable, which represent the values of surgical outcomes). Still, we do not know the parameter α of the curve. This assumption can be relaxed; please refer to Section 3.4.6 for more detail about learning multiple unknown parameters. Specifically, $\omega_i, i = 1, 2, 3$, can be known parameters (describing the existing experience) or unknown in which case they need to be learned along with $\alpha_i, i = 1, 2, 3$ (see Section 3.4.6).

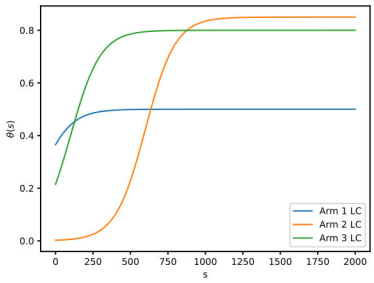


Figure 3.1: An example with three learning curves. All are Sigmoid functions with different full potentials, shape parameters, and starting experience levels.

Let $T_{a,t-1}$, sometimes written as $T_a(t-1)$, denote the number of times that arm a (or a's corresponding surgery) has been chosen (practiced) up to and including time $t-1$ (this may be different from $s_{a,t-1}$ in models with arm correlation, see Section B.5). We define the state of the SLT learning problem as (ℓ_t, S_t) where $\ell_t \in \mathcal{L}$ and $S_t := (T_{a,t-1}, s_{a,t-1})_{a\in\mathcal{A}} \in (\mathbb{N}_+ \times \mathbb{R}_+)^{|A|}$. Let σ_t be the decision rule at time t, i.e. $a_t = \sigma_t(S_t)$. A policy π $(\pi(t))$ is a series of decision rules, i.e. $\pi = \{\sigma_\tau\}_{\tau=1}^{T}$ $(\pi(t) = \{\sigma_\tau\}_{\tau=1}^{t})$. Given $a_t \in \mathcal{A}_{\ell_t}$, we obtain a random reward $r(\ell_t, a_t, S_t)$, e.g., 1-year graft survival. When $\mathcal{L} = \{\ell\}$, for simplicity, we denote $r_{a,s_{a,t}} := r(\ell, a_t, S_t)$. We assume that the reward is a discrete Bernoulli variable, with the mean being the hidden expertise or mastery level of the participating medical team(s) for a certain type of surgery, i.e. $\theta_a(s_{a,t-1})$, where $a \in \mathcal{A}_{\ell_t}$ is the surgery type/arm and $s_{a,t-1}$ is a's experience level prior t.

The objective of the SLT learning problem is to find the policy which maximizes the objective function, e.g., the 1-year graft survival, for large T:

$$\max_{\pi} \ \mathbb{E}\sum_{t=1}^{T} r(\ell_t, a_t^\pi, S_t^\pi). \tag{3.1}$$

3.3.2 The Multi-Armed Bandit Model

Here, we summarize important notation and explain how we map the SLT liver allocation problem to a MAB model with endogenously nonstationary reward curves. At each time t, we choose an arm $a_t \in \mathcal{A}_\ell$ (i.e., allocate a liver for an SLT surgery) and receive a random reward $r(\ell, a_t, S_t)$, that is, the outcome of the SLT surgery. A strategy π is a series of allocation actions or choices of arms; $\pi(t)$ denotes the series of actions from time 1 up to t. We call $\pi(t)$ the policy history at t.

Let $\theta_a^{\pi(t-1)}$ be an SLT arm a's unknown SLT performance level at time t, under a policy history $\pi(t-1)$, and $T_{a,t}$ or $T_a(t)$ denotes the number of times that arm $a \in \mathcal{A}_\ell$ was chosen prior to and including time t. Recall that $s_{a,t-1}$ or $s_a(t-1)$ denotes the experience level of arm a prior to time t. As we assume that arms are independent, $s_{a,t-1} = T_{a,t-1}$, and the hidden performance level of an SLT arm can be rewritten as $\theta_a(T_{a,t-1})$. The outcome of action a is a Bernoulli random variable with mean $\theta_a(T_{a,t-1})$.

3.3.3 Regret

We define the offline policy/the optimal full-information policy π_t^* which achieves the highest cumulative rewards, i.e., $\pi_t^* := \text{argmax}_\pi \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^\pi, S_\tau^\pi)$. To evaluate the utility loss due to lack of information on TC aptitudes where learning curves exist, we use a common objective in the bandit literature — minimizing total expected regret, that is, the expected deficit suffered relative to the optimal full-information

policy. For any fixed turn $t \in \mathcal{T}$, the regret is defined as

$$R_t = \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi_t^*}, S_\tau^{\pi_t^*}) - \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi}, S_\tau^{\pi}). \tag{3.2}$$

When arms are independent of each other and arms' aptitude parameters $\alpha_a, \forall a \in \mathcal{A}$ are known, the optimal policy as t grows large is trivial, always selecting the arm with the highest long-term aptitude; in other words, it always chooses $a^* := \text{argmax}_{a \in \mathcal{A}} \alpha_a$. For any given t, full-information dynamic programming can solve the offline policy, which may not be trivial for small t.

3.4 L-UCB Algorithm and Regret Bounds

In this section we study the MAB problem with learning curves embedded in the reward functions, as shown in Figure 3.2. Each transplant center has an unknown true aptitude that determines its hidden (unobservable) expertise or proficiency level $\theta_a(s_{a,t-1})$ when a center's experience level is $s_{a,t-1}$. The observable variables are its experience level $s_{a,t-1}$ at time t, and past and current outcome variables $r^t := r(\ell, a_t, S_t)$ (e.g., 1-year graft survival). Note that the outcomes are also affected by environmental variables $e_t \in \mathcal{P} \times \mathcal{L}$ (patient health condition, liver quality, etc.); these environmental variables are taken into account in our formulation of MAB.

In Section 3.4.1, we introduce the classical, vanilla UCB algorithm; in Section 3.4.2, we present the L-UCB algorithm; in Section 3.4.3, we prove L-UCB's regret bounds; in Section 3.4.4, we provide a generic method of moments (MoM) approach to construct unbiased L-UCB estimators; in Section 3.4.5, we discuss the use of biased estimators in L-UCB; in Section 3.4.6, we present a general approach using maximum likelihood

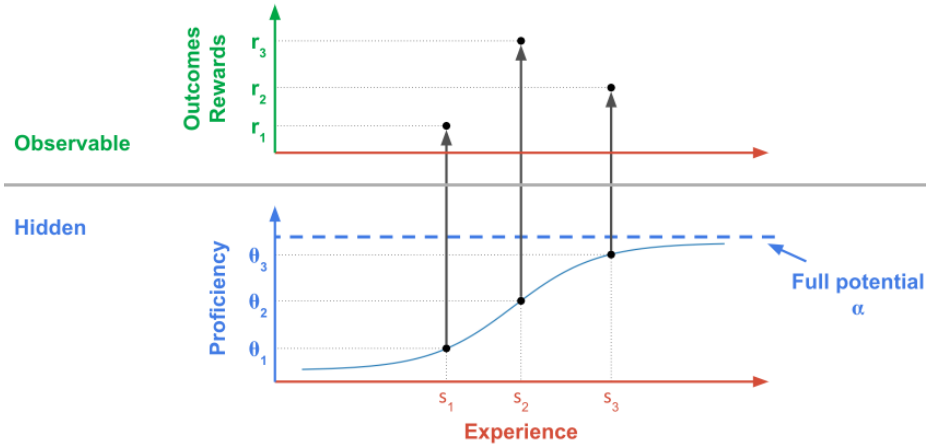


Figure 3.2: A graphical representation of the SLT learning MAB problem. The observable outcome of surgery, r, is a random function of the hidden proficiency level θ. For a specific arm, when its experience with a certain type of SLT surgery is s and s', its hidden proficiency levels would be θ and θ', while the observable (stochastic) outcomes are r and r', respectively.

estimation (MLE) and maximum a posteriori probability (MAP) to construct estimators for α and/or multiple unknown parameters; finally, in Section 3.4.7, we study the scenarios where the parametric form of learning curves may be unknown, and propose nonparametric algorithms within the L-UCB framework.

3.4.1 Upper Confidence Bound (UCB) Algorithms

The *upper confidence bound* (UCB) method is a class of algorithms for the MAB that give an asymptotically optimal solution achieving an $O(\log t)$ regret (Lattimore & Szepesvári, 2020). To illustrate this method, we start with a simplified scenario where $\theta_a(s) = \alpha_a, \forall s, a$; i.e., there is no learning present.

The standard, or vanilla UCB algorithm uses Hoeffding's inequality to derive upper confidence bounds on the unknown aptitudes; these bounds are greater than their *de facto* values with high probabilities. It then selects the arm with the maximal upper

bound. For any surgery a with unknown aptitude α_a that has been chosen n times and yielded random rewards $r_a^{(1)}, \ldots, r_a^{(n)}$, the vanilla UCB uses $\hat{\alpha}_{a,n} := \frac{1}{n}\sum_{i=1}^{n} r_a^{(i)}$ as the estimator of α_a, the empirical or sample mean. Recall that $T_a(t-1)$ denotes the number of times surgery a has been chosen prior to time t. Define the upper bound for the estimate of α_a as

$$B_{a,t,T_a(t-1)} := \hat{\alpha}_{a,T_a(t-1)} + \delta_{a,t,T_a(t-1)}, \quad \text{where} \quad \delta_{a,t,n} := \sqrt{\frac{2\log \eta(t)}{n}}.$$

We choose $\eta(t) = t$ in the vanilla UCB; then the algorithm is formally defined as

$$a_t = \begin{cases} \text{argmax}_a \, B_{a,t,T_a(t-1)} & \text{if } t > |\mathcal{A}| \\ t & \text{if } t \leq |\mathcal{A}| \end{cases} \tag{3.3}$$

3.4.2 The L-UCB Algorithm

Now we describe the L-UCB algorithm for MAB problems with learning curves embedded in the reward functions. Similar to the notation used in the vanilla UCB, we denote by $\hat{\alpha}_{a,n}$ the estimator of α_a after arm a has been chosen n times. But instead of being restricted to the empirical mean, the estimator $\hat{\alpha}_{a,n}$ can be any function of n random rewards $(r_a^{(\tau)})_{\tau=1}^{n}$ and the corresponding n experience levels $(s_{a,\tau})_{\tau=1}^{n}$ to the estimate of the value of α_a, where $r_a^{(\tau)}$ denotes the random reward obtained the τth time arm a was chosen, when the experience level was $s_{a,\tau}$. Thus in the L-UCB algorithm $\hat{\alpha}_{a,n}$ can utilize a broad class of mapping functions and estimators, including the empirical mean. Some other potential estimators are method of moments (MoM) estimators, maximum likelihood estimation (MLE) estimators, and maximum a pos-

teriori probability (MAP) estimators, which we discuss more in Section 3.4.4 $\sim$ 3.4.6. The estimator $\hat{\alpha}_{a,n}$ in L-UCB takes n additional arguments—the experience levels—compared to the estimator used in the vanilla UCB. In this section, the experience level $s_{a,\tau} = \tau - 1$, for all $\tau = 1, 2, \ldots, n$, i.e., the experience level is equivalent to the number of historical pulls. In Section 3.6.2 we discuss an extension in which arms are correlated; in such scenarios, $s_{a,\tau-1}$ and $T_a(\tau - 1)$ may not be equivalent. Note that the estimator $\hat{\alpha}_{a,n}$ can incorporate the parametric forms of learning curves, if such information is available. Alternatively, $\hat{\alpha}_{a,n}$ can be chosen without any knowledge of the learning curves' parametric form; we discuss nonparametric methods in Section 3.4.7.

Because $\hat{\alpha}$ might be a more complicated function than the empirical mean, we define the following properties over this function class.

Definition 3.4.1 (Bias of an estimator)**.** The bias of an estimator $\hat{\alpha}$ of parameter α is the difference between the expected value of the estimator and the true value of α; that is, $\mathbb{E}[\hat{\alpha}] - \alpha$.

Definition 3.4.2 (Unbiased estimator)**.** Estimator $\hat{\alpha}$ is unbiased if its bias is zero.

It should be noted that many widely used estimators are biased; for example, the MLE estimator of the Gaussian variance is biased.

Definition 3.4.3 (Gap of a sub-optimal arm a)**.** The gap of arm a is $\Delta_a := \max_{a' \in \mathcal{A}} \alpha_{a'} - \alpha_a$.

Remark: In our SLT learning problem, because all α_a's take values from a bounded set $\mathcal{U} := [0, 1]$, the sub-optimal gaps are also bounded throughout the planning horizon.

Definition 3.4.4 (Per-coordinate difference bound)**.** Suppose $\mathcal{X}$ is a sample space and $\varphi : \mathcal{X}^n \to \mathbb{R}$. If there exists $w_1, \ldots, w_n \geq 0$ such that

$$\sup_{x_1,\ldots,x_n,x_i^{'}\in\mathcal{X}} |\varphi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - \varphi(x_1, \ldots, x_{i-1}, x_i^{'}, x_{i+1}, \ldots, x_n)| \leq w_i \tag{3.4}$$

for $i \in \{1, 2, \ldots, n\}$, then $(w_i)_{i=1}^n$ is said to be a per-coordinate difference bound for φ.

Equation (3.4) states that any modification of the value of the ith coordinate changes the value of φ by at most w_i whatever values the other coordinates take. This φ can be any function including any aforementioned estimator $\hat{\alpha}_{a,n}$. For any φ, the per-coordinate difference bound doesn't have an upper bound (as $w_i = \infty$ satisfies (3.4)) but does have an infimum which varies with φ. When φ is independent of the ith coordinate, that is, changing the value of the ith coordinate solely never changes the value of φ, the infimum of w_i is zero. In this case, the ith coordinate is obsolete.

Definition 3.4.5 (Per-coordinate difference bound parameter)**.** If mapping function $\varphi : \mathcal{X}^n \to \mathbb{R}$ has per-coordinate difference bound $w_1, \ldots, w_n$, then we say φ has a per-coordinate difference bound with parameter $C_n^w := \frac{1}{n\sum_{i=1}^n w_i^2}$.

Remark: For any function φ, C_n^w is not unique and doesn't have a positive lower bound, i.e., C_n^w can be arbitrarily small, because $w_1 = w_2 = \cdots = w_n = +\infty$ is a per-coordinate difference bound of φ with parameter $C_n^w = 0$. However, as w_i has an infimum, C_n^w does have a supremum, which depends on the nature of φ.

Now we present pseudo-code of the L-UCB algorithm. Assume $\hat{\alpha}_{a,n}$, the estimator of α after arm a has been chosen n times, has a per-coordinate upper bound with parameter $C^w_{a,n}(> 0)$. Similar to the vanilla UCB, we define

$$\delta_{a,t,n} := \sqrt{\frac{2\log t}{nC^w_{a,n}}} \quad \text{and} \quad B_{a,t,T_a(t-1)} := \hat{\alpha}_{a,T_a(t-1)} + \delta_{a,t,T_a(t-1)}.$$

For simplicity, $B_{a,T_a(t-1)} := B_{a,t,T_a(t-1)}$. Denote by $b_{a,n}$ the bias of $\hat{\alpha}_{a,n}$, and assume there exists an $m_a \in \mathcal{T}$ for arm a such that $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC^w_{a,n}}}$ for all $n \geq m_a$ (we discuss m_a below).

Procedure 1: L-UCB Algorithm Pseudo Code

1: **Initialization:** Select each arm a m_a times

2: **Update statistic:** $B_{a,T_a(t-1)} \leftarrow \hat{\alpha}_{a,T_a(t-1)} + \sqrt{\frac{2\log t}{C^w_{a,T_a(t-1)}T_a(t-1)}}, \quad \forall a \in \mathcal{A}$

3: **Select arm:** $a_t \leftarrow \text{argmax}_a\, B_{a,t,T_a(t-1)}$, and update $T_{a_t,t}$

4: **Increment** t and **Go to Step 2**

Now, we discuss the behaviors of m_a when the bias $|b_{a,n}|$ shrinks at different rates and when $C^w_{a,n}$ has a positive lower bound, i.e. $C^w_a := \inf_{n\in\mathbb{N}_+} C^w_{a,n} > 0$. When estimator $\hat{\alpha}_{a,n}$ is unbiased for all n, we select arm a exactly once in the initialization, just as the vanilla UCB. When the bias $|b_{a,n}|$ decays at $O\left(\sqrt{\frac{1}{n}}\right)$ rate, i.e. there exists a constant K^b_a such that $|b_{a,n}| \leq K^b_a n^{-1/2}$ for any n, we can set $m_a = \lceil \exp(50(K^b_a)^2 C^w_a) \rceil$. If the bias doesn't decay we need to choose an alternative estimator with zero or decaying bias, unless the bias is known and can be corrected.

3.4.3 L-UCB Regret Bounds

In this subsection we derive the upper bound on the regret of the L-UCB algorithm.

Proposition 3.4.1 (Upper and lower bounds of the supremum of the per-coordinate difference bound)**.** Suppose $\mathcal{X}$ is a sample space, $\varphi : \mathcal{X}^n \to \mathbb{R}$ is a function whose image set is $[0, 1]$, and $\{\omega_i\}_{i=1}^n$ is any per-coordinate bound defined over $\mathcal{X}$. Then the supremum of the per-coordinate difference bound of φ over all possible values of $\omega_1, \ldots, \omega_n$, denoted by $C_n^* := \sup_{\omega_1,\ldots,\omega_n} C_n^w$, satisfies $\frac{1}{n^2} \leq C_n^* \leq 1$.

To prove the left inequality we note that $w_1 = 1, \ldots, w_n = 1$ is a per-coordinate difference bound of φ with parameter $C_n^w = \frac{1}{n^2}$. Because C_n^* is the supremum of all feasible C_n^w, we know $C^* \geq \frac{1}{n^2}$. The proof of the right inequality uses Chebyshev's sum inequality (Hardy et al., 1952). Please refer to Section B.1.2 for proof details.

Lemma 3.4.1 (Bounded Difference Inequality)**.** Suppose $\mathcal{X}$ is a sample space, and function $\varphi : \mathcal{X}^n \mapsto \mathbb{R}$ has per-coordinate difference bound $w_1, \ldots, w_n$ with parameter C_n^w, i.e. $w_1, \ldots, w_n > 0$ satisfy $C_n^w = \frac{1}{n\sum_{i=1}^n w_i^2}$ and

$$\sup_{x_1,\ldots,x_n,x_i^{'}} |\varphi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - \varphi(x_1, \ldots, x_{i-1}, x_i^{'}, x_{i+1}, \ldots, x_n)| \leq w_i \tag{3.5}$$

for $i \in \{1, \ldots, n\}$. Then,

$$P\left(\varphi - \mathbb{E}[\varphi] > \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n w_i^2}\right) = \exp\left(-2nC_n^w\varepsilon^2\right), \tag{3.6}$$

$$P\left(\varphi - \mathbb{E}[\varphi] < -\varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n w_i^2}\right) = \exp\left(-2nC_n^w\varepsilon^2\right). \tag{3.7}$$

For proof details, readers are referred to McDiarmid, 1998 (see their Theorems 1 and 2 and references therein). Lemma 3.4.1 states that the probability of the value of φ

being close to its expectation is higher when φ is less sensitive to its arguments, i.e., the upper bounds of the above probabilities are smaller if the w_i's are smaller and C_n^w is larger.

When φ is the empirical mean used in the vanilla UCB, we have $w_a^{(i)} = \frac{1}{n}$ for any i. Because the empirical mean achieves the maximum $C_n^w = 1$, we take it as the standard and compare other functions' behaviors with it. In this sense, nC_n^w can be thought of as a reduced number of samples: For the empirical mean, nC_n^w is precisely n, which is the number of samples governing the rate of decay of the bound. When the estimator φ has larger $w_a^{(i)}$s, we have C_n^w less than 1, and then the probabilistic bounds on $\varphi - \mathbb{E}[\varphi]$ are as tight as the corresponding bounds of the empirical mean with $nC_n^w < n$ samples, i.e., the estimator with C_n^w achieves the same accuracy with fewer samples compared to the empirical mean estimator, as $C_n^w < 1$.

For example, suppose $r_t \sim \text{Bernoulli}(\frac{\alpha t}{t+1})$ for $t \in \{1, \dots, T\}$, the higher the aptitude α and/or the proficiency level $s_t := t$, the more likely that a surgery is successful. The value of α is hidden, but t is known in each round. The estimator $\hat{\alpha} := \frac{1}{T}\sum_{t=1}^{T} \frac{t+1}{t} r_t$ is an unbiased estimator of α. This $\hat{\alpha}$ can be thought of as a weighted empirical mean (although the weights don't sum to one), so, similar to the empirical mean, this $\hat{\alpha}$ has per-coordinate difference bound $w_1 = \frac{1}{T}, \dots, w_t = \frac{t+1}{tT}, \dots, w_T = \frac{T+1}{T^2}$ with parameter $C_T^w = \frac{T}{T+2\sum_{t=1}^{T} t^{-1} + \sum_{t=1}^{T} t^{-2}}$; this C_T^w is less than 1 at any finite T and approaches 1 as T approaches infinity.

Theorem 3.4.1. Denote the reward of choosing arm a for the nth time by $r_a^{(n)}$. Suppose $r_a^{(1)}, r_a^{(2)}, \dots$ are independent of each other conditioned on the latent aptitude α_a. For each $n \in \mathcal{T}$, suppose estimator $\hat{\alpha}_{a,n}$ has a per-coordinate difference bound

$w_{a,n}^{(1)}, \ldots, w_{a,n}^{(n)}$ with parameter $C_{a,n}^{w}$, i.e. $w_{a,n}^{(1)}, \ldots, w_{a,n}^{(n)} \in \mathbb{R}_+$ satisfy $C_{a,n}^{w} := \frac{1}{n \sum_{i=1}^{n} (w_{a,n}^{(i)})^2}$ and

$$\sup_{r_a^{(1)}, \ldots, r_a^{(n)}, r'} |\varphi(r_a^{(1)}, r_a^{(2)}, \ldots, r_a^{(i-1)}, r_a^{(i)}, r_a^{(i+1)}, \ldots, r_a^{(n)}) - \varphi(r_a^{(1)}, r_a^{(2)}, \ldots, r_a^{(i-1)}, r', r_a^{(i+1)}, \ldots, r_a^{(n)})| \leq w_{a,n}^{(i)}, \quad \forall a \in \mathcal{A}, i \in \{1, \ldots, n\} \tag{3.8}$$

Let $t \in \mathcal{T}$ be any timestamp. When $C_{a,n}^{w}$ has a positive lower bound, i.e. $C_a^w := \inf_{n \in \mathbb{N}_+} C_{a,n}^{w} > 0$, and when the bias of $\hat{\alpha}_{a,n}$ satisfies $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2 \log n}{n C_a^w}}$, each sub-optimal arm is pulled in expectation at most

$$\mathbb{E}[T_a(t)] \leq \frac{8 \log t}{C_a^w \Delta_a^2} + 2\zeta(1.24) \tag{3.9}$$

times, where $\zeta(1.24) \approx 4.76$, and $\zeta(s)$ is the Riemann zeta function, i.e. $\zeta(s) = \sum_{i=1}^{\infty} i^{-s}$.

The expected cumulative regret of the L-UCB algorithm is bounded by

$$\mathbb{E}[R(t)] \leq \sum_{a \neq a^*} (\bar{r}_{a^*} - \underline{r}_a) \left(\frac{8 \log t}{C_a^{\omega} \Delta_a^2} + 2\zeta(1.24) \right). \tag{3.10}$$

Our proof adapts some techniques from the proof of bounds for vanilla UCB algorithms, but our results are applicable to a more general class of bandits. The primary differences/improvements of our result are: a) our regret bounds apply in a broader class of UCB algorithms that use any estimators φ that satisfy certain benign criteria in the L-UCB algorithms; b) our proof allows these estimators to be biased, up

to $\frac{1}{10}\sqrt{\frac{\log nC^w_{a,n}}{n}}$, where n is the sample size and is adequately large. Taken together, these innovations significantly expand the scope of MAB regret bounds, including those with embedded learning curves and a broad class of estimators; examples and discussions in Section 3.4.4 illustrate these benefits.

3.4.4 A Generic Method of Moment (MoM) Estimator: An Explicit Formula

For learning curves that satisfy the following $\theta(s) = \alpha g_\omega(s) + f(s)$, e.g., the learning curves in Example 3.4.1 and Section 3.7, a generic Method of Moments (MoM) estimator can be $\hat{\alpha}^{MoM}_n = \frac{1}{n}\sum_{s=1,g_\omega(s)\neq 0}^{n} \frac{r^{(s)}-f(s)}{g_\omega(s)}$. (We assume not all $g_\omega(s) = 0$; if so, $\alpha^{MoM}_n = 0$.) Note that α^{MoM} is unbiased, because $\mathbb{E}\alpha^{MoM}_n = \mathbb{E}\frac{1}{n}\sum_{s=1}^{n}\frac{r^{(s)}-f(s)}{g_\omega(s)} = \frac{1}{n}\sum_{s=1}^{n}\frac{\mathbb{E}r^{(s)}-f(s)}{g_\omega(s)} = \frac{1}{n}\sum_{s=1}^{n}\frac{\alpha g_\omega(s)}{g_\omega(s)} = \alpha$, assuming $g_\omega(s) \neq 0$. (If for certain s', $g_\omega(s') = 0$, we drop s' when taking the average.) $C^{w,MoM}_n = \frac{1}{n\sum_{s=1}^{n}(1/g_\omega(s))^2}$. α^{MoM}, as $r^{(s)} \in \{0,1\}$. In cases where ω is unknown (see Example 3.4.3 in Section 3.4.6), we can use estimates $\hat{\omega}_s$ to replace ω, i.e. $C^{w,MoM}_n = \frac{1}{n\sum_{s=1}^{n}(1/g_{\hat{\omega}_s}(s))^2}$.

Example 3.4.1 illustrates constructing an MoM estimator and applying L-UCB.

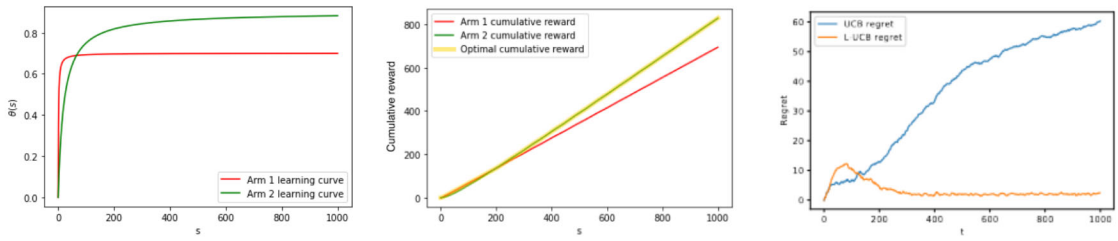
Example 3.4.1 (Incorporating information about the learning curve)**.** Consider a bandit with two independent arms, whose reward curves are illustrated in Figure 3.3a. Arm 1 has a learning curve $\theta_1(\alpha_1, s) = \alpha_1\frac{s}{s+1}$ where α_1's unknown true value is 0.7, while arm 2's learning curve is $\theta_2(\alpha_2, s) = \alpha_2\frac{s}{s+20}$ while α_2's unknown true value is 0.9. Suppose the random outcome $r^{(i)}_{a,s}$ is a Bernoulli variable with parameter $\theta_a(\alpha_a, s)$. We use the unbiased MLE estimators $\hat{\alpha}_{1,n} = \frac{1}{n}\sum_{i=1}^{n}\frac{i+1}{i}r^{(i)}_{1,i}$ and $\hat{\alpha}_{2,n} = \frac{1}{n}\sum_{i=1}^{n}\frac{i+20}{i}r^{(i)}_{2,i}$ in the L-UCB algorithm. The estimator for the vanilla UCB is $\hat{\alpha}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} r^{(i)}_{1,i}$.

In Figure 3.3c, it is clear that the L-UCB algorithm has significantly lower numerical regret when t is large enough (i.e., $t > 194$) because L-UCB incorporates the learning curve information in the early stages to identify the "best" arm in the long-term more efficiently.

It might be counterintuitive that the L-UCB's regret curve increases before decreasing; however, this is possible when the offline policy (i.e., the optimal policy solved by a clairvoyant who knows all parameters) is nonstationary in t. In Example 3.4.1, the "best" arm to play is dependent on the time horizon: If we only consider $t \leq 194$, the "optimal" strategy is always pulling arm 1, but in the long term ($t \geq 195$), the "optimal" policy is always pulling arm 2. The optimal offline policy may not be trivial for any given t, and in general, can be tricky to solve and sometimes intractable. The non-asymptotic regime is important for general MAB problems (Garivier et al., 2019). Nevertheless, in Example 3.4.1 and the SLT problem, we focus on identifying the best long-term arm, which is a special scenario in dynamic learning in which constantly pulling the arm with the highest full potential (α) is the optimal policy asymptotically.

For smaller t the vanilla UCB does the "correct" thing by playing the (temporally) more valuable arm 1. The "optimal" cumulative reward for each t in Figure 3.3b is the optimal cumulative reward obtained by a clairvoyant who knows the true parameters (including α_1, α_2) of all arms' learning curves. Because our L-UCB policy is designed to identify the best long-term arm (i.e., arm 2), in the short term ($t \leq 194$) where the short-term "optimal" policy is pulling arm 1, L-UCB may incur more temporal regret, as seen in Figure 3.3c.

It is possible in applications where learning is present that the vanilla UCB, without



(a) Learning curves **(b)** Expected cum. rewards **(c)** Regret comparison

Figure 3.3: Illustrations for Example 3.4.1. The regret results shown are averaged over 20 instances.

information on learning curves, may still result in $O(\log t)$ regret. However, this is typically not the case in general dynamic learning problems. Performances of the UCB and L-UCB are more extensively compared numerically in Sections 3.4.6 $\sim$ 3.4.7 and in Section 3.7.

3.4.5 Biased Estimators

We further illustrate the benefit of Theorem 3.4.1 by allowing the estimators to be biased.

Example 3.4.2 (Estimator bias)**.** Consider a bandit problem where rewards $(r_a^{(i)})_{i=1}^n$ are i.i.d. Bernoulli random variables with parameter p. To reduce the estimator's variance, people often use a MAP estimator with a Beta prior. Mathematically, the MAP estimator with prior $\text{Beta}(\alpha, \beta)$ is $\hat{\alpha}_n := \frac{h+\alpha-1}{n+\alpha+\beta-2}$, where h is the number of ones in rewards. With n samples, the bias of this MAP estimator is $\frac{(1-p)(\alpha-1)-p(\beta-1)}{n+\alpha+\beta-2}$, which is nonzero for most combinations of α, β, and p. As we obtain more samples, i.e., as n increases, the bias decays at $O(\frac{1}{n})$ rate. Hence, Theorem 3.4.1 is applicable to this case and guarantees an $O(\log t)$ regret. In contrast, the vanilla UCB cannot be guaranteed to work with this estimator.

This flexibility with respect to estimators yields one further advantage of the L-UCB algorithm: The vanilla UCB cannot guarantee identifying the "best" arm when the empirical mean is not an appropriate estimator for the metric of interest, for example, the variance or standard deviation of a random variable. In contrast, in the L-UCB algorithm, we can use any estimator, and if the premises of Theorem 3.4.1 are met, we immediately have that the regret is bounded by $O(\log t)$, thus providing much greater freedom in the choice of metric.

In cases where the bias is initially large but decays quickly, i.e., $|b_{a,n}| \leq C_a^b\sqrt{\frac{1}{n}}$ for some large constant $C_a^b \in \mathbb{R}_+$, the bounds on $T_a(t)$ and $R(t)$ in the theorem may not hold for small t, because the bias condition $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$ may not hold for $n = t$ in this case. Nevertheless, these bounds hold for any adequately large t. Because Theorem 3.4.1 is intended to show the scale of how many times we choose sub-optimal arms and the scale of the regret, we omit discussions of issues around these cases, such as the minimum t where the bounds hold for a given C_a^b.

In Theorem 3.4.1's premise, we assumed $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$. The biases of many standard estimators, e.g., the MLE estimator of logistic regression, is $O(\frac{1}{n})$. Thus, the premise of Theorem 3.4.1 holds for most common estimators (Lehmann & Casella, 2006). Note that in rare cases, verifying the bias conditions for some algorithm instances within our proposed L-UCB framework may be nontrivial; one may skip verifying the bias conditions, apply the L-UCB algorithms and see if they have logarithmic regrets empirically. See Section 3.4.6 for a detailed discussion about ways to verify the bias conditions for L-UCB with MLE and MAP.

3.4.6 MLE and MAP for Estimating Unknown Vector Parameters

When the parametric forms of the learning curves are known, but one or more parameters are unknown, a systematic approach for finding unknown parameters is to use the MLE or MAP. So far, we have illustrated several examples where we obtain explicit formulas for $\hat{\alpha}^{MoM}$. More generally, one may write down the likelihood function (for MLE) or posterior probability (for MAP) and apply optimization algorithms, such as gradient descent (Goodfellow et al., 2016), to get point estimates.

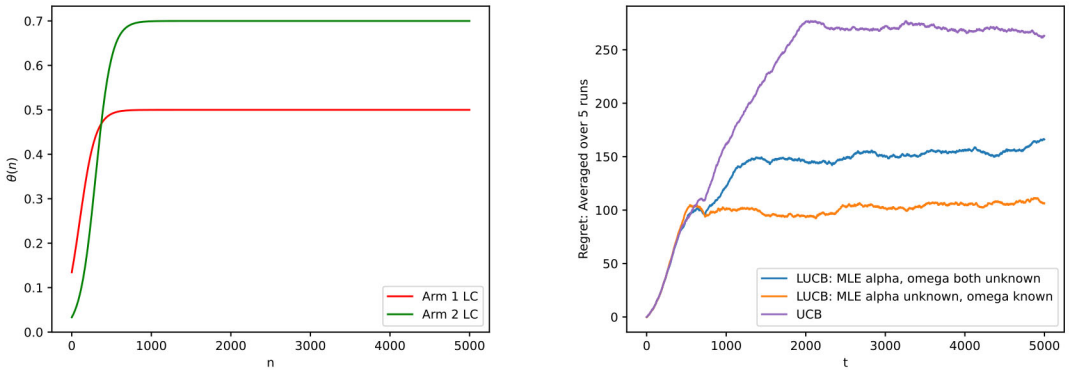
For general parametric learning curves, standard results for the MLE imply that it will satisfy the bias condition (assuming typical identification and regulatory conditions); see section 6.5 of Lehmann and Casella, 2006. We may also numerically verify the bias condition by taking the log scale in both the number of pulls (n) and the absolute values of empirical biases $|b_\alpha|$ and $|b_\omega|$ and fit the curves using linear regression, assuming the empirical biases are observable. The bias condition in Theorem 3.4.1, i.e., $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$, is equivalent to $\log|b_{a,n}| \leq -0.5\log n + \log\frac{1}{10}\sqrt{\frac{2\log n}{C_a^w}}$. A sufficient condition of the bias condition is that the slope of the fitted line is lower than -0.5. Note that the result of empirical verification of the bias conditions is instance specific.

In Example 3.4.3, we show the steps of estimating α and ω simultaneously using MLE and compare: (i) the regret of L-UCB with MLE estimators for both α and ω, (ii) UCB, and (iii) L-UCB with an MLE estimator for α and *known* ω. Comparing L-UCB estimating two unknown parameters with L-UCB with only α unknown shows that knowing ω reduces regret.

Example 3.4.3 (Estimate multiple unknown parameters.)**.** Consider a 2-armed ban-

dit with two "S"-shaped learning curves embedded in the reward function, respectively. Both learning curves share the parametric form: $\theta_{\alpha,\omega}(n) = \frac{\alpha}{1+\exp(-0.01n-\omega)}$. The true parameters for this example are $\alpha_1 = 0.5,\ \omega_1 = -2$ and $\alpha_2 = 0.7,\ \omega_2 = -4$.

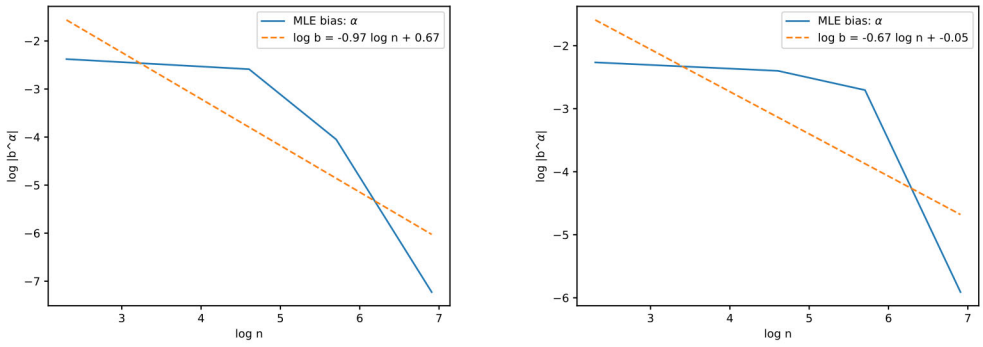
In Figure 3.5 and Figure B.1, we empirically verify that the MLE estimators' biases of α_n^{MLE} and ω_n^{MLE} are both $o\left(\sqrt{\frac{\log n}{n}}\right)$. Specifically, the biases of $\alpha_{1,n}^{MLE}$ and $\omega_{1,n}^{MLE}$ are $O(n^{-0.97})$ and $O(n^{-1.28})$, and the biases of $\alpha_{2,n}^{MLE}$ and $\omega_{2,n}^{MLE}$ are $O(n^{-0.67})$ and $O(n^{-1.80})$. Therefore, MLE estimators of the parameters meet the bias conditions in Theorem 3.4.1. We show the empirical verification figures for $\hat{\alpha}_{i,n}^{MLE},\ i = 1, 2$ in Figure 3.5; figures illustrating bias scales for $\hat{\omega}_{i,n}^{MLE}$ are deferred to the appendix.



(a) Learning curves **(b)** Regret comparison

Figure 3.4: The two learning curves (left). On the right, we compare the regrets (averaged over five runs) of the L-UCB with MLE estimators where both parameters are unknown (blue), L-UCB with $\hat{\alpha}^{MLE}$ where ω is known (orange), and vanilla UCB (purple). The L-UCB: MLE with only α unknown plateaus to regret of approximately 100 after about 1000 trials, with α and ω unknown plateaus to regret of approximately 150 after 1200 trials, and the vanilla UCB plateaus to approximately 250 after about 2000 trials. Thus we see the benefit of utilizing the learning curves' parametric forms and knowing ω, respectively.

Our L-UCB algorithm for the SLT problem only uses $\hat{\alpha}_{a,n}$'s for arm selection, while estimates of ω help construct the $\hat{\alpha}_{a,n}$'s and $C_{a,n}^{w}$'s and verify $\hat{\alpha}_{a,n}$'s bias conditions



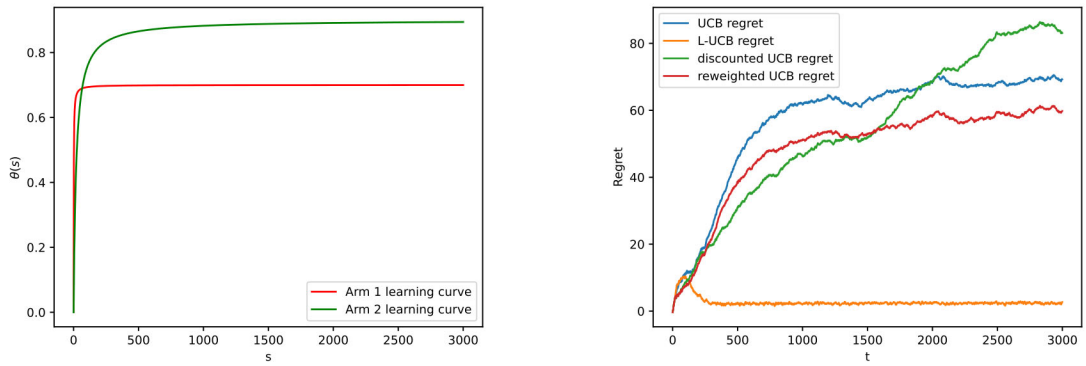
(a) The bias of $\alpha_{1,n}^{MLE}$ is $O(n^{-0.97})$ **(b)** The bias of $\alpha_{2,n}^{MLE}$ is $O(n^{-0.67})$

Figure 3.5: Verifying the bias scales of $\alpha_{1,n}^{MLE}$ and $\alpha_{2,n}^{MLE}$, MLE estimators for arm 1 and arm 2's learning curves. The bias scales are both $o\left(\sqrt{\frac{\log n}{n}}\right)$, satisfying the bias condition in Theorem 3.4.1.

for Theorem 3.4.1 to hold. In general, dynamic learning algorithms may use vector parameter estimates for arm selection; in such cases, our Theorem 3.4.1 may be applied element-wise to obtain theoretical bounds. In rare scenarios when verifying Theorem 3.4.1's premises is tricky, L-UCB with MLE or MAP is still applicable and can provide guidance for finding the optimal online policy.

3.4.7 L-UCB with Unknown Learning Curves

In order to apply MLE and MAP estimators, we assumed the parametric forms of the learning curves are known. When such information is not available, our L-UCB method can still be applied with an estimator $\hat{\alpha}$ that does not exploit the parametric form. For example, the vanilla UCB (a special case of L-UCB) is a nonparametric method, as it does not use any knowledge about the parametric form (or even that the expected rewards are non-stationary). There are other nonparametric estimators that adapt more quickly to nonstationary expected rewards: Instead of using an empirical



(a) Two learning curves

(b) Regret: L-UCB vs. nonparametrics

Figure 3.6: Same numerical setup as Example 3.4.1. The regret results shown are averaged over 20 instances. L-UCB obtains the lowest long-term regret, and reweighted UCB performs better than UCB in this instance.

mean where all historical observations are given the same weight $\frac{1}{n}$, we can assign more recent observations greater weight. For example, we may discount past observations, i.e., $\hat{\alpha}_{a,n}^{disc} := (\sum_{s=1}^{n} \delta_a^{n-s} r_a^{(s)})/\frac{1-\delta_a^n}{1-\delta_a}$, where $\delta_a \in (0,1)$. We may also make the weights a function of n, e.g., $\hat{\alpha}_{a,n}^{rew} := (\sum_{s=1}^{n} sr_a^{(s)})/\frac{n(n+1)}{2}$. Similar ideas have been studied in Garivier and Moulines, 2011, see their D-UCB and SW-UCB, for example. In Section B.8 we discuss the differentiation.

Figure 3.6b shows a comparison between L-UCB with MLE, vanilla UCB, discounted UCB ($\hat{\alpha}_{a,n}^{disc} := (\sum_{s=1}^{n} \delta_a^{n-s} r_a^{(s)})/\frac{1-\delta_a^n}{1-\delta_a}$ and $\delta = 0.9$ in this example), and reweighted UCB ($\hat{\alpha}_{a,n}^{rew} := (\sum_{s=1}^{n} sr_a^{(s)})/\frac{n(n+1)}{2}$); the latter two are special cases of L-UCB where the nonparametric estimators $\hat{\alpha}_{a,n}^{i}$ ($i = \{disc, rew\}$) do not utilize the parametric form of learning curves. The numerical setup is the same as those in Example 3.4.1 except that in Figure 3.6, we show a wider range of t.

Discounted UCB and reweighted UCB may be particularly useful when we know

the expected rewards are nonstationary because they put more weight on recent observations. Reweighted UCB assigns recent observations more weight than UCB, therefore may respond more quickly to the nonstationary environment. Interestingly, discounted UCB has lower regret than reweighted UCB initially ($t \leq 1500$), but accumulates more regret than all other three algorithms once $t > 1500$ before it shows signs of potentially converging around $t = 3000$. This could result from assigning too much, constant weight to recent observations and discounting the past too aggressively, therefore, not efficiently using all the historical data (e.g., the weight assigned to the oldest observation is δ^{n-1}, which decays fast and can become very close to 0). Reweighted UCB is more sample efficient because all the sample weights are between $\left[\frac{2}{n(n+1)}, \frac{2}{n+1}\right]$. All of the nonparametric UCB algorithms perform much worse than the L-UCB, again because the latter exploits the parametric forms of the reward curves.

3.5 Fairness and the FL-UCB Algorithm

In this section we introduce probabilistic fairness definitions within the SLT context. We add constraints that require specific arms to be chosen with no less than certain predefined probabilities; one can interpret probabilistic fairness as long-term average max-min fairness. By properly configuring a set of fair probabilities, we guarantee that donor livers are equitably distributed to a broader range of recipients, rather than being offered continuously to a very narrow group.

Our first type of probabilistic fairness we call *best-K θ-fairness*, or *BK-fairness*, where we require the best $K(\leq |\mathcal{A}|)$ arms be chosen with probability greater than or equal to a vector of predefined levels, θ^{BK}. Note that the choices of θ^{BK} are constrained to render the BK-fairness concept well-defined, i.e. $|\theta^{BK}|_1 \leq 1$ and $\theta^{BK} \geq 0$, where $|X|_1$

is the $L1$ norm of vector X.

Definition 3.5.1 (Best-K θ-Fairness/BK-fairness)**.** Any a in the set of the best-K arms, $\mathcal{A}^{BK}$, has to be chosen with probability no less than $\theta_a^{BK} - \epsilon$ when $t \to \infty$, for any $\epsilon \in \left(0, \min_{a \in \mathcal{A}_{BK}}\{\theta_a^{BK}\}\right)$, where $\sum_{a \in \mathcal{A}_{BK}} \theta_a^{BK} \leq 1$ and $\theta^{BK} \geq 0$.

We are interested in BK-fairness because more widespread use of SLT could bring benefits in practice. For example, if more transplant centers are capable of performing SLT, it could be easier to schedule surgeries and potentially facilitate logistics and reduce organ wastage.

The second type of fairness we define is *arbitrary arm fairness*, or *AA-Fairness*, which prioritizes a set of arbitrary arms, independent of surgeons' aptitudes or expertise. This could ensure that certain populations have access to organs, even if their outcomes are not among the K-best.

Definition 3.5.2 (Arbitrary-Arm θ-Fairness)**.** For a set of arbitrarily-selected arms, $\mathcal{A}_A$, the vector of probabilities of being chosen is no less than $\theta^A \in [0,1]^{|\mathcal{A}_A|}$, where $\left|\theta^A\right|_1 \leq 1$.

3.5.1 The FL-UCB Algorithm

We define a linear optimization program, FL-LP, as follows ($\mathcal{A}, \mathcal{A}_{BK}, \theta^{BK}, \mathcal{A}_A, \theta^A$ are inputs):

$$\max \quad \sum_{a \in \mathcal{A}} \alpha_a z_a \tag{3.11}$$

$$s.t. \quad z_a \geq \theta_a^A \qquad \forall a \in \mathcal{A}_A \tag{3.12}$$

$$z_a \geq \theta_a^{BK} \qquad \forall a \in \mathcal{A}_{BK} \tag{3.13}$$

$$\sum_{a \in \mathcal{A}} z_a = 1 \tag{3.14}$$

$$z_a \geq 0 \qquad \forall a \in \mathcal{A} \tag{3.15}$$

The solution to LP (3.11) - (3.15), z^*, gives the true optimal fair policy in an offline setting. In an online setting where α is not known, we use $B_{a,t,T_a(t-1)}$ instead of α_a, replacing (3.11) with

$$\max \quad \sum_{a \in \mathcal{A}} B_{a,T_a(t-1)} z_a. \tag{3.16}$$

Procedure 2: The FL-UCB Algorithm Pseudo Code

1: **Initialization:** Select each arm m_a times

2: **Update statistic:** $B_{a,T_a(t-1)} \leftarrow \hat{\alpha}_{a,T_a(t-1)} + \sqrt{\frac{2 \log t}{C_a^w T_a(t-1)}}, \quad \forall a \in \mathcal{A}$

3: **Select arm:**

4: Sort $\{B_{a,T_a(t-1)}\}_{a=1}^{|\mathcal{A}|} : B_{(1),T_{(1)}(t-1)} \geq B_{(2),T_{(2)}(t-1)} \geq$

$\ldots, B_{(K),T_{(K)}(t-1)}, \ldots, B_{(|\mathcal{A}|),T_{(|\mathcal{A}|)}(t-1)}$

5: $\mathcal{A}_{BK} \leftarrow \{B_{(1),T_{(1)}(t-1)}, B_{(2),T_{(2)}(t-1)}, \ldots, B_{(K),T_{(K)}(t-1)}\}$

▷ Construct a set ($\mathcal{A}_{BK}$) that contains the top-K indexes

6: $z^* \leftarrow \text{SolveFLLP}(\mathcal{A}, \mathcal{A}_{BK}, \theta^{BK}, \mathcal{A}_A, \theta^A)$

▷ The objective of SolveFLLP is (3.16); the solution satisfies BK- and AA-fairness

7: Choose arm $a \in \mathcal{A}$ with probability z_a^*

8: **Increment** t and **Go to Step 2**

Above is the pseudo code for our proposed FL-UCB algorithm, where the optimization

of FL-LP is called as a subroutine. The objective function of SolveFLLP in step 6 is (3.16).

3.5.2 The FL-UCB Regret Bounds

When $\theta^A \neq \mathbf{0}$, the difference between the offline (optimal) policy without fairness constraints and an optimal fair policy is, in general, $O(t)$, by the definition of BK-fairness and AA-fairness. (Only when $\mathcal{A}_{BK}$ and $\mathcal{A}_A$ contain only the optimal arm does this fail to hold.) We therefore define the *price of fairness* in the SLT context.

Definition 3.5.3. The price of fairness, or PoF, is the gap between the total reward of the optimal policy and the optimal fair policy.

We thus define the difference between the objective value of the optimal fair policy and a given fair policy, which we call the *F-regret*; it is incurred solely due to a lack of information about the arm parameters. Specifically, in the original definition of *regret*, π_t^* is defined as the offline policy over all possible policies; now, we are restricting the feasibility set by imposing fairness constraints. Since the price of fairness causes an inevitable linear loss, we focus on lowering the additional loss by efficiently using information, i.e., controlling the *F-regret*. When appropriate, we may alternatively use the terms F-regret and regret without ambiguity.

Next, we analyze the regret upper bound for the proposed FL-UCB algorithm. The regret lower bound for FL-UCB is $O(\log t)$ because the vanilla bandit is a special case, and its regret lower bound is $O(\log t)$ (Lai & Robbins, 1985). For convenience, we denote $a_{(i)}, i \in [|\mathcal{A}|]$, and $\alpha_{(i)}$ as the i-th best arm and its aptitude parameter, respectively, and let $\Delta_{a_{(i)},a_{(j)}} := \alpha_{(i)} - \alpha_{(j)}$, $i, j \in [|\mathcal{A}|]$. Recall that $r(\ell, a, (T_{a,t-1}, s_{a,t-1}))$ is the

random reward of pulling arm $a \in \mathcal{A}$ with experience level $s_{a,t-1}$ (when arms are mutually independent, $s_{a,t-1} = T_{a,t-1}$). We further define $\bar{r}_a = \sup_t r_a^{(t)}$ and $\underline{r}_a = \inf_t r_a^{(t)}$. Theorem 3.5.1 establishes bounds on the FL-UCB regrets.

Theorem 3.5.1. When LP (3.11) $\sim$ (3.15) has a unique solution:

(a) The expected number of times that the (non-degenerate) solution of the LP with objective (3.16) is different from that of (3.11) $\sim$ (3.15), satisfies

$$\sum_{a \neq a^*} \mathbb{E}[T_a] \leq \left(\sum_{a \neq a^*} \frac{8 \log t}{C_a^w \Delta_a^2} + \sum_{k=2}^{K} \sum_{i=1}^{|\mathcal{A}|-K} \frac{8 \log t}{C_a^w \Delta^2_{a_{(k)}, a_{(K+i)}}} + \frac{8 \log t}{C_a^w \Delta^2_{a_{(k-1)}, a_{(k)}}} \right) + (2|\mathcal{A}| - K)(K+1)\zeta(1.24)$$

(b) The F-regret is bounded by

$$\mathbb{E}[R(t)] \leq \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} \left(\bar{r}_{a^*} - \underline{r}_{a_{(i)}} \right) \left(\frac{8 \log t}{C_a^w \Delta^2_{a_{(k)}, a_{(k+i)}}} + 2\zeta(1.24) \right)$$

Parameter	Value	Comment
BK-Fairness	$(K, \theta^{BK}) = (10, 0.02)$	Uniform for every group
AA-Fairness	$\theta^A = 0.001$	Uniform for every group
Queueing	$\kappa = 0.5$	Equal weight

Table 3.1: Experiment parameters

Rates\OPO Regions	Region 4	Region 5
Liver arrival rates	$(1.24, 2.32, 2.98, 3.12)$	$(1.54, 2.43, 2.78, 3.32)$

Table 3.2: Experiment setup: livers arrival rates

Parameter\Team type	Average	Better	Note
Sg. Aptitude ($\mathcal{A}$)	$[0.6, 0.95]$	$[0.8, 0.99]$	$\alpha = \sup_t r(l_t)$
Sg. Base Perform. ($\mathcal{B}$)	$[0.05, 0.2]$	$[0.15, 0.3]$	Uniformly drawn; $\inf_t r(l_t)$
Sg. Learning Rate (Ω)	$\{5, 6, \ldots, 50\}$	$\{5, 6, \ldots, 20\}$	Uniformly drawn; slope

Table 3.3: Experiment setup: medical teams

3.6 Extensions

This section discusses extensions of our model to incorporate delayed feedback and arm correlation.

3.6.1 Delayed Feedback

Some SLT outcomes are not immediately observed after the surgery, e.g., 1-month and 1-year survival. When the true outcome $r_a^{(i)}$ is only observed after some delay, we may use perioperative data and clinical metrics to provide an initial outcome estimate, $\hat{r}_a^{(i)}$, and replace it with the true outcome $r_a^{(i)}$ when it becomes available.

Corollary 3.6.1. Let $k_a := O(1)$ be the maximum number of true rewards that haven't been revealed yet for arm a. Assume $\exists\ n_e > 0$, and the estimated outcome $\hat{r}_a^{(i)}$ and estimator function ϕ satisfy the property:

$$e_{a,n} := \phi(\hat{r}_a^{(n)}, \ldots, \hat{r}_a^{(n-k_a+1)}, r_a^{(n-k_a)}, \ldots, r_a^{(1)}) - \phi(r_a^{(n)}, \ldots, r_a^{(1)}) \leq \frac{1}{40}\sqrt{\frac{\log n}{nC_a^w}}, \quad \forall n > n_e \tag{3.17}$$

When all premises in Theorem 3.4.1 hold; then, even when feedback is delayed by k_a

for each arm a, each sub-optimal arm is pulled in expectation at most

$$\mathbb{E}[T_a(t)] \leq \frac{8 \log t}{C_a^w \Delta_a^2} + 2\zeta(1.063) \tag{3.18}$$

times, where $\zeta(1.063) \approx 16.45$, and $\zeta(s)$ is the Riemann zeta function, i.e. $\zeta(s) = \sum_{i=1}^{\infty} i^{-s}$. The expected cumulative regret of the L-UCB algorithm when feedback may be delayed is bounded by

$$\mathbb{E}[R(t)] \leq \sum_{a \neq a^*} (\bar{r}_{a^*} - \underline{r}_a) \left(\frac{8 \log t}{C_a^{\omega} \Delta_a^2} + 2\zeta(1.063) \right). \tag{3.19}$$

The proof is shown in Section B.3.

Note that the estimator error bound in (3.17) is relatively mild: It is much looser than the error decay rate of taking a sample average while having k_a delayed, unobserved outcomes, which is at the scale of $O\left(\frac{1}{n}\right)$. For learning curves of the form: $\theta = \alpha g_{\omega}(s)$ and an MoM estimator $\hat{\alpha}_n^{MoM} = \frac{1}{n} \sum_{s=1, g_{\omega}(s) \neq 0}^{n} \frac{r^{(s)}}{g_{\omega}(s)}$, the decay rate of the MoM estimator's $e_{a,n}$ is also $O\left(\frac{1}{n}\right)$ which satisfies (3.17). Our assumption that k_a does not scale with n (the number of arm pulls of arm a) is reasonable in the SLT application because, in practice, only a finite number of livers become available within any fixed period; that is, there are a finite number of arm pulls during the survival period (e.g., one-year).

3.6.2 Incorporating Feature-Based Rewards and Arm Correlation

Each transplant surgery outcome/bandit reward is determined by the surgical team's proficiency (that is unknown and needs to be learned), the patient's clinical (e.g., serum bilirubin, creatinine, and the international normalized ratio) and demographic information (e.g., age, BMI), and the donated liver's compatibility (e.g., size matching, ABO compatibility) and quality (e.g., donor age and health, cold ischemia time.) Thus, a natural extension of our MAB model is to formulate feature-based rewards for each transplant surgery: Each arm is fully characterized by a potentially high-dimensional vector consisting of known patient and liver attributes, and the central planner learns the relationship between surgical teams' experience and transplant outcomes. Moreover, we can further decompose surgical proficiency to capture overlaps in required skills across surgeries. Feature-based rewards and high dimensionality in dynamic learning problems have been studied in revenue management contexts (Ban & Keskin, 2021; Keskin et al., 2023). Exploring the salient surgery features and surgical teams' experience could be a promising direction and facilitate detailed characterization of likely correlated expected rewards for different arms.

In the appendix, we discuss a special type of arm correlation: Linear correlation, and discuss its impact on the optimal policy compared to a clairvoyant policy, and L-UCB/FL-UCB performances.

3.7 Numerical Study

We run numerical experiments based on real-world data to test the performance of our proposed algorithms. Specifically, we consider the training and selection of med-

ical teams as part of an SLT expansion effort coordinated by the OPTN, the central planner overseeing organ allocation in the US. We estimate parameters and generate outcomes based on Standard Transplant Analysis Research (STAR) files and Potential Transplant Recipient (PTR) dataset provided by Organ Procurement and Transplant Networks (OPTN) to capture current SLT practice. The "true" optimal cumulative reward for t is $\mathbb{E}\sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi_t^*}, S_\tau^{\pi_t^*})$, where $\pi_t^* := \text{argmax}_\pi \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi}, S_\tau^{\pi})$. In Section 3.7.2 we show that our proposed algorithms converge rapidly and demonstrate asymptotic advantages. The problem of showing theoretical bounds for small-t scenarios is beyond the scope of this chapter, as the offline policy may change as t grows, as we illustrated in Example 3.4.1.

3.7.1 Numerical Experiment Setup

The time horizon in this experiment is $\{1, 2, \ldots, 3600\}$; each time step corresponds to an arrival of a split liver graft and marks the beginning of a matching run (that may last for hours after the donor dies and donates their liver). Recall that more than 10% of all deceased-donor livers in the US are medically safe to split; there were 14905 deceased donors in total during 2022. We consider a geographical region that includes OPTN regions 2, 9, 10, 11, and Wisconsin and Illinois (see Section B.6 for details about allocating heterogeneous livers as parallel MABs). Around 8000 livers are donated annually, and 10 large transplant centers locate in the 500NM Circle. While in theory, all medically-safe livers can be split, in conversations with UCSF transplant surgeons, they suggested it would be helpful to consider a more gradual rollout in the initial phase of SLT expansion to accommodate surgical learning. Thus, if we assume that 150 or $\sim 2\%$ of the total deceased-donor livers will be split for 300 surgeries in a year for the first two years, and 500 or $\sim 6\%$ livers to be used for 1000

SLT surgeries annually in the third to fifth years, the time horizon $\{1, 2, \ldots, 3600\}$ would be around five years at the typical deceased liver donation level in the US. We focus on identifying the arm(s) with the highest aptitude(s) such that they would perform the best in the long term, i.e., expanding the base of SLT among transplant teams with the highest potential. We investigate how we can accelerate bandit learning by incorporating information about the learning curve structure via our proposed FL-UCB algorithms.

SLT has been primarily practiced in a few big TCs in the US since its development in the 1980s (Duke Health, 2021; Ge et al., 2020); thus, there is no historical data for widespread SLT learning. Nevertheless, based on findings from existing studies on medical learning and the nature of SLT surgeries (involving repetitive tasks such as dividing and connecting blood vessels), the medical teams' learning curves likely follow an 'S'-shaped structure (Le Morvan & Stock, 2005; Pusic et al., 2015). The bandit rewards or SLT outcomes are 1-year graft survivals, primarily dependent on surgeon proficiency and experience; 1-year graft survival may also correlate with donor and recipient age and the recipients' health conditions. These factors and surgical expertise collectively determine the MAB asymptotic rewards or the arm's full potential/aptitude. There has been ongoing research investigating how survival outcomes depend on TC expertise, high-dimensional demographics, and perioperative clinical metrics; unfortunately, there are no exact mappings from these factors to the outcomes. As discussed in Section 3.6.2, high-dimensional feature-based dynamic learning in SLT is a potential research direction. Here, we simulate the arm parameters and outcome distributions based on historical data without specifying the exact feature mapping.

Specifically, we formulate the reward functions of arms following the Sigmoid curve; each with hidden aptitude parameter, where the bounds of the range are estimated directly from the STAR files. Except for few big TCs that already perform SLTs regularly, most TCs need to learn SLT with limited existing experience/initial proficiency and overcome barriers in the initial phase of surgical learning (characterized by ω). Since we do not have a direct data source, as SLT has not been widely learned or practiced, we consulted UCSF surgeons, and they believe the range $\omega \in [1, 14]$ is realistic: Typically, after performing $6 \sim 15$ SLT surgeries, a medical team can be considered sufficiently experienced and the proficiency starts to stabilize. Factoring in existing experience and skills transferred from similar surgeries, we arrive at $\omega \in [1, 14]$. We assume that the ranges of α's that we draw from are the same as the ranges of long-term expected outcomes in historical data containing mostly traditional WLT and limited SLT surgeries, $(0.3, 0.95)$; recent findings show that SLT outcomes can be comparable to those of WLT (Hackl et al., 2018). The parameters are specified in Table 3.4, where the learning curves follow (3.20):

$$\theta(\alpha, s) = \frac{\alpha}{1 + \exp(-s + \omega)} \tag{3.20}$$

In one set of the simulation, we assume the true rewards, 1-year graft survivals, are not observed immediately after the surgery. Since the time horizon is $\{1, \ldots, 3600\}$ and we consider an approximately five-year, gradual rollout of SLT expansion. The delay in observing our true rewards are 300, 300, 1000, 1000, and 1000, for SLT surgeries performed in the 1st to 5th year, respectively. In other words, the true rewards, i.e., 1-year graft survival, is delayed 300-time steps if an arm pull takes place in the first two years and 1000-time steps in the third to fifth years. Before the true rewards are

Parameter	Value	Comment
Number of arms	50	Each arm has a learning curve
BK-Fairness	$(K, \theta^{BK}) = (1, 0.05)$	$K = 1$ implies no BK constraint
AA-Fairness	$\theta^{A} = 0.001$	Uniform for every arm; $\mathcal{A}_A = \mathcal{A}$
Aptitude/Full potential	$\alpha \in (0.3, 0.95)$	α unknown
Initial setup cost	$\omega \in [1, 14]$	Known, existing skills

Table 3.4: Experiment parameters. More details can be found in Section B.6.

observed, we have 1-year survival estimates based on perioperative clinical metrics and demographic information. The prediction accuracy for 1-year graft survival can be as good as around 85% (Kantidakis et al., 2020; Nitski et al., 2021); we choose to be more conservative in this simulation and assume the accuracy for the estimates is 0.6. See Section B.6 for more details.

Given the experimental configurations above, we compare the performances (i.e., regrets) of the FL-UCB algorithm with MLE against seven other bandit algorithms — vanilla UCB, discounted UCB, reweighted UCB, ϵ-greedy, explore-then-commit (ETC), vanilla Thompson sampling (TS), and learning-enhanced Thompson sampling (L-TS) where we infuse learning-curve information into the TS posterior updating function.

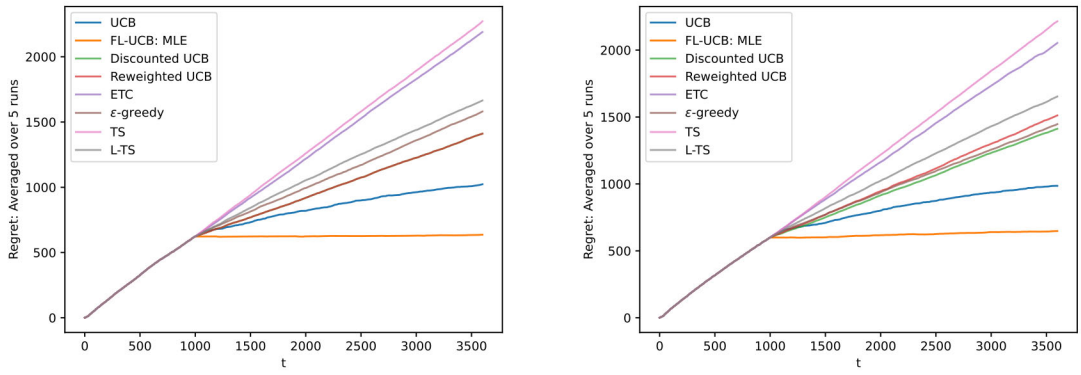
Vanilla UCB, discounted UCB, and reweighted UCB can be viewed as special cases of FL-UCB; they do not assume knowledge of the parametric form of learning curves, see Section 3.4.7. In our setting, the parametric forms of medical learning are known; FL-UCB with MLE can leverage this knowledge and accelerate bandit learning and achieve faster convergence compared to these nonparametric methods. Nevertheless,

in scenarios where either the existence of endogenous learning or the parametric form is unknown, FL-UCB with nonparametric estimators can usually generalize well. As our chapter focuses on FL-UCB with MLE and nonparametric estimators we do not elaborate on L-TS; nevertheless, numerical results show that L-TS performs consistently better than canonical TS when endogenous learning is present and is second only to FL-UCB when the true feedback is delayed. For all bandit algorithms, we start with 20 round robins and report regrets/rewards that are averages over five runs. We also provide numerical results on how the offline reward (i.e., optimal cumulative reward) and PoF (i.e., price of fairness) grow as functions of t.

3.7.2 Numerical Results

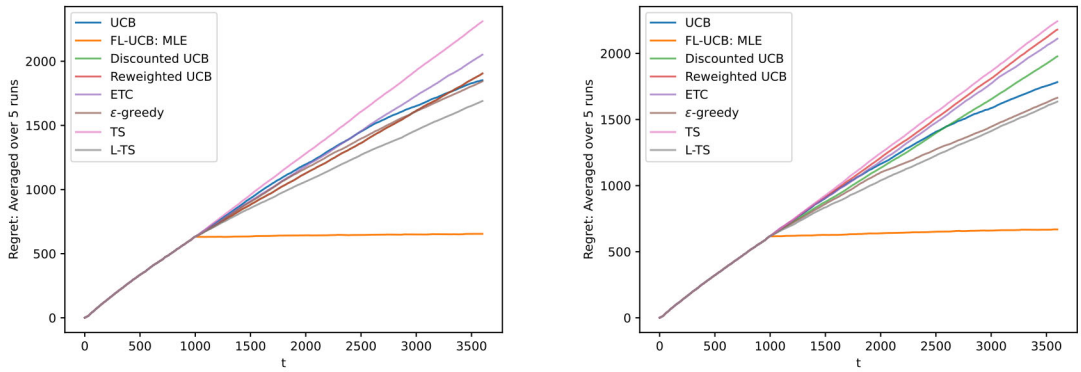
Figures 3.7 and 3.8 show the total regret of each algorithm as a function of t, the total number of surgeries performed; while figure 3.9 illustrates the optimal cumulative reward, PoF, and PoF percentage as a function of t. Specifically, Figure 3.7 demonstrates that when surgical learning occurs, FL-UCB outperforms the benchmarks as it has the lowest regrets and converges rapidly, whether the fairness constraints are imposed or not. Figure 3.8 shows that when the true rewards are delayed and 60%-accurate reward estimates are available, the advantage of FL-UCB is preserved: FL-UCB with the MLE estimator still outperforms other bandit algorithms and achieves similar regrets, while UCB and nonparametric L-UCB variants incur greater regrets compared to the no-delay simulations, regardless of the presence of fairness constraints.

Figure 3.9 shows that the PoF, the loss in utility in optimal fair solutions relative to optimal solutions without fairness constraints, is small (although it is still $O(t)$, as the PoF / Optimal cumulative reward ratio remains constant as t grows). In figure 3.10



(a) Regret w/o fairness **(b)** F-regret w/ fairness

Figure 3.7: Comparing FL-UCB regret against benchmarks when medical learning exists and assuming no delay in observing true rewards. FL-UCB with MLE estimation has the lowest regrets and converges rapidly.



(a) Regret w/o fairness **(b)** F-regret w/ fairness

Figure 3.8: Comparing FL-UCB regret against benchmarks when medical learning exists and rewards (i.e., 1-year graft survival) are delayed. We assume estimates based on demographics and perioperative clinical metrics are available and are 60% accurate. FL-UCB with MLE estimation learns efficiently in the initial round-robin exploration phase (where each arm observes 12 true outcomes and 8 estimated outcomes) and still has the lowest regret and converges fast. Meanwhile, UCB regrets are much higher when the true feedback is delayed.

we illustrate the breakdown of data points available at time t, $t \in \{1, \ldots, 3600\}$. For the first 300 time steps, all available information comes from outcome estimates,

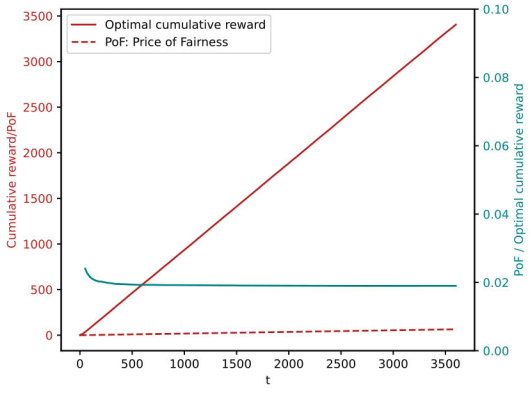


Figure 3.9: [PoF / Optimal cumulative reward] is constant, i.e., PoF is $O(t)$; when $t < 200$, the ratio could be subject to numerical instability.

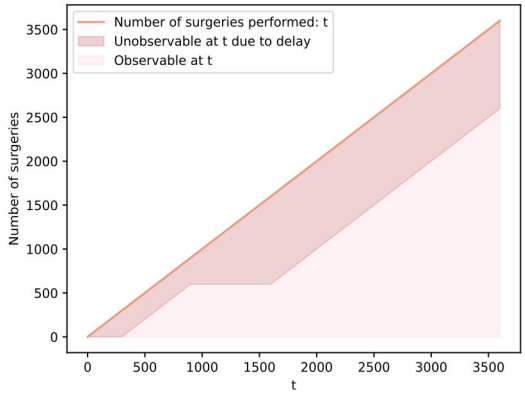


Figure 3.10: The delay in observing rewards. For each t, the true rewards are not revealed until after some delay and can only be estimated using a 60% accurate surrogate.

while in later stages, only a dwindling proportion of cumulative rewards is delayed and requires prediction. Specifically, the number of delayed rewards is t for $t \leq 300$, 300 for $t \in [301, 900]$, and $\min\{1000, t - 600\}$ when $t \geq 900$.

3.8 Concluding Remarks

To address the trade-off between exploration versus exploitation in SLT allocation, we formulated an MAB variant with endogenous learning curves embedded in the arm reward functions. Our model also enables the incorporation of learning curves, fairness constraints, and nonparmetrics (i.e., UCB variants that do not assume knowing the parametric forms). We propose UCB variants—the L-UCB and FL-UCB algorithms, which converge to the optimal offline policy incurring optimal total regret: $O(\log t)$ scale. Application of our model can potentially shed insights on strategies (e.g. liver allocation, transplant center/surgery selection) to expand SLT use in the US, as well as in other settings characterized by decision-making with experience-based learning,

such as call centers and franchising. Methodologically, our formulation and proposed UCB variants significantly extend the canonical UCB to bandit problems where the estimates of the unknown parameters can be different from the empirical mean.

There are several potential directions for future work that will generalize and deepen the conclusions of this chapter. Methodologically, we developed non-parametric methods, discounted UCB and reweighted UCB, and compared them against parametric L-UCB; results show that we can leverage the parametric forms of learning curves through MoM or MLE or MAP estimators in L-UCB to achieve lesser regrets. In scenarios where the parametric form is unknown, reweighted UCB and discounted UCB and vanilla UCB can be used, and the former two may sometimes perform better than vanilla UCB, but not always. An intriguing future direction is to study the selection of good nonparametric estimators under the FL-UCB framework and estimators' robustness in various applications. Another promising direction is to incorporate feature-based rewards that explore the correlation between the expected rewards of different arms, please refer to Section 3.6.2 for more details. Moreover, our proposed L-TS algorithm has fair numerical performance as shown in Section 3.7, it is possible that L-TS has theoretical regret bounds. One could investigate this in a separate paper. On the application side, an important extension is to include candidates' strategic accept/reject decisions when offered organs. Observable dynamics in patient queues can also be factored into the definition of the bandit rewards and subroutines of the L-UCB algorithm. Specifically, rewards can be redefined to penalize queueing delays and, more generally, account for any hidden cost not captured with surgical outcomes of surgeries. And incentives to practice SLTs are intentionally left to future research: The discussion on incentive compatibility is beyond the scope of

this chapter; the solution to this likely requires a package of policies.

Chapter 4

Human-Artificial Intelligence Teaming and Effects of System Load on the Screening of Child Maltreatment Reports

4.1 Introduction

4.1.1 Child Welfare Services

The Centers for Disease Control and Prevention estimates that at least 1 in 7 children in the US have experienced child abuse or neglect in the last few years. In 2020, 1,750 children in the US died of abuse and neglect; and the total lifetime economic burden of child abuse and neglect was estimated to be $592 billion in 2018. This economic burden rivals the cost of other high-profile public health problems, such as heart disease and diabetes. Abused or neglected children may suffer immediate physical injuries and are more likely to suffer emotional and psychological problems later in their lives, such as anxiety or post-traumatic stress. In the long term, maltreated children are at higher risk of future violence victimization and perpetration, substance abuse, delayed brain development, lower educational attainment, and limited employment opportunities (CDC, 2023).

Child welfare organizations are tasked with protecting children from abuse and neglect; their responsibilities include investigating child maltreatment allegations and providing services to children and families in need. However, not all allegations are substantiated, and unnecessary investigations may harm the families involved. More-

over, many child welfare organizations need to prioritize scarce resources and efforts to substantiated and serious referrals. Therefore, it is crucial that allegations are carefully screened before initiating formal investigations for the overall welfare of children, families, and the community.

A child maltreatment report or call is designated as a *referral*; one referral may involve more than one child and several allegations. Referrals are categorized into two types: child protective service (CPS) and general protective service (GPS). CPS reports are made by *mandated reporters*, i.e., adults who are required by law to report suspected child neglect or abuse. Mandated reporters are those adults who work or volunteer with children, including school employees, healthcare professionals, foster parents, and employees at other public services organizations. Many US states (e.g., Michigan, Pennsylvania) established laws that mandate CPS referrals to be investigated within 24 hours (Michigan HHS, 2023; Pennsylvania DHS, 2023a). In contrast, GPS referrals are only investigated by an assigned caseworker if *screened in*—accepted for investigation. Child abuse may take various forms; common abuse types include physical abuse, sexual abuse, emotional abuse (i.e., behaviors that harm a child's self-worth or emotional well-being, e.g., name-calling, shaming, rejecting, withholding love, and threatening), and neglect physically or emotionally (CDC, 2023).

4.1.2 Operations in a Child Welfare Organization

We partner with a child welfare organization (CWO) within a county in the United States. Figure 4.1 illustrates the workflow of the CWO. Specifically, their operations have three main stages: screening, investigation, and intervention. In the screening stage, all incoming GPS calls or reports are screened by call screeners in the CWO's

intake office. Social workers at the intake office are mainly tasked with answering calls, assessing referrals, and investigating CPS referrals. GPS referrals, if screened-in, will be investigated by one of the five regional offices located in five different geographic regions in the county. CPS referrals are mandatorily investigated by social workers, often at the intake office. When an investigation concludes a referral is in need of services, a *case* for the referral is opened, and a social worker is assigned to intervene, offer protection, provide service, and work up a long-term solution for the children and family involved.

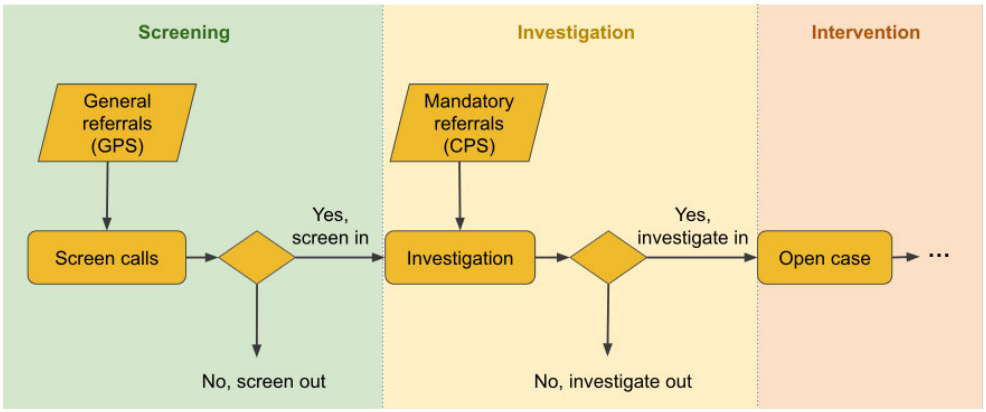


Figure 4.1: CWO workflow diagram.

A predictive risk model (PRM) has been designed and implemented to enhance the screening decision making process within the CWO's child welfare system. The PRM harnesses the power of hundreds of data elements to generate the *PRM score*, which quantifies the likelihood of a child being placed *out-of-home* (OOH): The PRM score takes integer values ranging from 1 to 20, with a higher value indicating a higher probability of OOH placement within two years Pennsylvania DHS, 2023b. The PRM is designed to complement clinical judgment by offering additional vital information to aid child welfare workers in making informed call-screening decisions. The PRM AI tool is generated and viewed only by call screeners; caseworkers who conduct

investigations and provide services are not able to see any information from PRM.

Stage 1: Screening. Figure 4.2 illustrates the workflow of the screening stage. If someone in the community has concerns about suspected child maltreatment, they may report it to the CWO's hotline. When a call comes in, a call screener at the intake office answers the call while recording information into the computerized system following the established protocols. For all calls—those falling within a protocol or not—the screener runs the PRM on their computer, which will take the input of historical data (including demographics and past interactions with the CWOs) and the call information, and output the assessed risk score. After obtaining the PRM risk score, a field screen might be conducted if the call screener would like to gather more information about the children, the household, and the allegations for screening decision[1].

Some referrals with specific characteristics (i.e., having a PRM score greater than 17 and involving a child aged 16 or under) fall into the *high-risk protocol* and, therefore, are designated to be screened in. In contrast, a small percentage of referrals fall into the *low-risk protocol* (i.e., having a PRM score no greater than 11 and no children under 12) and are encouraged to be screened out. Many other referrals fall into neither high-risk nor low-risk protocols. While the risk protocols guide screening decision-making, intake office supervisors can override the protocol-suggested screening outcomes at their discretion, given that they complete the override documentation.

[1]A field screen is typically conducted if one or more of the following conditions are met: a) The child maltreatment report involves children three years old and younger who are directly impacted by the allegations. (b) If a report is the fourth referral associated with the same household within two years, yet there has not been any previous investigation into the household. And (c) a report involving children who receive education (through homeschooling, distance learning, or remote learning) at home.

However, the computer information system shows the PRM score to the call screener only when a referral does not follow either risk protocol. For a referral that follows a high-risk protocol, the human-system interface displays “High-Risk Protocol: High Risk and Children Under 16 on Referral." Similarly, for a referral that follows a low-risk protocol, the human-system interface displays “Low-Risk Protocol: Low Risk and All Children Aged 12+ on Referral" and “recommended screen out."

Call screeners recommend screening in or screening out referrals. based on their assessment, the risk protocols or PRM risk scores, and field screens when necessary. All call screeners' recommendations go through their supervisors, who make the final screening decisions based on the risk protocols, PRM scores, relevant data, and call screener recommendations. The call screeners and supervisors often discuss the evidence and screening rationale and then jointly make the screening decisions. Currently, 48.05% of incoming referrals are screened in for investigation.

Stage 2: Investigation. All screened-in GPS referrals and occasionally some CPS referrals are investigated by one of the five regional offices in different geographical regions within a US county. The intake office investigates most CPS referrals. Caseworkers must conclude any investigation within 60 days, though best practice encourages a 30-day completion of any CPS investigation. The investigation determines whether a CPS report is founded (i.e., there is a court action), indicated (i.e., there is substantial evidence that maltreatment occurred), unfounded (i.e., existing evidence does not meet the criteria for maltreatment), or pending (i.e., the CWO investigation cannot be concluded in 60 days because criminal or juvenile court action is initiated). GPS investigations conclude with whether allegations are valid or not

based on collected evidence. If the caseworker(s) and their supervisor(s) believe there is an ongoing risk of child maltreatment in the household, then the referral may be accepted for service, and a child welfare case will be opened. If a child welfare case is not opened for the family, other community-based resources might still be offered to assist the household if needed.

Stage 3: Intervention. Upon opening a case (or equivalently, being accepted for service), a child welfare case worker will arrange a conference meeting with the family and their identified support (e.g., friends and other family members). At the conference meeting, they will discuss the family goals and devise a plan for services so the child can remain safely within the household. Subsequent meetings are held with the same stakeholders to ensure that the family is making acceptable progress toward the goals. The investigation caseworker who conducts the investigation often remains with the family to provide service.

The CWO and other supervising bodies periodically review ongoing cases. In scenarios where they believe a child can no longer safely remain in a household, an out-of-home placement or removal from the home will be considered. Other interventions and services the CWO provides include foster care placement, adoption, permanent legal custodianship, and reunification.

4.1.3 Research Overview

In this project, we examine the interplay between human workers and the PRM, specifically assessing the influence of system load on child welfare screening recommendations and decisions. Recent research underscores the benefits of human-AI col-

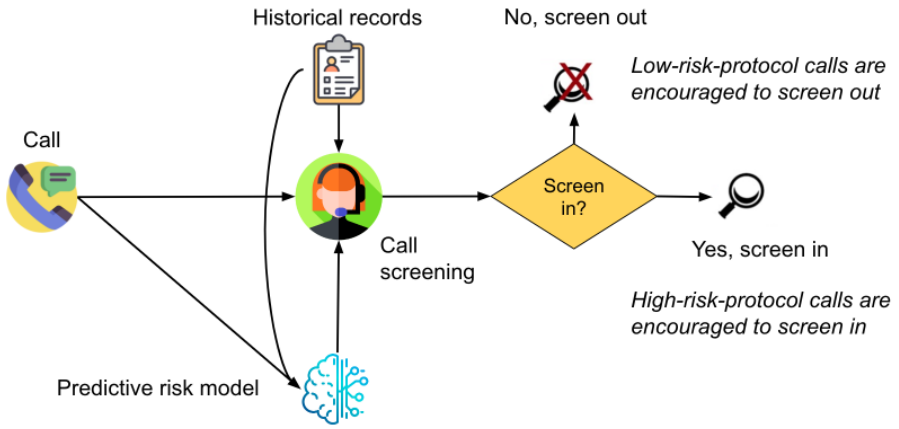


Figure 4.2: Workflow diagram: the screening step.

laboration in diminishing disparities and errors. For example, Fogliato et al., 2022 highlights how human discretion in high-stakes contexts like child welfare can counterbalance algorithmic inaccuracies and reduce disparities. However, operational challenges have received scant attention in the existing literature. Our study fills the gap: We explicitly focus on the role of system load in human-AI collaboration in a high-intensity service environment with capacity constraints. Empirical evidence suggests that human workers are more likely to deviate from load-agnostic PRM recommendations when the workload level is either very high or low. Our findings reveal that human workers adeptly recognize the implications of system load on both organizational efficiency and service quality, thereby making screening choices, informed by clinical judgement and PRM's risk scores, to ensure workload sustainability.

The rest of this chapter is organized as follows: Section 4.2 summarizes relevant papers and positions our work in the existing literature. Section 4.4 presents our main analysis results. Section 4.5 discusses our approaches to alleviate endogeneity concerns and demonstrates the robustness of our main results. Section 4.6 details the

roadmap for this ongoing project and enumerates our next steps.

4.2 Literature Review

This work is relevant to two main streams of literature: the child welfare system and human-AI teaming.

Child welfare systems. The child welfare system in the United States is comprised of a network of services and policies aimed at protecting children from maltreatment such as abuse and neglect. Recent literature discusses its structure while highlighting the challenges and areas that could be improved (Slaugh, 2024).

The U.S. child welfare system is decentralized. Each state has its own system, though all under federal laws and guidelines. Foster care, a critical component of child welfare, involves removing children from homes where they are at substantial risk of maltreatment and providing them with substitute care. Other essential aspects of the child welfare system in the US include foster care and adoption (Olberg et al., 2021).

One significant challenge is the overrepresentation of children of color, particularly Black children, in the child welfare system. Literature on racial disproportionality and disparities sheds light on the systemic and societal factors contributing to this issue and calls for targeted reforms (Doe & Clark, 2020). Additionally, aging out of foster care without adequate support leads to higher risks of homelessness and unemployment among these young adults (Wilson, 2019).

Existing literature underscores the complexity of the U.S. child welfare system, the myriad challenges it faces, and policy-making. However, research focusing on the oper-

ational improvement for the upstream services (e.g., call screening and investigation) in child welfare organizations is sparse (Slaugh, 2024). This chapter fills this gap by conducting a detailed empirical study of call screening at a child welfare organization in a US county.

Human-AI teaming. Many US states have implemented algorithms to assist child welfare operations (Saxena et al., 2020). Child welfare operations adopted these algorithms to reduce costs and, ideally, improve operational efficiency, equity, and service quality. Cheng et al., 2022 showed that screening decisions are more equitable when combining the strengths of AI algorithms and humans' clinical judgement. Fogliato et al., 2022 showed that call workers adjust their behaviors after the deployment of the AI tool. Call workers are capable of integrating complementary AI recommendations with their own judgement—Evidence show that they are less likely to adhere to erroneous AI recommendations. However, existing work on human-AI collaboration in a child welfare context does not consider vital operational factors such as workload and capacity.

In a different context, Snyder et al., 2022 conducted a behavioral study that investigates humans' algorithm understanding and reliance under different pressure levels. The authors implemented laboratory experiments for a large-scale personalized recommendation context. Results show that greater time pressure increases human reliance on algorithms in general. Considering heterogeneous algorithm performances, humans rely more on superior algorithms as their ability to discern algorithm performance also improves under high load. To our best knowledge, we are the first to study human-AI teaming under various workload conditions utilizing a real-world datasets.

4.3 Empirical Setting and Data Description

This section presents the research setting, dataset, summary statistics, and the data processing procedure. We describe the key variables, dependent variables, control variables, and outcome labels in detail.

4.3.1 Research Setting

We focus our analyses on the screening stage of CWO operations. Specifically, we examine when human screeners' decisions align with or deviate from AI recommendations, and how call screening decisions are influenced by the workload. To study these research questions, we conduct empirical analyses on the referral level using the general form shown in Equations (4.1) and (4.2):

$$\text{Screening Decision} \sim \text{Workload} + \text{Controls} \tag{4.1}$$

$$\text{Deviation from AI} \sim \text{Worlload} + \text{Controls} \tag{4.2}$$

In particular, we would like to see how workload influences screening decisions and deviation from AI. Below, we introduce the data and preprocessing procedures (Section 4.3.2), the workload variables (Section 4.3.3), the dependent variables (Section 4.3.4), the outcome labels (Section 4.3.5), and controls (Section 4.3.6).

4.3.2 Data Description

Our collaborating CWO granted us access to their private datasets, which are stored on a secure remote server hosted by a large research institution in the US. The datasets contain referral-level data from January 1, 2017 to November 11, 2022. For each

incoming child maltreatment report, the datasets record the following information: the number of children involved and the child(ren)'s ID(s) in the system, demographic information about the household, zipcode, allegation type(s), abuse or neglect type(s), and the reporter's relationship to the household. From an operational standpoint, the system also records the date and time of receiving the report, the call screener's ID, and the supervisor's ID. The call screener also assesses the safety concerns and risks associated with each referral, and they document their individually assessed initial risk evaluations in three categories: High risk, medium risk, and low risk. Call screeners run the PRM tool after entering information about the child maltreatment report, and the system automatically outputs and stores the PRM score. Based on the PRM score, risk protocols, and risk evaluation (which sometimes involves a field investigation), the screener makes a screening recommendation and enters it into the database. A supervisor then authorizes a final screening decision, which may differ from a screener's recommendation, after reviewing all the referral information mentioned above.

Besides referral information, our datasets also include investigation and case details. Recall that screened-in referrals are investigated and often assigned to another caseworker at a regional office. We have an assignment table that includes the caseworker IDs and times of investigation assignments. Investigation outcomes are summarized in service decisions: "accept for service," which will be followed by opening a case and providing services by the CWO and community partners, or "do not accept for service," closing the investigation without providing services. While our analyses focus on the screening stage and discuss screening decisions' impact on the investigation stage, our datasets contain additional details regarding the services provided to cases

and case assignments.

Table 4.1 describes the summary statistics of relevant variables in the datasets.

Table 4.1: Summary statistics

	N	Mean	Median	Max	Min	SD
Case load	2140	1898	1028	2129	1637	167.45
Investigation load	2140	1047	1156	1468	530	164.54
Case load per worker in 10	2140	0.90	0.91	1.02	0.76	0.07
Inv. load per worker in 10	2140	0.46	0.46	0.59	0.26	0.07
Referral load	2140	41.93	1028	2129	1637	20.42
Number of screeners	2140	13.94	15	18	3	4.74
Calls per screener	2140	2.85	2.88	5.5	1	0.66
Investigation workforce	2140	227.7	230	275	177	19.28
PRM score	64870	13.95	15	20	1	4.57
Number of children	64870	5.14	5	22	1	2.06
Min. child age	64870	6.29	5	88	0	5.33

4.3.3 Key Variables

The key variables of our analyses are the workload at the intake office and across the entire organization. The workload at the intake office is described in the variable “referral load." “Referal load" describes the number of incoming calls/referrals that are received by the intake office. We also track the “number of screeners" that received at least one call on a given day; there is always more than one screener on duty within our time window. Similarly, based on the start and end dates of investigations and cases, we obtain the number of active “investigation load" and “case load" in the CWO

on each day.

4.3.4 Dependent Variables

The dependent variables of our main analyses include the “screen-in" decisions by supervisors and “deviation," whether the final screening decisions differ from those implied by PRM scores. In the main analyses, we define “deviation" as the follows: If the PRM score is equal or above 15, but the referral is screened out, then we label the screening decision a deviation from PRM recommendation. Similarly, if the PRM score is less than 15, but the referral is screened in, then the screening decision also deviates from PRM. Otherwise, there is no deviation. Note that the definition of deviation is mainly used for our analyses of human behaviors; it is different from that of PRM high or risk protocols which are implemented by the CWO and are applicable to only a fraction of referrals. On average, 36.03% of the time the final screening decisions deviate from PRM at our collaborating CWO.

4.3.5 Referral Outcome Labels

Consistent with CWO’s most recent documentation (Allegheny County DHS, 2024), we define the outcome of referrals as follows:

- True positive: If a referral is screened in (i.e., positive), and the investigation concludes that it requires service, i.e., “service decision" is true, then the initial screening decision outcome is labeled as true positive or TP.

- False positive: If a referral is screened in (i.e., positive), and the investigation determines that no service is necessary, i.e., “service decision" is false, then the initial screening decision outcome is labeled as false positive or FP.

- True negative: If a referral is screened out (i.e., negative), and no child on the referral is removed from the household in the next 90 days, then the initial screening decision outcome is labeled as true negative or TN.

- False negative: If a referral is screened in (i.e., positive), and at least one child on the referral is removed from the household in the next 90 days, then the initial screening decision outcome is false negative or FN.

The ultimate goal of the CWO is to provide services to families in need; this means the CWO aims to have a low FN rate. Meanwhile, the CWO strives to reduce FP rate, the reasons are twofold: First, investigating unsubstantiated and unfounded claims might cause unnecessary hassles and even harm to families. Second, the actual workload per caseworker at CWOs often exceeds the ideal levels; CWOs are therefore incentivized to reduce the number of investigations that do not lead to providing services.

The CWO must strike a delicate balance at the screening stage: If screening in all referrals, the FN rate will be zero, but this is likely infeasible due to the capacity and budget constraints and will increase the FP rate, which means a higher level of unnecessary inconvenience or even interference to families involved. If screening out more referrals, the likelihood of having a FN outcome increases. The PRM tool is designed to help making informed screening decisions by providing the percentile catogories of predicted OOH placement likelihood. As a result of deploying PRM and other accumulative effort, the CWO has maintained a remarkable FN rate of 2.28%. (Note that the initial screen-out decision for an FN referral may be actually accurate, as many things can change in 90 days.) However, the FP rate is 18.74%, which is higher than our collaborating CWO's expectation. There is an incentive within the

CWO to reduce the FP rate while maintaining the low FN rate.

4.3.6 Control Variables

Our CWO datasets enable us to use a comprehensive set of control variables in our empirical analyses. We describe our control variables below. The summary statistics for these control variables are presented in Table 4.1.

Demographics: We control for the demographics of the reporter of child maltreatment, child(ren), and perpetrator associated with referrals. If a child is suspected of being maltreated, then entire household which the child belongs to is included in the referral. A household includes the victim(s) and all children with the same mother, as well as other adults living with the mother or children. All victims, children, reporters, and perpetrators are given unique IDs and their demographics data and interactions with the organization are recorded into CWO's database. Specifically, we account for the number of child(ren) associated with the household, the child(ren)'s age(s), race(s), household zipcode(s), the reporter's relationship with the victim(s).

Allegations and risks: For each referral, we control for the allegation type(s) and the PRM score, which describes the risk of at least one child being removed from the household in the next two years. We also flag families that are linked to an active case with the CWO with the variable "active family." About 16.13% of all referrals are linked to active families. These control variables together with the PRM scores help account for the nature of allegations and overall risks.

Organization workload and workforce: To better capture the organization's workforce and capacity, we include the "number of screeners" as a control. Simi-

larly, based on how many workers are actively working on investigations and cases, we will track the “investigation workforce" and “case workforce" on daily levels and include them as controls once schedule data becomes available.

Time and holidays: We control for the year, month, day of the week (DoW), and holidays for our main analyses. We also present additional robustness checks, for which we additionally account for weekends as well as winter and summer vacations.

4.4 Empirical Investigation on Workload's Impact in Call-Screening

This section discusses the main analyses of workload effects on call screening decisions. Section 4.4.1 describes the model specification, and Section 4.4.2 presents the main regression results and their implications on human-AI teaming in multi-stage operations.

4.4.1 Model Specification

In our primary empirical analyses, we investigate the effects of workload levels—both during the call-screening phase and across the entire system—on call-screening decisions and the collaboration between screeners and PRM. These analyses focus on GPS referrals, where call screeners and their supervisors have decision-making discretion. (Recall that CPS referrals are mandated to be investigated.) Our main regression specifications for the referral-level analyses are as follows: Let Y_j denote the supervisor-approved call screening decision for referral j arriving at time t (1 for

screened-in and 0 for screened-out).

$$\begin{aligned}\text{logit}(Y_j) =& \alpha + \beta_1 \text{Investigation\_Load}_t + \beta_2 \text{System\_Load}_t \\ &+ \text{Referral\_Controls} + \text{Time\_Controls} + \text{Screener\_FE} + \epsilon_j,\end{aligned} \tag{4.3}$$

Let Z_j represent the deviation from the PRM suggested screening decision (1 for deviation and 0 for alignment) .

$$\begin{aligned}\text{logit}(Z_j) = \kappa &+ \gamma_1 \text{Investigation\_Load}_t + \gamma_2 \text{Investigation\_Load}_t^2 \\ &+ \gamma_3 \text{System\_Load}_t + \gamma_4 \text{System\_Load}_t^2 \\ &+ \text{Referral\_Controls}_j + \text{Time\_Controls}_t + \text{Screener\_FE}_i + \epsilon_j,\end{aligned} \tag{4.4}$$

We incorporate comprehensive controls across referral characteristics, time, and screener dimensions; see details in Table 4.2.

4.4.2 Main Regression Results

In the main analyses, we focus on GPS referrals for children no younger than seven years old, filed between January 2018 and March 2022. Table 4.2 presents the workload's impact on the probability of call screening decisions and deviations from AI recommendations. The results suggest a U-shape relationship between deviation and workload. In particular, human agents are more likely to diverge from the AI recommendations when the system workload is either high or low, rather than moderate. This might seem counter-intuitive at first. Yet a closer look reveals that while the AI tool does not adjust for varying workloads, human agents seem to factor in the workload when making screening decisions. Specifically, they lean toward admitting more low-risk cases when the case load is low and rejecting more high-risk cases when the

case load is high. We reason that this behavior has the potential to enhance overall system performance, as it helps maintain a sustainable system workload.

4.5 Robustness Analysis

In our primary empirical analyses, as presented in Equation 4.3, it is important to address concerns related to endogeneity. One element of particular concern is the referral load and investigation load. The system load may not be directly connected to the focal referral j under investigation and, therefore, may be less problematic. However, the referral load requires careful scrutiny. The underlying reason is that there could be confounding variables influencing both the incoming referral volume and their screening decisions since both are directly associated with the focal cases. It is crucial to ensure that these external factors do not inadvertently bias our findings. Below, we discuss the approaches taken to alleviate potential endogeneity concerns associated with the referral load, investigation load, and case load.

First, in Section 4.4, we intentionally employ the shift-level workload (i.e., the daily average for all call-screeners) as our primary measure for the “referral load" instead of the workload for individual call screeners. This is because individual workload might be subject to potential endogeneous assignment rules and practices. For instance, call screeners could be assigned to certain types of referrals: Some might specialize in high-complexity (high-risk) reports while others focus on low-risk ones. If screeners for low-risk referrals consistently handle more cases due to their simpler nature, we could see a trend where increased individual workloads correlate with decreased screen-in rates. However, this does not imply that a system-wide increase in referrals leads to lower screen-in rates. To avoid this endogeneity concern arising from assignments, the

Table 4.2: Workload's impact in call-screening

	Screen-In	Deviation
Key Variables:		
Case load	-0.026^{***} (0.003)	-0.231^{***} (0.065)
Case load (sq)		0.001^{***} (0.0002)
Investigation load	0.0004 (0.002)	-0.003 (0.012)
Investigation load (sq)		-0.00001 (0.0001)
Controls:		
Referral load	-0.007^{***} (0.002)	0.002 (0.002)
Number of call screeners	0.025^{***} (0.009)	0.002 (0.008)
PRM score	0.146^{***} (0.005)	0.088^{***} (0.005)
Number of children	0.060^{***} (0.014)	-0.022^{*} (0.012)
Child min. age	-0.057^{***} (0.008)	-0.006 (0.007)
Active family	2.548^{***} (0.074)	-1.224^{***} (0.058)
Race	YES	YES
Zipcode	YES	YES
Allegation	YES	YES
Call screener ID	YES	YES
Reporter relationship	YES	YES
Holidays	YES	YES
Year, month, DoW	YES	YES
Observations	17,923	17,923

Note. "Investigation load" and "case load" are measured by the numbers of all active investigations and active cases (divided by 10) on a day, respectively. "Investigation load (sq)" and "caseload (sq)" are the squared terms of "investigation load" and "case load," respectively. "Referral load" is measured by the number of referrals received by the CWO on a day; "Number of call screeners" is the number of actively working call screeners on a day. "Number of children" is the number of children in the household in which maltreatment to at least one child is reported in the referral; "child min. age" is the minimum age of these children. "Active family" indicates whether the referral is associated with a family currently under investigation related to prior referrals or has been receiving ongoing services provided by the CWO. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

shift-level load is chosen to serve as a more reliable measure.

Nevertheless, even under the shift-level measure, there could be other endogeneity challenges. For example, a public awareness campaign or a traumatic event in the community may lead to an increased number of child abuse/neglect calls. This would increase the overall shift-level load, but on average these reports might be of lower risks in nature.

To further alleviate such concerns, we consider using matching and weighting methods for call screening, investigation, and case workload. This is an ongoing effort, and preliminary results are promising. In addition, we also run the main analyses by controlling or excluding weekends, summer and winter vacations, as well as excluding holidays. Results show that our main findings are robust.

4.6 Conclusion and Future Directions

This work studies the collaboration between human agents and the PRM in call screening decision-making. Our findings reveal that human agents are more likely to deviate from AI recommendations when faced with high or low workloads. Interestingly, while the PRM does not adjust for varying workloads, human agents appear to consider workload when making screening decisions, resulting in a U-shaped relationship between deviation and workload. We will next study if an increased investigation load might affect the completion and accuracy of each individual investigation. These insights suggest that human workers effectively complement the AI's recommendations by incorporating important operational considerations such as maintaining a sustainable workload.

We have conducted extensive robustness checks which demonstrate that our main findings on the workload effect and the U-shape deviation patterns are strongly robust. We are exploring causal relationships between workload and screening decisions as well as human-AI collaborations through matching and weighting.

We aim to build upon our findings to offer recommendations to enhance the collaboration between human workers and AI by (i) incorporating measures of workload, (ii) proposing strategies to effectively improve screening decisions as well as the screening-investigation workflow, thus (iii) maintaining a sustainable system load. We are working on providing evidence to support that, by incorporating workload, call screeners can improve system welfare. We plan to enrich this analysis by considering load-aware screening protocols and how they might influence not only efficiency but also screening accuracy. Preliminary simulation results show that optimizing the risk protocol thresholds for default screen-ins and screen-outs can reduce the rate of screen-ins while maintaining and reducing false positive and false negative rates. This improvement requires minimal changes to the existing information system at the CWO. Other promising avenues to effectively incorporate load in screening decisions include displaying downstream traffic, average workload per caseworker, investigation load, and average case durations to better inform call screeners of the actual business levels. Alternatively, the AI output could include ranking the risks and urgency of the incoming referral compared to incoming and existing referrals/investigations/-cases in the system. However, these approaches require a greater amount of changes to the existing system and intricate analysis of caseworkers' mental models in making screening decisions.

Broadly speaking, our work contributes to the discussion on human-AI teaming in high-stakes decision-making within organizations with budget and capacity constraints. Like the PRM used in our partner organization, many deployed AI tools are programmed for a prediction task and cannot identify or address important operational constraints. Human workers can complement AI tools by accounting for operational considerations (e.g., managing a sustainable workload). This finding again emphasizes the significance of effective human-AI collaboration; it also points to directions to enhance AI tools designed for use in high-stakes contexts.

Chapter 5

Conclusion and Future Directions

This dissertation revolves around the operational challenges encountered in liver allocation systems and the screening of child maltreatment reports. Through thorough analysis utilizing data-driven models, the study provides insightful observations and practical recommendations for effectively addressing these challenges. Key themes explored include dynamic resource allocation amidst capacity constraints and the interaction between human decision-making and technology (e.g., sophisticated surgical procedure, AI).

Effective solutions to operational challenges in public service operations require a deep understanding of methodologies and practical needs for each specific context. Below, I summarize three primary directions for future research.

First, one could explore effective operational strategy designs that incorporate human and organizational learning in decision-making. Experience-based learning is ubiquitous. Besides medical learning in SLT, surgical teams need to learn nonconventional procedures such as robot-assisted surgeries by performing them, staff members in a call center need to handle customer calls to improve their ability to resolve customer issues efficiently and courteously, and new franchisees learn to operate smoothly over time by serving customers. Such human and organizational learning, while necessary and important in the long term, may come with a short-term cost and affect other stakeholders as well as system performance. It is worthwhile to further study service systems where human or organizational learning is present. For example, improve

facility planning and operating room scheduling by considering the varied and evolving experience levels of surgeons and supporting staff as well as conventional and nonconventional surgeries.

Second, it would be helpful to study operational improvements in child welfare organizations. Through our conversations with the practitioners in child welfare services, we learned they also hope to optimize staffing decisions and downstream assignment of social workers to cases. In particular, the organization is interested in the one-caseworker model, i.e., one worker is in charge of all investigations, services, and the placement of each child abuse case. Assigning a single worker for each case enables consistent support for the family and may reduce overhead, for example, from case transfers. On the flip side, the one-caseworker model may fail to capitalize on workers' specialized skills and create conflicting incentives for workers. To fully leverage the benefit of the one-caseworker model and overcome its inherent challenges, one may adopt a data-driven model to jointly optimize workforce planning and caseworker assignment as well as supervision. Such efforts could both bolster child welfare and reduce staff turnover.

Third, exploring the nonasymptotic regime and addressing the challenge of "small data" present compelling opportunities for future research. An intriguing avenue in this regard is examining sequential decision-making under uncertainty within the nonasymptotic framework. More specifically, it is worth exploring whether the insights gained from Chapter 3, particularly those related to Multi-Armed Bandit (MAB) scenarios featuring endogenously nonstationary reward curves, extend beyond the bounds of the asymptotic regime. Such investigations hold promise for aiding decision-

makers navigating dynamic environments, particularly in scenarios requiring resource allocation over shorter to medium-term horizons. Moreover, in contemporary data-driven applications, while vast quantities of data are often available at an aggregate level, the granularity of data or its relevance to specific decision-making categories may be limited. For example, the PRM tool described in Chapter 4 relies on child-specific data input to predict the out-of-home placement risks associated with each referral. For some children, limited data is available to the PRM tool to output a well-informed prediction. Therefore, an exciting area for future exploration involves devising strategies to effectively leverage knowledge and insights from data-rich contexts to address challenges posed by “small data" situations.

Bibliographic references

Achilleos, Mayer, Mcmaster, Mirza, Buckels, & Pirenne. (2003). Encouraging results of split-liver transplantation. British journal of surgery, 85(4), 494–497. https://doi.org/10.1046/j.1365-2168.1998.00605.x

Akan, M., Alagoz, O., Ata, B., Erenay, F. S., & Said, A. (2012). A Broader View of Designing the Liver Allocation System. Operations research, 60(4), 757–770. https://doi.org/10.1287/opre.1120.1064

Akshat, S., Ma, L., & Raghavan, S. (2023). Improving broader sharing to address geographic inequity in liver transplantation. Manufacturing & service operations management, 25(4), 1509–1526.

Alagoz, O., Hsu, H., Schaefer, A. J., & Roberts, M. S. (2010). Markov decision processes: A tool for sequential decision making under uncertainty. Medical decision making, 30(4), 474–483.

Alagoz, O., Maillart, L. M., Schaefer, A. J., & Roberts, M. S. (2007a). Choosing among living-donor and cadaveric livers. Management science, 53(11), 1702–1715.

Alagoz, O., Maillart, L. M., Schaefer, A. J., & Roberts, M. S. (2007b). Determining the acceptance of cadaveric livers using an implicit model of the waiting list. Operations research, 55(1), 24–36.

Alban, A., Chick, S. E., & Zoumpoulis, S. I. (2022). Learning personalized treatment strategies with predictive and prognostic covariates in adaptive clinical trials. Available at ssrn 4160045.

Allegheny County DHS. (2024). Allegheny family screening tool [Accessed: 2024-03-19].

Anderer, A., Bastani, H., & Silberholz, J. (2022). Adaptive clinical trial designs with surrogates: When should we bother? Management science, 68(3), 1982–2002.

Anderson, G., et al. (2022). Strengths-based practice in child welfare: A systematic literature review. Child and family social work, 27(3), 304–319.

Argote, L., & Epple, D. (1990). Learning curves in manufacturing. Science, 247(4945), 920–924.

Arlotto, A., Chick, S. E., & Gans, N. (2014). Optimal hiring and retention policies for heterogeneous workers who learn. Management science, 60(1), 110–129.

Ata, B., Ding, Y., & Zenios, S. (2021). An achievable-region-based approach for kidney allocation policy design with endogenous patient choice. Manufacturing & service operations management, 23(1), 36–54.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2-3), 235–256.

Ban, G.-Y., & Keskin, N. B. (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. Management science, 67(9), 5549–5568.

Barshes, N. R., Lee, T. C., Udell, I. W., O'Mahoney, C. A., Carter, B. A., Karpen, S. J., & Goss, J. A. (2005). Adult liver transplant candidate attitudes toward graft sharing are not obstacles to split liver transplantation. American journal of transplantation. https://doi.org/10.1111/j.1600-6143.2005.00946.x

Bellman, R. (1966). Dynamic programming. Science, 153(3731), 34–37.

Berger-Tal, O., & Avgar, T. (2012). The glass is half-full: Overestimating the quality of a novel environment is advantageous. Plos one, 7(4).

Bertsimas, D., Farias, V. F., & Trichakis, N. (2011a). The price of fairness. Operations research, 59(1), 17–31.

Bertsimas, D., Farias, V. F., & Trichakis, N. (2011b). The Price of Fairness. Operations research, 59(1), 17–31. https://doi.org/10.1287/opre.1100.0865

Bertsimas, D., Farias, V. F., & Trichakis, N. (2013). Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. Operations research, 61(1), 73–87.

Bertsimas, D., & Niño-Mora, J. (2000). Restless bandits, linear programming relaxations, and a primal-dual index heuristic. Operations research, 48(1), 80–90.

Bertsimas, D., Papalexopoulos, T., Trichakis, N., Wang, Y., Hirose, R., & Vagefi, P. A. (2020). Balancing efficiency and fairness in liver transplant access: Tradeoff curves for the assessment of organ distribution policies. Transplantation, 104(5), 981–987.

Bertsimas, D., & Tsitsiklis, J. N. (1997). Introduction to linear optimization (Vol. 6). Athena Scientific Belmont, MA.

Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. Advances in neural information processing systems, 27, 199–207.

Besbes, O., Gur, Y., & Zeevi, A. (2019). Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. Stochastic systems, 9(4), 319–337.

Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press.

Cauley, R. P., Vakili, K., Fullington, N., Potanos, K., Graham, D. A., Finkelstein, J. A., & Kim, H. B. (2013). Deceased-Donor Split-Liver Transplantation in Adult Recipients: Is the Learning Curve Over? Journal of the american college of surgeons, 217(4), 672–684. https://doi.org/10.1016/J.JAMCOLLSURG.2013.06.005

CDC. (2023). Fast facts: Preventing child abuse and neglect [Accessed: 2023-06-13].

Cheng, H.-F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., Perer, A., Holstein, K., Wu, Z. S., & Zhu, H. (2022). How child welfare workers reduce racial disparities in algorithmic decisions. Proceedings of CHI 2022, 1–22.

Cheung, W. C., Simchi-Levi, D., & Zhu, R. (2019). Learning to optimize under non-stationarity. The 22nd international conference on artificial intelligence and statistics, 1079–1087.

Cheung, W. C., Simchi-Levi, D., & Zhu, R. (2020). Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. International conference on machine learning, 1843–1854.

Chick, S. E., Gans, N., & Yapar, Ö. (2022). Bayesian sequential learning for clinical trials of multiple correlated medical interventions. Management science, 68(7), 4919–4938.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration.

Philosophical transactions of the royal society b: biological sciences, 362(1481), 933–942.

Committee, O. K. T., et al. (2009). Kidney allocation concepts: Request for information. 2008.

Corno, V., Colledan, M., Dezza, M. C., Guizzetti, M., Lucianetti, A., Maldini, G., Pinelli, D., Giovanelli, M., Zambelli, M., Torre, G., & Strazzabosco, M. (2006). Extended right split liver graft for primary transplantation in children and adults. Transplant international. https://doi.org/10.1111/j.1432-2277.2006.00323.x

Cox, D. R., & Oakes, D. (2018). Analysis of survival data. Chapman; Hall/CRC.

Darr, E. D., Argote, L., & Epple, D. (1995). The acquisition, transfer, and depreciation of knowledge in service organizations: Productivity in franchises. Management science, 41(11), 1750–1762.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441(7095), 876–879.

den Boer, A. V., & Keskin, N. B. (2022). Dynamic pricing with demand learning and reference effects. Management science, 68(10), 7112–7130.

Doe, C., & Clark, D. (2020). Racial disproportionality and disparities in the child welfare system. Journal of social issues, 76(4), 765–788.

Dong, J., & Perry, O. (2020). Queueing models for patient-flow dynamics in inpatient wards. Operations research, 68(1), 250–275.

Duke Health. (2021). Duke health blog [Accessed: 2023-03-20].

Eliassen, S., Jørgensen, C., Mangel, M., & Giske, J. (2007). Exploration or exploitation: Life expectancy changes the value of learning in foraging strategies. Oikos, 116(3), 513–523.

Emre, S., & Umman, V. (2011). Split liver transplantation: An overview. Transplantation proceedings, 43(3), 884–887. https://doi.org/10.1016/j.transproceed.2011.02.036

Fecteau, A., Diamond, I. R., Song, C., Losanoff, J. E., Anand, R., Ng, V., & Millis, J. M. (2007). Impact of Graft Type on Outcome in Pediatric Liver

Transplantation. Annals of surgery, 246(2), 301–310. https://doi.org/10.1097/sla.0b013e3180caa415

Fogliato, R., De-Arteaga, M., & Chouldechova, A. (2022). A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. Available at ssrn.

Garivier, A., Ménard, P., & Stoltz, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. Mathematics of operations research, 44(2), 377–399.

Garivier, A., & Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. Arxiv preprint arxiv:0805.3415.

Garivier, A., & Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. International conference on algorithmic learning theory, 174–188.

Ge, J., Perito, E. R., Bucuvalas, J., Gilroy, R., Hsu, E. K., Roberts, J. P., & Lai, J. C. (2020). Split liver transplantation is utilized infrequently and concentrated at few transplant centers in the united states. American journal of transplantation, 20(4), 1116–1124.

Gittins, J., Glazebrook, K., & Weber, R. (2011). Multi-armed bandit allocation indices. John Wiley & Sons.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Grover, A., Markov, T., Attia, P., Jin, N., Perkins, N., Cheong, B., Chen, M., Yang, Z., Harris, S., Chueh, W., et al. (2018). Best arm identification in multi-armed bandits with delayed feedback. International conference on artificial intelligence and statistics, 833–842.

Hackl, C., Schmidt, K. M., Süsal, C., Döhler, B., Zidek, M., & Schlitt, H. J. (2018). Split liver transplantation: Current developments. World journal of gastroenterology, 24(47), 5312.

Halfin, S., & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. Operations research, 29(3), 567–588.

Hardy, G., Littlewood, J., & Polya, G. (1952). Inequalities cambridge univ. Press, cambridge, (1988).

Haussler, D., Kearns, M., Seung, H. S., & Tishby, N. (1996). Rigorous learning curve bounds from statistical mechanics. Machine learning, 25(2), 195–236.

Hu, Y., Chan, C. W., & Dong, J. (2021). Optimal scheduling of proactive service with customer deterioration and improvement. Management science.

Illeris, K. (2002). The three dimensions of learning.

Jacko, P. (2010). Restless bandits approach to the job scheduling problem and its extensions.
Modern trends in controlled stochastic processes: theory and applications, 248–267.

Jarvis, P. (2006). Towards a comprehensive theory of human learning (Vol. 1). Psychology Press.

Joulani, P., Gyorgy, A., & Szepesvári, C. (2013). Online learning under delayed feedback. International conference on machine learning, 1453–1461.

Kamath, P. S., & Kim, W. R. (2007). The model for end-stage liver disease (meld). Hepatology, 45(3), 797–805.

Kantidakis, G., Putter, H., Lancia, C., Boer, J. d., Braat, A. E., & Fiocco, M. (2020). Survival prediction models since liver transplantation-comparisons between cox models and machine learning techniques.
Bmc medical research methodology, 20, 1–14.

Kasiske, B. L., Pyke, J., & Snyder, J. J. (2020). Continuous distribution as an organ allocation framework. Current opinion in organ transplantation, 25(2), 115–121.

Keskin, N. B., & Li, M. (2021). Selling quality-differentiated products in a markovian market with unknown transition probabilities.
Available at ssrn 3526568.

Keskin, N. B., Li, Y., & Song, J.-S. (2022). Data-driven dynamic pricing and ordering with perishable inventory in a changing environment.
Management science, 68(3), 1938–1958.

Keskin, N. B., Li, Y., & Sunar, N. (2023). Data-driven clustering and feature-based retail electricity pricing with smart meters. Available at ssrn 3686518.

Keskin, N. B., & Zeevi, A. (2017). Chasing demand: Learning and earning in a changing environment. Mathematics of operations research, 42(2), 277–307.

Kim, S.-P., Gupta, D., Israni, A. K., & Kasiske, B. L. (2015). Accept/decline decision module for the liver simulated allocation model. Health care management science, 18, 35–57.

Kim, T. W., Roberts, J., Strudler, A., & Tayur, S. (2021). Ethics of split liver transplantation: Should a large liver always be split if medically safe? Journal of medical ethics.

Kondi, A., Mystakidou, K., Kostopanagiotou, G., Contis, J., Kehagias, D., Gamaletsos, E., Smyrniotis, V., & Theodoraki, K. (2005). Hemodynamic interaction between portal vien and hepatic artery flow in small-for-size split liver transplantation. Transplant international, 15(7), 355–360. https://doi.org/10.1111/j.1432-2277.2002.tb00178.x

Krishnasamy, S., Sen, R., Johari, R., & Shakkottai, S. (2016). Regret of queueing bandits. Advances in neural information processing systems, 29, 1669–1677.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1), 4–22.

Larranaga, M., Ayesta, U., & Verloop, I. M. (2013). Dynamic fluid-based scheduling in a multi-class abandonment queue. Performance evaluation, 70(10), 841–858.

Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.

Lau, L., Kankanige, Y., Rubinstein, B., Jones, R., Christophi, C., Muralidharan, V., & Bailey, J. (2017). Machine-learning algorithms predict graft failure after liver transplantation. Transplantation, 101(4), e125.

Le Morvan, P., & Stock, B. (2005). Medical learning curves and the kantian ideal. Journal of medical ethics, 31(9), 513–518.

Lee, K. W., Cameron, A. M., Maley, W. R., Segev, D. L., & Montgomery, R. A. (2008). Factors affecting graft survival after adult/child split-liver transplantation: Analysis of the UNOS/OPTN data base. American journal of transplantation, 8(6), 1186–1196. https://doi.org/10.1111/j.1600-6143.2008.02211.x

Lefrancois, G. R. (2019). Theories of human learning. Cambridge University Press.

Lehmann, E. L., & Casella, G. (2006). Theory of point estimation. Springer Science & Business Media.

Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). Optimal control. John Wiley & Sons.

Maema, A., Takayama, T., Sano, K., Sugawara, Y., Makuuchi, M., Hui, A.-M., & Imamura, H. (2003). Impaired volume regeneration of split livers with partial venous disruption: a latent problem in partial liver transplantation. Transplantation, 73(5), 765–769. https://doi.org/10.1097/00007890-200203150-00019

Marudanayagam, R., Shanmugam, V., Sandhu, B., Gunson, B. K., Mirza, D. F., Mayer, D., Buckels, J., & Bramhall, S. R. (2010). Liver retransplantation in adults: A single-centre, 25-year experience. Hpb, 12(3), 217–224.

McDiarmid, C. (1989). On the method of bounded differences. Surveys in combinatorics, 141(1), 148–188.

McDiarmid, C. (1998). Concentration. In Probabilistic methods for algorithmic discrete mathematics (pp. 195–248). Springer.

Meek, C., Thiesson, B., & Heckerman, D. (2002). The learning-curve sampling method applied to model-based clustering. Journal of machine learning research, 2(Feb), 397–418.

Michigan HHS. (2023). Children's protective services investigation process [Accessed: 2023-06-19]. https://www.michigan.gov/mdhhs/adult-child-serv/abuse-neglect/childrens/investigation/timeframes/childrens-protective-services-investigation-process

Mogul, D. B., Perito, E. R., Wood, N., Mazariegos, G. V., VanDerwerken, D., Ibrahim, S. H., Mohammad, S., Valentino, P. L., Gentry, S., & Hsu, E. (2020). Impact of acuity circles on outcomes for pediatric liver transplant candidates. Transplantation, 104(8), 1627–1632.

NHS. (2022). National liver offering scheme [Accessed: 2022-02-18].

Nitski, O., Azhie, A., Qazi-Arisar, F. A., Wang, X., Ma, S., Lilly, L., Watt, K. D., Levitsky, J., Asrani, S. K., Lee, D. S., et al. (2021). Long-term mortality risk stratification of liver transplant recipients: Real-time application of deep

learning algorithms on longitudinal data. The lancet digital health, 3(5), e295–e305.

Olberg, N., Dierks, L., Seuken, S., Slaugh, V. W., & Ünver, M. U. (2021). Search and matching for adoption from foster care. Arxiv preprint arxiv:2103.10145.

OPTN & UNOS. (2016). Split Versus Whole Liver Transplantation (tech. rep.).

OPTN & UNOS. (2022a). UNOS liver data.

OPTN & UNOS. (2022b). UNOS liver policy.

Pecora, P. J., Whittaker, J. K., Barth, R. P., Borja, S., & Vesneski, W. (2018). The child welfare challenge: Policy, practice, and research. Routledge.

Pennsylvania DHS. (2023a). Child protective services laws [Accessed: 2023-06-19].

Pennsylvania DHS. (2023b). Predictive risk modeling in child welfare in allegheny county [Accessed: 2023-10-12].

Perito, E. R., Roll, G., Dodge, J. L., Rhee, S., & Roberts, J. P. (2019). Split Liver Transplantation and Pediatric Waitlist Mortality in the United States: Potential for Improvement. Transplantation, 103(3), 552–557. https://doi.org/10.1097/TP.0000000000002249

Pusic, M. V., Boutis, K., Hatala, R., & Cook, D. A. (2015). Learning curves in health professions education. Academic medicine, 90(8), 1034–1042.

Puterman, M. L. (2014). Markov decision processes: Discrete stochastic dynamic programming. John Wiley & Sons.

Rawls, J. (1999). A theory of justice: Revised edition. Harvard university press.

Rawls, J. (2001). Justice as fairness: A restatement. Harvard University Press.

Reagans, R., Argote, L., & Brooks, D. (2005). Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. Management science, 51(6), 869–881.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the american mathematical society, 58(5), 527–535.

Rockafellar, R. T. (1970). Conjugate convex functions in optimal control and the calculus of variations. Journal of mathematical analysis and applications, 32(1), 174–222.

Sandıkçı, B., Maillart, L. M., Schaefer, A. J., Alagoz, O., & Roberts, M. S. (2008). Estimating the patient's price of privacy in liver transplantation. Operations research, 56(6), 1393–1410.

Sandıkçı, B., Maillart, L. M., Schaefer, A. J., & Roberts, M. S. (2013). Alleviating the patient's price of privacy through a partially observable waiting list. Management science, 59(8), 1836–1854.

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2020). A human-centered review of algorithms used within the us child welfare system. Proceeding of CHI 2020, 1–15.

Schumann, C., Lang, Z., Mattei, N., & Dickerson, J. P. (2019). Group fairness in bandit arm selection. Arxiv preprint arxiv:1912.03802.

Schumpeter, J., & Backhaus, U. (2003). The theory of economic development. In Joseph alois schumpeter (pp. 61–116). Springer.

Shi, P., Chou, M. C., Dai, J. G., Ding, D., & Sim, J. (2016). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. Management science, 62(1), 1–28.

Slaugh, V. (2024). Child welfare operations. In G. Berenguer & M. Sohoni (Eds.), Nonprofit operations and supply chain management: Theory and practice. Springer Nature.

Snyder, C., Keppler, S., & Leider, S. (2022). Algorithm reliance under pressure: The effect of customer load on service workers. Available at ssrn 4066823.

Spada, M., Gridelli, B., Colledan, M., Segalin, A., Lucianetti, A., Petz, W., Riva, S., & Torre, G. (2000). Extensive Use of Split Liver for Pediatric Liver Transplantation: A Single-Center Experience. Liver transplantation, 6(4), 415–428. https://doi.org/10.1053/JLTS.2000.7570

Stulberg, J. J., Huang, R., Kreutzer, L., Ban, K., Champagne, B. J., Steele, S. R., Johnson, J. K., Holl, J. L., Greenberg, C. C., & Bilimoria, K. Y. (2020). Association between surgeon technical skills and patient outcomes. Jama surgery, 155(10), 960–968.

Su, X., & Zenios, S. A. (2006). Recipient Choice Can Address the Efficiency-Equity Trade-off in Kidney Transplantation: A Mechanism Design Model. Management science, 52(11), 1647–1660. https://doi.org/10.1287/mnsc.1060.0541

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Tang, Y., Scheller-Wolf, A., Tayur, S., Perito, E., & Roberts, J. (2019). Split liver transplantation: A decision support tool. Ssrn.

Tang, Y. S., Li, A., Scheller-Wolf, A. A., & Tayur, S. R. (2021). Multi-armed bandits with endogenous learning curves: An application to split liver transplantation. Available at ssrn 3855206.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4), 285–294.

Tunç, S., Sandıkçı, B., & Tanrıöver, B. (2022). A simple incentive mechanism to alleviate the burden of organ wastage in transplantation. Management science.

UNOS. (2020, August). United networks of organ sharing data.

UNOS. (2021). System notice: Liver and intestinal organ distribution based on acuity circles implemented feb. 4. 2020.

Van Loan, C. (1978). Computing integrals involving the matrix exponential. Ieee transactions on automatic control, 23(3), 395–404.

Vulchev, A., Roberts, J. P., & Stock, P. G. (2004). Ethical issues in split versus whole liver transplantation. https://doi.org/10.1111/j.1600-6143.2004.00630.x

Whitt, W. (2006). Fluid models for multiserver queues with abandonments. Operations research, 54(1), 37–54.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. Journal of applied probability, 25(A), 287–298.

Wilson, F. (2019). Aging out of foster care: A national comparison of homelessness and unemployment risks. Youth studies quarterly, 38(1), 22–34.

Wojcicki, M., Milkiewicz, P., & Silva, M. (2008, October). Biliary tract complications after liver transplantation: A review. https://doi.org/10.1159/000144653

Wojcicki, M., Silva, M. A., Jethwa, P., Gunson, B., Bramhall, S. R., Mayer, D., Buckels, J. A., & Mirza, D. F. (2006). Biliary complications following adult right lobe ex vivo split liver transplantation. Liver transplantation, 12(5), 839–844. https://doi.org/10.1002/lt.20729

Xie, J., Zhu, T., Chao, A.-K., & Wang, S. (2017). Performance analysis of service systems with priority upgrades. Annals of operations research, 253(1), 683–705.

Yang, L., Yang, J., & Ren, S. (2021). Contextual bandits with delayed feedback and semi-supervised learning (student abstract). Proceedings of the aaai conference on artificial intelligence, 35(18), 15943–15944.

Zenios, S. A., Chertow, G. M., & Wein, L. M. (2000a). Dynamic allocation of kidneys to candidates on the transplant waiting list. Operations research, 48(4), 549–569.

Zenios, S. A., Chertow, G. M., & Wein, L. M. (2000b). Dynamic allocation of kidneys to candidates on the transplant waiting list. Operations research, 48(4), 549–569.

Zenios, S. A., Chertow, G. M., & Wein, L. M. (2003). Dynamic Allocation of Kidneys to Candidates on the Transplant Waiting List. Operations research, 48(4), 549–569. https://doi.org/10.1287/opre.48.4.549.12418

Appendix A

Appendix for Chapter 2

A.1 Alternative Transplantation Objectives

This section presents the analytical results/formulation for three alternative transplantation objectives: minimizing the total number of patient deaths (TNPD), minimizing the number of patient deaths after transplant (NPDAT), and minimizing organ wastage (OW).

A.1.1 TNPD and NPDAT

While we focused on NPDWT and QALY objectives in the main text, our fluid limit decomposition method can be applied to other commonly used transplantation objectives. For TNPD (2.10) and TNPDAT (2.11) objectives defined in Section 2.3, Proposition A.1.1 presents the closed-form objective functions $\mathscr{U}^{TNPD}$, $\mathscr{U}^{NPDAT}$ which are parts of the decomposed LPs that give optimal decision rules, respectively.

Proposition A.1.1. When $\mathbf{\Psi}$ is non-singular, $\mathscr{U}^{TNPD}$ and $\mathscr{U}^{NPDAT}$ can be simplified to

$$\mathscr{U}_t^{\text{TNPD}} = \mathbf{d}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \left(\sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell\right) + \sum_\ell \boldsymbol{\zeta}^\ell \mathbf{P}^\ell \mathbf{u}^\ell + \bar{\boldsymbol{\zeta}}^\ell \bar{\mathbf{P}}^\ell \mathbf{s}^\ell \tag{A.1}$$

$$\mathscr{U}_t^{\text{NPDAT}} = \sum_\ell \boldsymbol{\zeta}^\ell \mathbf{P}^\ell \mathbf{u}^\ell + \bar{\boldsymbol{\zeta}}^\ell \bar{\mathbf{P}}^\ell \mathbf{s}^\ell \tag{A.2}$$

Note that $\mathscr{U}_t^{\text{TNPD}} = \mathscr{U}_t^{\text{NPDAT}} + \mathscr{U}_t^{\text{NPDWT}}$, because by definition, the total number of patient deaths equals the number of patient deaths while waiting for transplants plus the number of patient deaths during and after transplants.

A.1.2 Minimizing Organ Wastage

Another important transplantation objective is to minimize organ wastage (OW).

$$\min_{(\mathbf{U},\mathbf{S})} \quad \text{OBJ}^{\text{OW}} := \sum_{\ell} \int_0^T \Big((1 - \mathbf{P}^{\ell}) \mathbf{u}^{\ell}(t) + (1 - \bar{\mathbf{P}}^{\ell}) \mathbf{s}^{\ell}(t) \Big) \, dt \tag{A.3}$$

Solving a fluid optimization problem with (A.4) as the objective is equivalent to maximizing the dynamic index $\mathscr{U}_t^{\text{OW}}$ for all $t \in \mathcal{T}$:

$$\mathscr{U}_t^{\text{OW}} := \sum_{\ell} \mathbf{P}^{\ell} \mathbf{u}^{\ell}(t) + \bar{\mathbf{P}}^{\ell} \mathbf{s}^{\ell}(t) \tag{A.4}$$

A.1.3 Multi-Objective Framework with More Than Two Objectives

When we have $m \in \mathbb{N}$ $(m > 2)$ objectives in a multi-objective framework, we assign each maximizing objective a non-positive weight and each minimizing objective a non-negative weight. Denote $\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^m$ as the weights for objectives $\text{OBJ}^1, \ldots, \text{OBJ}^m$. Let $\mathbf{w}^1, \ldots, \mathbf{w}^m$ be the weight of penalties for violating (soft) fairness constraints. If we set one of $\mathbf{w}^k, k \in [m]$ to be positive infinity, then we have a hard fairness constraint. The soft-constraint multi-objective fluid optimization problem is as follows:

$$\min_{(\mathbf{U},\mathbf{S}),\xi \in \mathcal{F}} \quad \text{OBJ}^{\text{Multi}} := \sum_{k=1}^{m} \boldsymbol{\eta}^k \text{OBJ}^k + \left(\sum_{k=1}^{m} \boldsymbol{\eta}^k \mathbf{w}^k \right)^{\top} \int_{t=0}^{T} \boldsymbol{\xi}(t) dt \tag{A.5}$$

$$s.t. \quad (2.2) \sim (2.5), (2.7), (2.12) \tag{A.6}$$

In liver transplantation, reducing (pre-transplant, post-transplant, and total) patient deaths and maximizing QALY are well-accepted objectives. However, these transplantation objectives are not often aligned. When using the multi-objective formulation, it is important to adjust the scale accordingly, which entails estimating the QALY equivalent of saving sick patients from dying. Moreover, this formulation allows us to incorporate input from the transplantation community in choosing the weight parameters.

A.2 Extensions to the Fluid Models in Section 2.3

This section discusses extensions to the fluid optimization problems we studied in the main texts. Our fluid model is generic and compatible with various extensions.

A.2.1 Probability of Getting Transplants and Alternative Fairness Formulation

In Sections 2.3 and 2.4, we adopt the max-min fairness concept from Rawls, 1999 and enforce fairness by lower-bounding the probabilities of getting a transplant (before leaving the queue) for patients from different groups. Our fluid limit decomposition method works as well with alternative fairness formulations in Zenios et al., 2000b based on different fairness notions (e.g. envy-freeness). More generally, our fluid decomposition method works for any fairness constraints that are linear in $(\mathbf{U}(t), \mathbf{S}(t))$ and $\mathbf{x}(t)$, or any fairness objective function that is polynomial in $(\mathbf{U}(t), \mathbf{S}(t))$ and $\mathbf{x}(t)$.

Generally speaking, the choices of fairness concepts (e.g. max-min, envy-free, etc.) depend on how well they suit the problem context. Some fairness concepts may not be of the first-order importance in our liver transplantation contexts: for example, an adult patient may not mind if a child gets a liver before she does, as long as she knows that she has a good chance of getting a satisfactory liver offer for transplant before too long. Thus, max-min fairness may be more appropriate in liver allocation. We confirmed this conjecture through conversations with transplant surgeons: The key fairness concern is a reasonable chance of getting a transplant for EVERY candidate group in liver allocation. Thus, we choose the max-min fairness notion and choose linear soft/hard constraints to incorporate fairness on a group level.

Another commonly used fairness formulation is the cumulative max-min fairness. Specifically, we may use the following constraints instead of (2.6):

$$\int_{t=0}^{T} \sum_{\ell} \mathbf{u}^{\ell}(t) + \sum_{\ell} \mathbf{Z}\mathbf{v}^{\ell}(t) \geq \mathbf{\Theta}\boldsymbol{\lambda} T \quad \forall t.$$

In the interior case, the optimal solution to the original fluid optimization problem (2.1) $\sim$ (2.7) is equivalent to:

$$\min_{(\mathbf{U},\mathbf{S}),\xi\in\mathcal{F}} \quad \int_{t=0}^{T} \mathbf{d}^{\top} \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \left(\sum_{\ell} \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) + \bar{\mathbf{P}}^{\ell}\mathbf{s}^{\ell}(t)\right) \tag{A.7}$$

$$s.t. \quad \mathbf{u}^{\ell}(t), \mathbf{s}^{\ell}(t) \geq 0 \quad \forall \ell, t \in [0,T] \tag{A.8}$$

$$\mathbf{1}_{I^{\cdot}J}\mathbf{u}^{\ell}(t) + \mathbf{1}_{I^2\times J^2}\mathbf{s}^{\ell}(t) \leq \boldsymbol{\mu}^{\ell}(t) \quad \forall \ell, t \in [0,T] \tag{A.9}$$

$$\mathbf{1}_{I^2\cdot J^2}\mathbf{s}^{\ell}(t) \leq \bar{\boldsymbol{\mu}}^{\ell}(t) \quad \forall \ell, t \in [0,T] \tag{A.10}$$

$$\int_{t=0}^{T} \sum_{\ell} \mathbf{u}^{\ell}(t) + \sum_{\ell} \mathbf{Z}\mathbf{v}^{\ell}(t) \geq \mathbf{\Theta}\boldsymbol{\lambda} T \quad \forall t \tag{A.11}$$

Because (A.11) requires the cumulative livers allocated to each patient class to be greater than some predefined proportions, we cannot directly decompose the fluid optimization problem to greedy decision rules. Nevertheless, the transformed fluid optimization problem (A.7) $\sim$ (A.11) contains only one linking constraint (A.11) while the others are standard nonnegative and capacity constraints. Moreover, we can still find a boundary case allocation policy by adding (2.26) to the constraints. We concentrate on the interior case in this paper, yet it is likely in some special scenarios in the boundary case, (A.7) $\sim$ (A.11) plus (2.26) gives an optimal solution to the original fluid optimization problem (2.1) $\sim$ (2.7). The same analysis applies to alternative transplantation objectives.

The cumulative max-min fairness constraint formulation is indeed practical, mainly because most organizations periodically review system performance on efficiency and fairness. Given the same fairness levels, solutions satisfying our stricter fairness constraints automatically meet the cumulative fairness requirement. Moreover, when the review period in the cumulative constraint is small enough, our fairness definition becomes very close to the cumulative fairness constraint.

In any case, the objective values obtained under our any-t fairness formulation are lower bounds for maximization problems or upper bounds for minimization problems. Specifically, we showed the benefit of wider use of SLT under the “sickest first" allocation rules by comparing “all-split, sickest first" with “few split, sickest first", an approximation of the current OPTN policy. Moreover, “optimal split, optimal allocation" and “all split, optimal allocation" consistently outperform “no-split" policies. Particularly, we showed that the “all-split, optimal allocation"

policy is optimal given only 10% livers are splittable. And ‘all-split, optimal allocation" under stricter fairness constraints performs better than “no-split" policies with no fairness constraints.

Our fairness formulation is intuitive and patient-centered, as it implies that the timing within a review period at which a candidate arrives to the waitlist should not affect their minimum probability of getting a transplant. Although one may think it can be a strict fairness requirement, it is very applicable. Figures 2.3 and 2.4 show that our stricter fairness criteria can be met with a small PoF through broader SLT use and optimally allocating organs to recipients. Moreover, one can choose the proper parameter for the fairness notion: Enforcing a 68% fairness with our definition is not necessarily stricter than maintaining a 75% fairness with the cumulative fairness level.

A.2.2 Patient Strategic Behaviors: Multiple Listing

We did not explicitly consider patient strategic behaviors in the base formulation, such as *multiple listing* and endogenous accept/reject decisions in our baseline model. Below, we show that modifying the patient arrival parameters to address multiple listing suffices. Moreover, we can slightly modify our fluid models to incorporate endogenous patient choices as functions of our allocation policy.

Multiple listing (also known as *Multi-listing*) refers to the same transplant candidate listing themselves at multiple TCs across different geographical locations, in the hope that they could get a transplant earlier at one of these TCs when a local donor liver becomes available. In our fluid model, we assumed that there is no

multi-listing, i.e., When allowing multi-listing, there are two cases: In the first case, a patient multi-listed at multiple TCs, but we would categorize him into the same patient group in any one of their listed TC. In this case, it makes no difference whether this patient is multi-listed or not in the fluid model, as this patient belongs to the same patient group. In the second case, a patient may have listed themselves at two TCs and may belong to two patient groups simultaneously. We argue that this case can be fully captured by our fluid model as well, because our fluid model assumes that all patient queues are non-empty, and we do not differentiate patients within the same patient group. Our parameter estimates may change when we factor multi-listing into our model to avoid double counting. That being said, for individual candidates, by multi-listing their chances of getting a transplant earlier increase (as they get more options); from the central planner's perspective, exactly who in the patient group gets the liver is not a first-order concern and re-estimation of parameters are needed to capture the expected percentages of multi-listing.

A.2.3 Patient Strategic Behaviors: Endogenous Accept/Reject Decision

A.2.3.1 Related Work.

Transplant candidates' accept/reject decisions on liver offers depend on several factors: their health conditions (e.g. they may need an immediate transplant to sustain life), the suitability of the donor liver (e.g. size and blood/tissue type matching), the cold ischemia time (i.e. time elapsed since harvesting the liver from its deceased donor, the shorter the better), the quality of the liver (e.g. young, healthy donors died from accidents are usually preferred), and the anticipation of future offers (e.g. they may choose to wait for a better offer).

In the literature, researchers have studied patient strategic decisions using discrete, infinite-horizon Markov decision processes (MDPs). For example, Alagoz et al., 2007b modeled patient accept/decline decision using an MDP and summarized patient decisions based on a) the patient's current and likely future health conditions, b) the current liver offer, and c) the patient's current and future prospects for organ offers. Alagoz et al., 2007a proceeded with structural analysis and Sandıkçı et al., 2008 further found that having a more transparent waiting list helps candidates make better accept/reject decisions. Sandıkçı et al., 2013 formed a partially observable Markov decision process (POMDP) for each candidate, as patient future offer prospects could only be estimated based on aggregated information about the waitlists. In Subsection 2.5.3, we also briefly discussed related papers using game-theoretic analysis, reduced models, and simulation-based structural models. None of the existing papers solves the analytical, dynamic, and transient optimal organ allocation problem in the presence of patient strategic accept/reject decisions.

While patients' anticipation of future organ offers may be modeled analytically, rigorous empirical studies are needed to investigate whether candidates' accept/reject decisions are truly endogenous and how patients respond to various organ allocation policies.

A.2.3.2 Proof for Subsection 2.5.3.

Below we show how we can apply the fluid decomposition technique to a case where $\mathbf{P}$ and $\bar{\mathbf{P}}$ are functions of $(\mathbf{U}(t), \mathbf{S}(t))$, i.e., $\mathbf{P}(\mathbf{U}(t), \mathbf{S}(t))$ and $\bar{\mathbf{P}}(\mathbf{U}, \mathbf{S})$. Fluid

dynamics equation (2.7) becomes

$$\dot{\mathbf{x}}(t) = \boldsymbol{\lambda} - \sum_{\ell} \mathbf{P}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{u}^{\ell}(t) - \bar{\mathbf{P}}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{s}^{\ell}(t) + \boldsymbol{\Psi}\mathbf{x}(t) \quad \forall t \tag{A.12}$$

(2.18) becomes

$$F(\mathbf{U}(\tau), \mathbf{S}(\tau)) := \lambda - \sum_{\ell} \left(\mathbf{P}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{u}^{\ell}(\tau) + \bar{\mathbf{P}}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{s}^{\ell}(\tau) \right). \tag{A.13}$$

Note that (2.17) holds regardless of the explicit form of $F(\mathbf{U}(\tau), \mathbf{S}(\tau))$. Also, (2.22) $\sim$ (2.24) all hold under (A.13) except that $\mathbf{P}$ and $\bar{\mathbf{P}}$ are replaced with the general $\mathbf{P}(\mathbf{U}(t), \mathbf{S}(t))$ and $\bar{\mathbf{P}}(\mathbf{U}(t), \mathbf{S}(t))$. Therefore, we can replace Proposition 3 with the following Proposition A.2.1:

Proposition A.2.1. When $\boldsymbol{\Psi}$ is non-singular and replacing (2.7) with (A.13), $\mathscr{U}_t^{NPDWT}$ and $\mathscr{U}_t^{QALY}$ can be written as

$$\mathscr{U}_t^{\mathrm{NPDWT}} = \mathbf{d}^{\top} \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I} \right) \boldsymbol{\Psi}^{-1} \left(\sum_{\ell} \mathbf{P}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{u}^{\ell}(t) + \bar{\mathbf{P}}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{s}^{\ell}(t) \right) \tag{A.14}$$

$$\mathscr{U}_t^{\mathrm{QALY}} = -\mathbf{q}^{\top} \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I} \right) \boldsymbol{\Psi}^{-1} \left(\sum_{\ell} \mathbf{P}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{u}^{\ell}(t) + \bar{\mathbf{P}}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{s}^{\ell}(t) \right) \tag{A.15}$$

$$+ \sum_{\ell} \mathbf{O}^{\ell}\mathbf{P}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{u}^{\ell}(t) + \bar{\mathbf{O}}^{\ell}\bar{\mathbf{P}}^{\ell}(\mathbf{U}(t), \mathbf{S}(t))\mathbf{s}^{\ell}(t)$$

Proposition A.2.1 shows that the fluid decomposition technique applies to a fluid model with any integrable functions $\mathbf{P}(\mathbf{U}(t), \mathbf{S}(t))$ and $\bar{\mathbf{P}}(\mathbf{U}(t), \mathbf{S}(t))$. This result is powerful, as we are able to remove the differential constraints and reduce the fluid

optimization problem over a continuous time horizon to simpler math programs (not necessarily linear programs, given a general $\mathbf{P} \notin \bar{\mathbf{P}}$) to solve optimal decision rules with the presence of endogenous and strategic patient choices. Corollary A.2.1 illustrates a case where $\mathbf{P}(\mathbf{U}(t), \mathbf{S}(t))$ and $\bar{\mathbf{P}}(\mathbf{U}(t), \mathbf{S}(t))$ are linear in $(\mathbf{U}(t), \mathbf{S}(t))$.

Corollary A.2.1. If $\mathbf{P}^\ell(\mathbf{U}(t), \mathbf{S}(t)) = \mathbf{p}_0 + \mathbf{p}_1 \cdot \mathbf{u}^\ell(t)$ and $\bar{\mathbf{P}}^\ell(\mathbf{U}(t), \mathbf{S}(t)) = \bar{\mathbf{p}}_0 + \bar{\mathbf{p}}_1 \cdot \mathbf{s}^\ell(t)$, $\mathscr{U}_t^{NPDWT}$ and $\mathscr{U}_t^{QALY}$ are quadratic programs, more specifically

$$\mathscr{U}_t^{\text{NPDWT}} = \mathbf{d}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \left(\sum_\ell (\mathbf{p}_0 + \mathbf{p}_1 \cdot \mathbf{u}^\ell(t))\mathbf{u}^\ell(t) + (\bar{\mathbf{p}}_0 + \bar{\mathbf{p}}_1 \cdot \mathbf{s}^\ell(t))\mathbf{s}^\ell(t) \right) \tag{A.16}$$

$$\begin{aligned} \mathscr{U}_t^{\text{QALY}} = -\mathbf{q}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \left(\sum_\ell (\mathbf{p}_0 + \mathbf{p}_1 \cdot \mathbf{u}^\ell(t))\mathbf{u}^\ell(t) + (\bar{\mathbf{p}}_0 + \bar{\mathbf{p}}_1 \cdot \mathbf{s}^\ell(t))\mathbf{s}^\ell(t) \right) \\ + \sum_\ell \mathbf{O}^\ell(\mathbf{p}_0 + \mathbf{p}_1 \cdot \mathbf{u}^\ell(t))\mathbf{u}^\ell(t) + \bar{\mathbf{O}}^\ell(\bar{\mathbf{p}}_0 + \bar{\mathbf{p}}_1 \cdot \mathbf{s}^\ell(t))\mathbf{s}^\ell(t) \end{aligned} \tag{A.17}$$

Corollary A.2.1 is obtained by directly applying Proposition A.2.1. We can also easily arrive at the conclusion that if $\mathbf{P}(\mathbf{U}(t), \mathbf{S}(t))$ and $\bar{\mathbf{P}}(\mathbf{U}(t), \mathbf{S}(t))$ are polynomials of degree n_1 and n_2 respectively, then $\mathscr{U}_t^{NPDWT}$ and $\mathscr{U}_t^{QALY}$ are math programs with polynomial objective functions of degree $n_1 + 1$ and $n_2 + 1$, respectively.

Notably, our fluid decomposition technique removes (2.7), placing the term in (2.7) that is independent of $\mathbf{x}$ and $\dot{\mathbf{x}}$, i.e., $F(\cdot)$, into the decomposed math programs' objective functions. If $F(\cdot)$ is a black-box mapping or a more complex form, solving

the decomposed math programs may require advanced optimization methods and techniques well beyond the scope of this paper.

We note that, we do not solve for the exact endogenous form of patients' strategic accept/reject decisions. Our work is complementary to existing work which either formulates the patient MDP accept/reject decisions (Alagoz et al., 2007a) for dynamic decision making or studies the system equilibria (Tunç et al., 2022) for steady-state analysis. In other words, these papers propose the exact forms of $\mathbf{P}(\mathbf{U}(t), \mathbf{S}(t))$ and $\bar{\mathbf{P}}(\mathbf{U}(t), \mathbf{S}(t))$; while our fluid decomposition technique can evaluate policy outcomes for any such forms as long as $F(\cdot)$ is an integrable function.

A.2.4 Sequential Organ Offering and Provisional Offers

In practice, cadaveric whole livers are offered to waitlisted candidates sequentially. Below we show that sequential liver offering can be easily factored into the fluid model. Suppose a type-ℓ liver is offered to n patients of type ij and the liver is eventually accepted. Each ij-patient's decision is independent of others and with probability π_{ij}^{ℓ}, they accept a type-ℓ whole liver offer. Given these notations, we have

$$\mathbf{P}_{ij}^{\ell} = 1 - (1 - \pi_{ij}^{\ell})^n \tag{A.18}$$

For split livers offered sequentially to n_1 type-ij candidates and n_2 type-$i'j'$ candidates. Each ij-patient's decision is independent of others and with probability $\bar{\pi}_{ij}^{\ell}$, they accept a type-ℓ split liver offer. There are four possible cases:

- Case 1: All patients reject type-ℓ split liver offers. Pr(Case 1) =

$(1-\bar{\pi}^{\ell}_{ij})^{n_1}(1-\bar{\pi}^{\ell}_{i'j'})^{n_2}$.

- Case 2: At least one type-ij patient accepts a type-ℓ split liver while all type-$i'j'$ patients reject. $\Pr(\text{Case 2}) = \left[1-(1-\bar{\pi}^{\ell}_{ij})^{n_1}\right](1-\bar{\pi}^{\ell}_{i'j'})^{n_2}$.

- Case 3: At least one type-$i'j'$ patient accepts a type-ℓ split liver while all type-ij patients reject. $\Pr(\text{Case 3}) = (1-\bar{\pi}^{\ell}_{ij})^{n_1}\left[1-(1-\bar{\pi}^{\ell}_{i'j'})^{n_2}\right]$.

- Case 4: Both type-ℓ split livers are accepted. $\Pr(\text{Case 4}) = 1-\sum_{i=1}^{3}\Pr(\text{Case } i)$.

We define $\bar{\mathbf{P}}^{\ell}_{ij,i'j'}$ as the probability of at least one partial liver accepts, i.e.,

$$\bar{\mathbf{P}}^{\ell}_{ij,i'j'} = 1-(1-\bar{\pi}^{\ell}_{ij})^{n_1}(1-\bar{\pi}^{\ell}_{i'j'})^{n_2} \tag{A.19}$$

In Case 1, the liver is usually wasted. In Cases 2 and 3, the liver is likely transplanted to the accepting candidate, using a reduced-size liver transplantation (RLT) technique if necessary.

In some scenarios, a *provisional offer* (i.e. an organ offer before other patients assigned previously declining the same organ), may be extended in order to reduce organ wastage. We can incorporate provisional offers in our fluid model by choosing the corresponding n^P, n_1^P and n_2^P that capture the number of total organ offers including provisional offers. No further changes are needed, because eventually, the organ is offered sequentially.

A.2.5 Broader Geographic Sharing

As UNOS is moving toward a more continuous allocation scoring system (Bertsimas et al., 2020; Kasiske et al., 2020), one important extension to the fluid model is enabling geographic sharing. We can easily incorporate geographic sharing by setting liver and patient categories that include larger geographical regions within the same group. Note that we cannot guarantee a strictly continuous, boundaryless sharing, as the liver and patient types are categorical by nature in the fluid model; but we can be very close to a continuous one by carefully choosing the parameters that define different liver and patient categories.

A.2.6 Retransplantation

The retransplantation rate is typically below 10% in liver transplantation (Marudanayagam et al., 2010). Our base model assumes that all patients who accept whole or partial livers leave the waitlist system, regardless of the surgery outcomes: The base model does not explicitly include retransplantation in our model formulation as the retransplantation rate is below 10% (Marudanayagam et al., 2010). Nevertheless, we can easily incorporate retransplantation in our fluid model. The only change is to add $\mathbf{T}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{T}}\mathbf{s}^\ell(t)$ on the right hand side of (2.7), where $\mathbf{T}^\ell \in \mathbb{R}^{I \cdot J}$ and $\bar{\mathbf{T}}^\ell \in \mathbb{R}^{I^2 \cdot J^2}$ are the retransplantation probability matrices for WLT and SLT, respectively. In other words, we replace (2.7) with (A.20):

$$\dot{\mathbf{x}}(t) = \boldsymbol{\lambda} - \sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) - \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t) + \boldsymbol{\Psi}\mathbf{x}(t) + \mathbf{T}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{T}}\mathbf{s}^\ell(t) \quad \forall t \tag{A.20}$$

A.2.7 Medical Learning and SLT Expertise

Despite SLT's potential to relieve the acute shortage of donated livers in the US, it is underused in part because few surgeons in the US have learned to perform SLT. One barrier for young surgeons to acquire the skills to perform SLT is the need to perform actual SLT surgeries to become proficient, and the lower success rate such early surgeries have. Further, because SLT is a delicate operation, even with practice, some medical teams may still have only mixed success.

Based on the fluid model and the fluid limit decomposition method described in this work, Y. S. Tang et al., 2021 study the donated liver allocation problem in a setting where surgeons with different potential abilities may learn SLT, becoming skilled over time. The authors formulate a multi-armed bandit (MAB) model, in which learning curves are embedded in the reward functions, to address the trade-off between discovering and developing talents (exploration) and utilizing a defined group of already-skilled surgeons (exploitation). To solve their MAB learning model, Y. S. Tang et al., 2021 propose the L-UCB, FL-UCB, and QFL-UCB algorithms, all variants of the upper confidence bound (UCB) algorithm, enhanced with additional features such as learning, fairness, queueing dynamics (decomposed fluid limits), and arm dependence. They prove that the regrets of the proposed algorithms, that is, the loss in total rewards due to lack of information about surgeons' aptitudes, are bounded by $O(\log t)$. They also show that the proposed algorithms have superior numerical performance compared to standard bandit algorithms in settings where learning exists. Y. S. Tang et al., 2021 provide insights into potential strategies to increase the proliferation of SLT and other technically-difficult medical procedures.

A.3 Sufficient Conditions for the Interior Case

The patient fluid queue lengths in our SLT context are likely always nonempty for several reasons. First, the national liver waitlists are overloaded. Second, the fluid model is a first-order approximation based on FLLN. It is most suitable for strategic planning on a high level with broad patient classes. That entails the proper estimation of model parameters and careful choices of granularity levels.

A sufficient condition for our original fluid optimization problem to stay in the interior of the state space is

$$\lambda_{ij}(t) \geq \sum_{\ell \in \mathcal{L}} \left(P^{\ell}_{ij} \bar{u}^{\ell}_{ij}(t) + \sum_{i',j'} \left(\bar{P}^{\ell}_{ij,i'j'} \bar{s}^{\ell}_{ij,i'j'}(t) + \bar{P}^{\ell}_{i'j',ij} \bar{s}^{\ell}_{i'j',ij}(t) \right) \right) \tag{A.21}$$

In (A.21), $\bar{u}^{1:L}_{ij}(t)$, $\bar{s}^{1:L}_{ij,i'j'}(t)$ and $\bar{s}^{1:L}_{i'j',ij}(t)$ are solved for each $i \in \mathcal{I}$ and $j \in \mathcal{J}$ to the following optimization problem for all $t \in \mathcal{T} \setminus \{T\}$, i, and j:

$$\max_{(\mathbf{U}(t),\mathbf{S}(t))} \quad \sum_{\ell \in \mathcal{L}} \left(P^{\ell}_{ij} \bar{u}^{\ell}_{ij}(t) + \sum_{i',j'} \left(\bar{P}^{\ell}_{ij,i'j'} \bar{s}^{\ell}_{ij,i'j'}(t) + \bar{P}^{\ell}_{i'j',ij} \bar{s}^{\ell}_{i'j',ij}(t) \right) \right) \tag{A.22}$$

$$\bar{\mathbf{u}}^{\ell}(t), \bar{\mathbf{s}}^{\ell}(t) \geq 0 \quad \forall \ell, t \in [0,T] \tag{A.23}$$

$$\mathbf{1}_{I.J} \bar{\mathbf{u}}^{\ell}(t) + \mathbf{1}_{I^2 \times J^2} \bar{\mathbf{s}}^{\ell}(t) \leq \boldsymbol{\mu}^{\ell}(t) \quad \forall \ell, t \in [0,T] \tag{A.24}$$

$$\mathbf{1}_{I^2.J^2} \bar{\mathbf{s}}^{\ell}(t) \leq \bar{\boldsymbol{\mu}}^{\ell}(t) \quad \forall \ell, t \in [0,T] \tag{A.25}$$

$$\sum_{\ell} \bar{\mathbf{u}}^{\ell}(t) + \sum_{\ell} \mathbf{Z} \bar{\mathbf{s}}^{\ell}(t) \geq \boldsymbol{\Theta} \boldsymbol{\lambda}(t) \quad \forall t \tag{A.26}$$

The sufficient condition (A.21) says that the maximum number successful surgeries that can be performed for any wait list cannot meet the incoming demand, subject to capacity and fairness constraints.

A.4 Application of Our Results to WLT and Kidney Allocation

A.4.1 Explicit Dynamic Indexes for the Optimal Policy in LT

We first note that if we set $\bar{\boldsymbol{\mu}}^{\ell} = 0$ for all $\ell \in \mathcal{L}$ in (2.5), then $\mathbf{s}^{\ell} = 0, \forall \ell \in \mathcal{L}$ and therefore the fluid optimization problem (2.1) $\sim$ (2.7) reduces to the fluid optimization problem (P_κ) in Akan et al., 2012 with $\kappa = 0$. Thus our results yield an exact and explicit solution for solving (P_κ) with $\kappa = 0$ in Akan et al., 2012 in the interior case. In a multi-objective framework, i.e., $\kappa \in [0, 1]$ in the optimization problem P_κ of Akan et al., 2012, solving (P_κ) is equivalent to solving the reduced LP (2.34) $\sim$ (2.35) with $\bar{\boldsymbol{\mu}}^{\ell} = 0, \forall \ell$.

Akan et al., 2012 derived the dual control problem of (2.1) $\sim$ (2.7) and pointed out that the optimal policy could be characterized by the dual solution, i.e., the *shadow prices*. From there, they showed that the optimal policy of the primal problems are dynamic index policies, maximizing indexes that are functions of the shadow prices. However, the dual control problems which give the indexes are neither smaller or easier than the primal ones, as they contain ordinary differential inequalities in the constraints in addition to linking constraints. Moreover, to obtain the dual solutions, one needs to search through the entire primal dual spaces, for the interior case and the boundary case. Our result show that in the interior of the primal problem's state space, the optimal allocation policy greedily optimizes over decomposed objectives (the dynamic indexes), which compactly summarize the contribution of each action to the overall objective value.

A.4.2 Structural Properties and Explicit Solutions to Fluid Models in Kidney Allocation

Using our results, the optimal policy of the fluid model that Zenios et al., 2000b used (see their Equation (19)) to describe the kidney allocation system can likewise be decomposed and reduced to standard quadratic programs (QPs). Equity objectives in their work are incorporated into the objective function, instead of appearing in the constraints: This formulation is equivalent to modifying our soft-constraint single-efficiency objective (QALY) with $\boldsymbol{\Theta} = 0$.

We show that there exists a closed-form expression for their objective (19) and using our fluid-limit decomposition technique, we can find the optimal solution of the proposed fluid model with objective (19) in the interior of the state space. Specifically, we can decompose and reduce the complex fluid control problem to standard QPs, which are solved in polynomial time when the QPs are convex. The optimal policy consists of decision rules that optimize over explicit and finite-dimensional dynamic indexes for each $t \in \mathcal{T}$, and are easy to describe and implement. Moreover, we find analytically tractable optimal decision rules and optimal policies without assumptions or approximations in the interior case.

The quadratic term in Zenios et al., 2000b's Equation (19) results from their choice of fairness objective. Generally speaking, the choices of fairness concepts (e.g. max-min, envy-free, etc.) should best fit the problem context: The formulation used to mathematically translate these concepts should consider the accuracy in describing the concept, elegance in design/formulation, and computational efficiency.

Below we show the proof. In notation consistent with our previous definitions and compatible with theirs, solving the QP below gives the optimal decision rules at each $t \in \mathcal{T}$:

$$\begin{aligned}\max_{\mathbf{U}(t)} \sum_{\ell} & \gamma^{\top} \tilde{D}\mathbf{u}^{\ell}(t) - \beta h^{\top} \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1}\mathbf{P}^{\ell}\mathbf{u}^{\ell}(t) \\ & - (1-\beta) \int_{t}^{T} \left[\exp((\tau - t)\boldsymbol{\Psi})(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t))\right]^{\top} R \left(\exp((\tau - t)\boldsymbol{\Psi})(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t))\right) d\tau \\ & - (1-\beta) \left[\exp[t\boldsymbol{\Psi}]\mathbf{x}(0)\right]^{\top} R \left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1}(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}) \\ & - (1-\beta) \left[\left(\exp((T-t)\boldsymbol{\Psi}) - \mathbf{I}\right) \boldsymbol{\Psi}^{-1}(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell})\right]^{\top} R \exp[t\boldsymbol{\Psi}]\mathbf{x}(0) \end{aligned} \tag{A.27}$$

$$s.t.\ (2.3), (2.4), (2.7), (2.26) \tag{A.28}$$

where $h \in \mathbb{R}^{|\mathcal{I}|\cdot|\mathcal{J}|}$ is defined as the vector of QALY scores assigned to patient groups, $\beta \in [0,1]$ is the weight of the efficiency objective, and $\gamma \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|}$ is the Lagrange multiplier vector, and $\mathbf{R}$ is an approximated parameter measuring waiting times at the equilibrium allocation rates under the FCFS policy used by Zenios et al., 2000b.

Notice that the second line in (A.27) contains a matrix integral; this can be transformed to closed-form expressions through matrix calculus. Before we dive into the derivation, recall that the matrix $\boldsymbol{\Psi}$ is based on real data, therefore $\boldsymbol{\Psi}$ is diagonalizable with probability 1; and it is indeed diagonalizable in our estimation using UNOS/OPTN data from 2009 - 2019. Consistent with our discussion on $\boldsymbol{\Psi}$'s non-singularity, in the case that $\boldsymbol{\Psi}$ is not diagonalizable (which occurs with probability 0), we can add a small enough noise matrix/perturbation $\boldsymbol{\epsilon} \in \mathbb{R}^{IJ \times IJ} \to \mathbf{0}$ so that $(\boldsymbol{\Psi} + \boldsymbol{\epsilon})$ is diagonalizable and $\lim_{\boldsymbol{\epsilon} \to 0} \boldsymbol{\Psi} + \boldsymbol{\epsilon} = \boldsymbol{\Psi}$. Summarizing above, we can safely assume that $\boldsymbol{\Psi}$ is diagonalizable, i.e., there exists

a diagonal matrix $\mathbf{D} \in \mathbb{R}^{IJ \times IJ}$ and an invertible matrix $\mathbf{V}$, s.t. $\mathbf{\Psi} = \mathbf{V}^{-1}\mathbf{D}\mathbf{V}$. Note that our proof holds even if Ψ is not diagonalizable; we can use Jordan matrices instead of diagonal $\mathbf{D}$ matrices.

Using the definition of the matrix exponential, we have

$$\begin{aligned} e^{t\mathbf{\Psi}} &= \sum_{k=0}^{\infty} \frac{1}{k!}(t\mathbf{\Psi})^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\left(t\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{V}\right)^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\boldsymbol{V}^{-1}\left(t\boldsymbol{D}\right)^k\boldsymbol{V} \\ &= \boldsymbol{V}^{-1}\left(\sum_{k=0}^{\infty} \frac{1}{k!}(t\boldsymbol{D})^k\right)\boldsymbol{V} \\ &= \boldsymbol{V}^{-1}e^{t\mathbf{D}}\mathbf{V}, \end{aligned}$$

Thus, we know $e^{t\mathbf{\Psi}}$ is also diagonalizable. Therefore, assuming $\beta \in [0, 1)$ (the closed-form expression for (A.27) is trivial when $\beta = 1$), we can equivalently write the second line in (A.27) as $-\frac{1}{1-\beta}\sum_\ell \mathbf{A}^\ell$ where $\mathbf{A}^\ell$ is defined as follows

$$\begin{aligned} \mathbf{A}^\ell &= \int_t^T \left[\exp((\tau - t)\mathbf{\Psi})(\boldsymbol{\lambda} - \mathbf{P}^\ell\mathbf{u}^\ell(t))\right]^\top \mathbf{R}\left(\exp((\tau - t)\mathbf{\Psi})(\boldsymbol{\lambda} - \mathbf{P}^\ell\mathbf{u}^\ell(t))\right) d\tau \\ &= \int_t^T \left[(\mathbf{V}^{-1})\top \exp((\tau - t)\mathbf{D})\mathbf{V}(\boldsymbol{\lambda} - \mathbf{P}^\ell\mathbf{u}^\ell(t))\right]^\top \mathbf{R}\left(\mathbf{V}^{-1}\exp((\tau - t)\mathbf{D})\mathbf{V}(\boldsymbol{\lambda} - \mathbf{P}^\ell\mathbf{u}^\ell(t))\right) d\tau \\ &= (\boldsymbol{\lambda} - \mathbf{P}^\ell\mathbf{u}^\ell(t))^\top \left(\int_t^T \mathbf{V}^\top \exp((\tau - t)\mathbf{D})(\mathbf{V}^{-1})^\top\mathbf{R}\mathbf{V}^{-1}\exp((\tau - t)\mathbf{D})\mathbf{V}d\tau\right)(\boldsymbol{\lambda} - \mathbf{P}^\ell\mathbf{u}^\ell(t)) \end{aligned}$$

Now, we look at the individual elements of $\mathbf{A}^\ell$ and show that we have closed-form expressions for each A_{ij}, $i \in [I], j \in [J]$. For convenience, define $\mathbf{B}^\ell = \int_t^T \mathbf{V}^\top \exp((\tau - t)\mathbf{D})\mathbf{V}^{-1}\mathbf{R}\mathbf{V}^{-1}\exp((\tau - t)\mathbf{D})\mathbf{V}d\tau$. Denote $\mathbf{D}$'s diagonal

elements as the scalars d_{kk}, where $k \in \{1, 2, \ldots, I \cdot J\}$:

$$
\begin{aligned}
B_{ij}^{\ell} &= \sum_{k\in[I],l\in[J]} \int_t^T (\mathbf{V}^\top)_{ik} \exp((\tau - t)d_{kk}) \left((\mathbf{V}^{-1})^\top \mathbf{R}\mathbf{V}^{-1}\right)_{kl} \exp((\tau - t)d_{ll})(\mathbf{V})_{lj} d\tau \\
&= \sum_{k\in[I],l\in[J]} (\mathbf{V}^\top)_{ik} \left((\mathbf{V}^{-1})^\top \mathbf{R}\mathbf{V}^{-1}\right)_{kl} (\mathbf{V})_{lj} \int_t^T \exp((\tau - t)d_{kk}) \exp((\tau - t)d_{ll}) d\tau \\
&= \sum_{k\in[I],l\in[J]} (\mathbf{V}^\top)_{ik} \left((\mathbf{V}^{-1})\top \mathbf{R}\mathbf{V}^{-1}\right)_{kl} (\mathbf{V})_{lj} \int_t^T \exp((\tau - t)[d_{kk} + d_{ll}] d\tau \\
&= \sum_{k\in[I],l\in[J]} (\mathbf{V}^\top)_{ik} \left((\mathbf{V}^{-1})\top \mathbf{R}\mathbf{V}^{-1}\right)_{kl} (\mathbf{V})_{lj} \frac{1}{d_{kk} + d_{ll}} \left[\exp((T - t)(d_{kk} + d_{ll})) - 1\right]
\end{aligned}
$$

From the above derivation, we have explicit closed-form expressions for all B_{ij}^{ℓ}'s, where $i \in [I], j \in [J]$. Thus we can write $\mathbf{B}^{\ell}$ and $\mathbf{A}^{\ell}$ explicitly. Specifically,

$$
\mathbf{A}^{\ell} = -(1 - \beta)(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t))^\top \mathbf{B}^{\ell}(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t))
$$

And line 2 in (A.27) can be written in closed-form as follows:

$$
\sum_{\ell\in\mathcal{L}} \mathbf{A}^{\ell} = -\sum_{\ell\in\mathcal{L}} (1 - \beta)(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t))^\top \mathbf{B}^{\ell}(\boldsymbol{\lambda} - \mathbf{P}^{\ell}\mathbf{u}^{\ell}(t))
$$

A.4.3 Exact Optimal Solution, Reduced Computational Complexity, and Structural Properties

Not only are our solutions exact in the interior space, they also significantly reduce computational complexity. The original fluid control problems have ODEs in the constraints, and such constraints are by nature continuous. Akan et al., 2012 relied on solving the the dual control which requires additional discretization of the dual space for both the interior case and the boundary case.

Besides loss in solution quality as a result of an additional discretization of the continuous dual control problem, one needs to solve a huge discretized LP or QP (Akan et al., 2012; Zenios et al., 2000b) with $O(T_N|\mathcal{I}||\mathcal{J}||\mathcal{K}|)$ variables and $O(T_N|\mathcal{I}||\mathcal{J}||\mathcal{K}|)$ constraints for the decision rule at $t \in \mathcal{T}$. While in our decomposed LP, we only need to solve a small LP with $O(|\mathcal{I}||\mathcal{J}||\mathcal{K}|)$ variables and constraints. Note that in practice, $O(|\mathcal{I}||\mathcal{J}||\mathcal{K}|)$ is less than 10^4 even under the most granular classification, but T_N can easily go up to $O(10^6)$ scale. Thus, our decomposition results reveal that the fluid model-optimal policy's decision rules are solutions to standard LPs—this finding significantly reduces the complexities of solving the fluid control problem with ODEs in the constraints. Moreover, we can easily derive the optimal decision rule at any $t \in \mathcal{T}$ in the interior of the state space, without the need to discretize the continuous space.

Our decomposed optimal decision rules imply and corroborate the structural properties found in Akan et al., 2012 (i.e. that the optimal policy contains decision rules maximizing some dynamic indexes), and offers a simple and fast approach to find the exact and explicit dynamic indexes without involving the dual control problem. All monotonicity results found in the previous literature become even more clear and straightforward, with our closed-form expressions for the dynamic indexes (see Section A.7 in Appendix for detail). We also provide new insights, for example, the convexity and piece-wise linearity of the dynamic index values as functions of resource and fairness constraints. Our exact solutions via fluid limit decomposition illuminate the full potential and inherent properties of the fluid approximation and fluid model-based optimization in organ transplantation.

A.5 Singular Patient Health Transition Matrix

When patient health transition matrix $\mathbf{\Psi}$ is estimated from real data numerically, $\mathbf{\Psi}$ is non-singular with very high probability and with probability 1 with infinite precision. Even if we get a singular $\mathbf{\Psi}$, we can add an arbitrarily small perturbation/error matrix to break the singularity, as described in Section 2.4.

For theoretical completeness, we show that even with a singular $\mathbf{\Psi}$, we can still remove the matrix integration and write the LP objectives (which are functions of $\mathbf{x}(t)$, $\mathbf{u}^{\ell}(t)$, and $\mathbf{s}^{\ell}(t)$) in closed form. The only task is to remove the integral of the matrix exponential in the expression for $\mathbf{x}(t)$. When $\mathbf{\Psi}$ is singular, using the Jordan form, we can rewrite it as

$$\mathbf{\Psi} = \mathbf{V}^{-1} \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \mathbf{V}$$

where $\mathbf{B}$ is non-singular, and $\mathbf{A}$ is strictly upper triangular. $\mathbf{V}$ is an invertible matrix. Using the definition of the matrix exponential, we have

$$e^{t\mathbf{\Psi}} = \mathbf{V}^{-1} \begin{pmatrix} e^{t\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & e^{t\mathbf{B}} \end{pmatrix} \mathbf{V}$$

Applying Proposition 3 to the non-singular $\mathbf{B}$, the integral of the matrix

exponential above can be written as

$$\int_a^b e^{t\mathbf{\Psi}} dt = \mathbf{V}^{-1} \begin{pmatrix} \int_a^b e^{t\mathbf{A}} dt & \mathbf{0} \\ \mathbf{0} & \left(e^{b\mathbf{B}} - e^{a\mathbf{B}}\right) \mathbf{B}^{-1} \end{pmatrix} \mathbf{V}$$

Denote A's dimension as n (in our problem, $\mathbf{\Psi}$ is a square matrix and its $n = I \cdot J$) Since $\mathbf{A}$ is strictly upper triangular, thus $\mathbf{A}^{n+k} = \mathbf{0}$, $\forall k \in \mathbb{N} \cup \{0\}$. Thus, $\int_0^T e^{t\mathbf{A}} dt$ can be written in closed-form:

$$\begin{aligned} \int_0^T e^{t\mathbf{A}} dt &= T\left(\mathbf{I} + \frac{\mathbf{A}T}{2!} + \frac{(\mathbf{A}T)^2}{3!} + \cdots + \frac{(\mathbf{A}T)^{n-1}}{n!} + \cdots\right) \\ &= T\left(\mathbf{I} + \frac{\mathbf{A}T}{2!} + \frac{(\mathbf{A}T)^2}{3!} + \cdots + \frac{(\mathbf{A}T)^{n-1}}{n!}\right) \end{aligned}$$

Therefore,

$$\int_a^b e^{t\mathbf{A}} dt = \mathbf{I}(b-a) + b\left(\frac{\mathbf{A}b}{2!} + \frac{(\mathbf{A}b)^2}{3!} + \cdots + \frac{(\mathbf{A}b)^{n-1}}{n!}\right) - a\left(\frac{\mathbf{A}a}{2!} + \frac{(\mathbf{A}a)^2}{3!} + \cdots + \frac{(\mathbf{A}a)^{n-1}}{n!}\right)$$

Summarizing above, even when $\mathbf{\Psi}$ is singular, we can still write the objectives of our decomposed LP in closed form.

A.6 Proofs for Analytical Results in Section 2.4.5

A.6.1 Proof for Proposition 5

Proof. Recall that

$\boldsymbol{\mu}^\ell(t) \in \mathbb{R}_+^{I \cdot J}, \bar{\boldsymbol{\mu}}^\ell(t) \in [\mathbf{0}, \boldsymbol{\mu}^\ell(t)], \mathbf{\Theta} \in [0,1]^{I \cdot J}, \boldsymbol{\lambda}(t) \in \mathbb{R}_+^{I \cdot J}, \forall \ell \in \mathcal{L}, t \in \mathcal{T}$. The feasible set of $f(\cdot)$ is convex. We first prove that

$g_t^{\text{Multi}}(\boldsymbol{\mu}^\ell, \bar{\boldsymbol{\mu}}^\ell, \boldsymbol{\Theta}\boldsymbol{\lambda}, \boldsymbol{\lambda}, \ell \in \mathcal{L}) := \max_{(\mathbf{U},\mathbf{S})} \mathscr{U}_t^{\text{Multi}}$ is piece-wise linear concave in the RHS (e.g. $\boldsymbol{\mu}^\ell, \bar{\boldsymbol{\mu}}^\ell, \boldsymbol{\Theta}\boldsymbol{\lambda}, \boldsymbol{\lambda}, \ell \in \mathcal{L}$). The proof follows the global sensitivity analysis from Bertsimas and Tsitsiklis, 1997 (see their Section 5.2, Equation 5.2). According to our fluid limit decomposition in the fully overloaded setting and the exchangeability in integration dimensions (2.24), $\text{OBJ}^{\text{Multi}}$ is a definite integral of $g_t^{\text{Multi}}(\boldsymbol{\mu}^\ell, \bar{\boldsymbol{\mu}}^\ell, \boldsymbol{\Theta}\boldsymbol{\lambda}, \boldsymbol{\lambda}, \ell \in \mathcal{L})$ from 0 to T. Concavity is preserved by integrals (Boyd & Vandenberghe, 2004). □

A.6.2 Proof for Corollary 2.5.1

Proof. According to Proposition 5, $f(\boldsymbol{\mu}^1, ..., \boldsymbol{\mu}^{|\mathcal{L}|}, \bar{\boldsymbol{\mu}}^1, \ldots, \bar{\boldsymbol{\mu}}^{|\mathcal{L}|}, \boldsymbol{\Theta}\boldsymbol{\lambda}, \boldsymbol{\lambda})$ is concave, thus $\frac{\partial f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\boldsymbol{\Theta}\boldsymbol{\lambda},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}}$ (the marginal benefit of one additional donor liver) is monotonically non-increasing in $\boldsymbol{\mu}$, and $\frac{\partial f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\boldsymbol{\Theta}\boldsymbol{\lambda},\boldsymbol{\lambda})}{\partial \bar{\boldsymbol{\mu}}}$ (the marginal benefit of one additional split-table donor liver) is monotonically non-increasing in $\bar{\boldsymbol{\mu}}$. □

A.6.3 Proof for Corollary 2.5.2

Proof. First, because the larger $\boldsymbol{\Theta} \geq 0$ is, the more restrictive (2.12) is, the smaller the feasible set the LP (2.34) $\sim$ (2.35) has. As a result, the objective function value (2.34) potentially decreases, thus $f(\boldsymbol{\mu}^1, ..., \boldsymbol{\mu}^{|\mathcal{L}|}, \bar{\boldsymbol{\mu}}^1, \ldots, \bar{\boldsymbol{\mu}}^{|\mathcal{L}|}, \boldsymbol{\Theta}\boldsymbol{\lambda}, \boldsymbol{\lambda})$ potentially goes down when $\boldsymbol{\Theta}$ increases. Therefore, $\text{PoF} = 1 - \frac{f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\boldsymbol{\Theta}\boldsymbol{\lambda},\boldsymbol{\lambda})}{f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\mathbf{0},\boldsymbol{\lambda})}$ is non-decreasing.

According to Proposition 5, $f(\boldsymbol{\mu}^\ell, \bar{\boldsymbol{\mu}}^\ell, \boldsymbol{\Theta}, \boldsymbol{\lambda})$ is concave. Because $f(\boldsymbol{\mu}^\ell, \bar{\boldsymbol{\mu}}^\ell, \mathbf{0}, \boldsymbol{\lambda}) > 0$ is a fixed number, $\frac{f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\boldsymbol{\Theta}\boldsymbol{\lambda},\boldsymbol{\lambda})}{f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\mathbf{0},\boldsymbol{\lambda})}$ is concave, and $\text{PoF} = 1 - \frac{f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\boldsymbol{\Theta}\boldsymbol{\lambda},\boldsymbol{\lambda})}{f(\boldsymbol{\mu}^1,...,\boldsymbol{\mu}^{|\mathcal{L}|},\bar{\boldsymbol{\mu}}^1,...,\bar{\boldsymbol{\mu}}^{|\mathcal{L}|},\mathbf{0},\boldsymbol{\lambda})}$ is convex. □

A.6.4 Proof for Proposition 4

Proof. First, we present two lemmas that will be useful in the proof of monotonicity of the dynamic indices:

Lemma A.6.1. For a square matrix A and any $t \in \mathbb{R}_{++}$, $\frac{de^{tA}}{dt} = e^{tA}A$.

Lemma A.6.2. If matrices A and B commute, i.e., $AB = BA$, then $e^{A+B} = e^A e^B$.

Let $\mathbf{I}_{IJ,IJ}/\mathbf{I}$ denote the identity matrix of dimension $IJ \times IJ$. Because $\mathbf{I}(\mathbf{\Psi} + \mathbf{I}) = (\mathbf{\Psi} + \mathbf{I})\mathbf{I}$ and $\mathbf{I}^{-1} = \mathbf{I}$; thus $\mathbf{I}^{-1}(\mathbf{\Psi} + \mathbf{I}) = (\mathbf{\Psi} + \mathbf{I})\mathbf{I}^{-1}$, and $(T-t)\mathbf{I}^{-1}(T-t)(\mathbf{\Psi} + \mathbf{I}) = (T-t)(\mathbf{\Psi} + \mathbf{I})(T-t)\mathbf{I}^{-1}$. According to Lemma A.6.2, we have $\exp((T-t)\mathbf{\Psi}) = \exp((T-t)(\mathbf{\Psi} + \mathbf{I} - \mathbf{I})) = \exp((T-t)(\mathbf{\Psi} + \mathbf{I}))\exp(-(T-t)\mathbf{I})$.

Given the two lemmas, for $D(t)$:

$$
\begin{aligned}
\frac{dD(t)}{dt} &= \frac{d\left(\mathbf{d}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right)\mathbf{\Psi}^{-1}\sum_\ell \mathbf{P}^\ell\right)}{dt} \\
&= \frac{d\left(\mathbf{d}^\top \exp((T-t)\mathbf{\Psi})\mathbf{\Psi}^{-1}\sum_\ell \mathbf{P}^\ell\right)}{dt} \\
&= \mathbf{d}^\top \left(\frac{d\left(\exp((T-t)\mathbf{\Psi})\right)}{dt}\right)\mathbf{\Psi}^{-1}\sum_\ell \mathbf{P}^\ell \\
&= \mathbf{d}^\top \left(-\exp((T-t)\mathbf{\Psi})\mathbf{\Psi}\right)\mathbf{\Psi}^{-1}\sum_\ell \mathbf{P}^\ell \qquad \text{(apply Lemma A.6.1)} \\
&= -\mathbf{d}^\top \exp((T-t)\mathbf{\Psi})\sum_\ell \mathbf{P}^\ell \\
&= -\mathbf{d}^\top \exp((T-t)(\mathbf{\Psi} + \mathbf{I} - \mathbf{I}))\sum_\ell \mathbf{P}^\ell \\
&= -\mathbf{d}^\top \exp((T-t)(\mathbf{\Psi} + \mathbf{I}) - (T-t)\mathbf{I})\sum_\ell \mathbf{P}^\ell
\end{aligned}
$$

$$= -\mathbf{d}^\top \exp((T-t)(\mathbf{\Psi}+\mathbf{I}))\exp(-(T-t)\mathbf{I})\sum_\ell \mathbf{P}^\ell \qquad \text{(Apply Lemma A.6.2)}$$

Let us look at $\exp((T-t)(\mathbf{\Psi}+\mathbf{I}))$ in the last line above. Recall that $\mathbf{\Psi}$'s diagonal elements are in $[-1,0)$, while its nondiagonal elements are in $[0,1)$. Thus, all of $(\mathbf{\Psi}+\mathbf{I})$'s elements are in $[0,1)$. As a result, $(\mathbf{\Psi}+\mathbf{I})^k \geq \mathbf{0}$ and $[(T-t)(\mathbf{\Psi}+\mathbf{I})]^k \geq \mathbf{0}$ for any $k \in \mathbb{N}^*$ and $t \in [0,T]$. By definition,

$$\exp((T-t)(\mathbf{\Psi}+\mathbf{I})) = \sum_{k=0}^{\infty} [(T-t)(\mathbf{\Psi}+\mathbf{I})]^k / k! \geq \mathbf{0}$$

Since $-(T-t)(\mathbf{I})$ is a diagonal matrix, thus $\exp(-(T-t)(\mathbf{I}))$ is also a diagonal matrix and its diagonal elements are $e^{t-T} > 0$. Since $\mathbf{d}$ and $\mathbf{P}^\ell$'s elements are nonnegative, we have

$$\frac{dD(t)}{dt} = -\mathbf{d}^\top \exp((T-t)(\mathbf{\Psi}+\mathbf{I}))\exp(-(T-t)(\mathbf{I}))\sum_\ell \mathbf{P}^\ell \leq \mathbf{0}$$

This tells us that $D(t)$ is nonincreasing in t. Replace $\mathbf{P}^\ell$ with $\bar{\mathbf{P}}^\ell$, the same proof goes through for $\bar{D}(t)$. Similarly, replace $\mathbf{d}$ above with $(-\mathbf{q})$, we have $Q(t)$ is nondecreasing. Replacing $\mathbf{d}$ above with $(-\mathbf{q})$ and $\mathbf{P}^\ell$ with $\bar{\mathbf{P}}^\ell$, we get $\bar{Q}(t)$'s non-decreasing proof. □

In proving monotonicity we showed above that $\exp(a\mathbf{\Psi}) > 0$ for any $a \geq 0$. This also implies that the NPDWT dynamic index is nonnegative:

$$\mathscr{U}_t^{\text{NPDWT}} := \mathbf{d}^\top \left\{ \int_t^T \exp[(\tau - t)\mathbf{\Psi}] d\tau \right\} \left(\sum_\ell \mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t) \right) \geq \mathbf{0}$$

The first term in the QALY dynamic index is nonpositive:

$$\max_{\mathbf{U}(t),\mathbf{S}(t)} \mathscr{U}_t^{\text{QALY}} := -\sum_{\ell} \mathbf{q}^\top \left\{ \int_t^T \exp[(\tau - t)\mathbf{\Psi}] d\tau \right\} \left(\mathbf{P}^\ell \mathbf{u}^\ell(t) + \bar{\mathbf{P}}^\ell \mathbf{s}^\ell(t) \right) \leq \mathbf{0}$$

Proposition 4 essentially tells us that as t increases, the absolute values of dynamic indices's coefficient vectors shrink. This suggests earlier allocation decisions have a larger impacts in absolute value on the cumulative system objective value.

A.6.5 Proof for Proposition 6

Proof. Based on the explicit objective function of LP (2.34) $\sim$ (2.35), when $\bar{D}^\ell_{ij,i'j'}(t) e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} D^\ell_{i''j''}(t) e^{IJ}_{i''j''}$ and $\bar{Q}^\ell_{ij,i'j'}(t) e^{IJ\times IJ}_{ij,i'j'} + (\bar{H}^\ell_{ij} + \bar{H}^\ell_{i'j'}) \bar{P}^\ell_{ij,i'j'} e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} Q^\ell_{i''j''}(t) e^{IJ}_{i''j''} + H^\ell_{i''j''} P^\ell_{i''j''} e^{IJ}_{i''j''}$ hold, the coefficient of splitting ℓ and transplanting it to $(ij, i'j')$ dominates, thus the objective value of a decision rule that allocates as many type-ℓ livers as possible to a $(ij, i'j')$-pair for SLT is higher than that of a decision rule allocates more-than-necessary type-ℓ livers for WLT. Constraint (2.5) ensures that the allocation does not exceed the splittable-liver capacity; this constraint is ordinary and thus won't affect the optimal splitting decisions. (2.6) prevents the allocation from infringing the fairness guarantee; under certain choices of $\mathbf{\Theta}$, this could mean forcing the decision rules to deviate from the optimal splitting decisions without fairness constraints. (2.26) assures that the fluid limits to always stay non-negative (i.e. never assign livers to empty fluid queues). This constraint is not active in the interior case; for completeness, we include it here and it serves as a constraint for the boundary case heuristics. □

A.6.6 Proof for Corollary 2.5.3

Proof. When for all $\ell \in \mathcal{L}$, $\exists\, i, i' \in \mathcal{I}, j, j' \in \mathcal{J}$, s.t. (a) $\bar{D}^{\ell}_{ij,i'j'}(t)e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} D^{\ell}_{i''j''}(t)e^{IJ}_{i''j''}$, and b) $\bar{Q}^{\ell}_{ij,i'j'}(t)e^{IJ\times IJ}_{ij,i'j'} + (\bar{H}^{\ell}_{ij} + \bar{H}^{\ell}_{i'j'})\bar{P}^{\ell}_{ij,i'j'}e^{IJ\times IJ}_{ij,i'j'} \geq \max_{i'',j''} Q^{\ell}_{i''j''}(t)e^{IJ}_{i''j''} + H^{\ell}_{i''j''}P^{\ell}_{i''j''}e^{IJ}_{i''j''}$ hold, apply Proposition 6: The coefficient of splitting ℓ and transplanting it to their corresponding $(ij, i'j')$ dominates. Thus, for every ℓ, the objective value of a decision rule that allocates as many type-ℓ livers as possible to their corresponding $(ij, i'j')$-pair for SLT is higher than that of a decision rule allocates more-than-necessary type-ℓ livers for WLT. Of course, the decision rules are subject to the other constraints and may have to deviate from always splitting at some t's. Please refer to the proof for Proposition 6 for discussions on the influences of constraints. □

A.7 Analytical Results Analogous to Propositions 1 and 2 from Akan et al. 2012

With our closed-form objective functions and constraints, we can solve the decomposed LPs with standard solvers and perform sensitivity analysis on the optimal decision rules using the explicit LPs. For example, if we want to study the impact of increasing d^{ℓ}_{ij} or H^{ℓ}_{ij} on our optimal decision: when we increase/decrease the parameters just a little bit, the base of the solution may not change; however, our optimal solution and the base may change when we further increase/decrease the parameters to some point. Below we present two corollaries analogous to Propositions 1 and 2 from-Akan et al., 2012 but in our fluid formulation with SLT

and fairness. Note that we only demonstrate a subset of sufficient conditions, more general results are possible.

The UNOS policies usually give precedence to the sickest patients. Corollary A.7.1 characterizes sufficient scenarios where such a strategy is optimal in minimizing patient deaths on the waitlists. Corollary A.7.1 is not a direct translation or extention of Proposition 1 from Akan et al., 2012 in our context, due to the difference in formulation of the transition matrix $\mathbf{\Psi}$, but it provides comparable and broader insights.

Corollary A.7.1. Suppose that $P^{\ell}_{ij} = P^{\ell}_{i'j} = P^{\ell}_{j}$ for all i, i', j, ℓ for WLT, $\bar{P}^{\ell}_{ij,i'j'} = \bar{P}^{\ell}_{i''j,i'j'} = \bar{P}^{\ell}_{j,i'j'}$, and $\bar{P}^{\ell}_{i'j',ij} = \bar{P}^{\ell}_{i'j',i''j} = \bar{P}^{\ell}_{i'j',j}$ for SLT, $\forall i, i', i'', j, j', \ell$. In an optimal solution to (2.1) $\sim$ (2.7), within the same static group, patients' relative priorities are set in the order of their death rates $\mathbf{d}$, i.e., giving the priority to the sickest patients (who are mostly likely to die in the next 90 days), provided that $d_{ij} > d_{i(j-1)}, \forall i, j = \{2, \ldots, J\}$ and $y := \mathbf{d}^{\top} \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1} \in \mathbb{R}^{1 \times IJ}$ satisfies $y_{ij} > y_{i(j-1)}, \forall i, j = \{2, \ldots, J\}$, $\forall t \in (0, T-t]$.

The proof of Corollary A.7.1 uses the decomposed dynamic indexes. The high-level idea is that y summarizes the total reduced deaths (including the immediate removals and the reduced future deaths as a result of shortened waitlists) from t to T per unit of liver allocated to each of the patient classes.

Proof. Because $d_{ij} > d_{i(j-1)}, \forall i, j = \{2, \ldots, J\}$ and $y_{ij} > y_{i(j-1)}, \forall i, j = \{2, \ldots, J\}$, $\forall t \in (0, T-t]$ and the assumptions on $\mathbf{P}$ and $\bar{\mathbf{P}}$,
$C := \mathbf{d}^{\top} \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1}\mathbf{P} \in \mathbb{R}^{1 \times IJ}$ and

$\bar{C} := \mathbf{d}^\top \left(\exp((T-t)\mathbf{\Psi}) - \mathbf{I}\right) \mathbf{\Psi}^{-1}\bar{\mathbf{P}} \in \mathbb{R}^{1\times IJ}$ satisfy: $C_{ij} > C_{i(j-1)}$, $\bar{C}_{ij,i'j'} > C_{i(j-1),i'j'}$, and $\bar{C}_{i'j',ij} > C_{i'j',i(j-1)}$, $\forall i, i', j, j', t$ and $j \in \{2, \dots, J\}$. Therefore, the marginal benefit or the weight of class ij is larger than that of patient class $i(j-1)$, for all $i, j \in \{2, \dots, J\}$; the optimal solution will prioritize those patient classes with larger marginal benefit, equivalently, the sickest under the premises of Corollary A.7.1. □

The UNOS policies sometimes prioritize certain static patient groups when other conditions are the same, e.g. children over adults. Corollary A.7.2 characterize when such policies are optimal and generalizes Proposition 2 from Akan et al., 2012.

Corollary A.7.2. Suppose that there exists a permutation $r^\ell(\cdot)$ of $\mathcal{I}$ such that $q^\ell_{r^\ell(1)j} \geq q^\ell_{r^\ell(2)j} \geq \cdots \geq q_{r(I)^\ell j}$ for all $j \in \mathcal{J}, \ell \in \mathcal{L}$, and for each $\ell \in \mathcal{L}$, and a subset of the following conditions holds

$$H^\ell_{r^\ell(1)j} \geq H^\ell_{r^\ell(2)j} \geq \cdots \geq H^\ell_{r^\ell(I)j}, \quad \forall j \in \mathcal{J} \tag{A.29}$$

$$\bar{H}^\ell_{r^\ell(1)j,i'j'} \geq \bar{H}^\ell_{r^\ell(2)j,i'j'} \geq \cdots \geq \bar{H}^\ell_{r^\ell(I)j,i'j'}, \quad \forall j, j' \in \mathcal{J}, i' \in \mathcal{I} \tag{A.30}$$

$$\bar{H}^\ell_{i'j',r^\ell(1)j} \geq \bar{H}^\ell_{i'j',r^\ell(2)j} \geq \cdots \geq \bar{H}^\ell_{i'j',r^\ell(I)j}, \quad \forall j, j' \in \mathcal{J}, i' \in \mathcal{I} \tag{A.31}$$

$$P^\ell_{r^\ell(1)j} \geq P^\ell_{r^\ell(2)j} \geq \cdots \geq P^\ell_{r^\ell(I)j}, \quad \forall j \in \mathcal{J} \tag{A.32}$$

$$\bar{P}^\ell_{r^\ell(1)j,i'j'} \geq \bar{P}^\ell_{r^\ell(2)j,i'j'} \geq \cdots \geq \bar{P}^\ell_{r^\ell(I)j,i'j'}, \quad \forall j, j' \in \mathcal{J}, i' \in \mathcal{I} \tag{A.33}$$

$$\bar{P}^\ell_{i'j',r^\ell(1)j} \geq \bar{P}^\ell_{i'j',r^\ell(2)j} \geq \cdots \geq \bar{P}^\ell_{i'j',r^\ell(I)j}, \quad \forall j, j' \in \mathcal{J}, i' \in \mathcal{I} \tag{A.34}$$

a. If (A.29) and (A.32) hold, $H^\ell_{r^\ell(1)j} > \max\{\bar{H}^\ell_{r^\ell(1)j,i'j'}, \bar{H}^\ell_{i'j',r^\ell(1)j}\}$, and $P^\ell_{r^\ell(1)j} > \max\{\bar{P}^\ell_{r^\ell(1)j,i'j'}, \bar{P}^\ell_{i'j',r^\ell(1)j}\}$, an optimal solution to (2.8) $\sim$ (2.9) only assigns livers of type ℓ to static patient type $r^\ell(1) = \operatorname{argmax}_i H^\ell_{ij}$ if $x_{ij} > 0$ for

all i, j and $t \in \mathcal{T}$.

b. If (A.30) and (A.33) hold, $\bar{H}^{\ell}_{r^{\ell}(1)j,i'j'} > \max\{H^{\ell}_{r^{\ell}(1)j}, \bar{H}^{\ell}_{i'j',r^{\ell}(1)j}\}$, and $\bar{P}^{\ell}_{r^{\ell}(1)j,i'j'} > \max\{P^{\ell}_{r^{\ell}(1)j}, \bar{P}^{\ell}_{i'j',r^{\ell}(1)j}\}$, an optimal solution to (2.8) $\sim$ (2.9) only assigns partial livers of type ℓ to static patient type $r^{\ell}(1) = \text{argmax}_i \bar{H}^{\ell}_{ij}$ as primary recipients if $x_{ij} > 0$ $\forall i, j$ and $t \in \mathcal{T}$.

c. If (A.31) and (A.34) hold, $\bar{H}^{\ell}_{i'j',r^{\ell}(1)j} > \max\{\bar{H}^{\ell}_{r^{\ell}(1)j,i'j'}, H^{\ell}_{r^{\ell}(1)j}\}$, and $\bar{P}^{\ell}_{i'j',r^{\ell}(1)j} > \max\{\bar{P}^{\ell}_{r^{\ell}(1)j,i'j'}, P^{\ell}_{r^{\ell}(1)j}\}$, an optimal solution to (2.8) $\sim$ (2.9) only assigns partial livers of type ℓ to static patient type $r^{\ell}(1) = \text{argmax}_i \bar{H}^{\ell}_{ij}$ as secondary recipients if $x_{ij} > 0$ $\forall i, j$ and $t \in \mathcal{T}$.

The proof is straight-forward, because between static classes there is no transition. Therefore, we can ignore $\mathbf{\Psi}$ in (2.27) $\sim$ (2.28) and based on the explicit LP formulation, draw the monotonicity conclusions directly.

A.8 Current SLT Practice

For completeness, we describe the practice of SLT in the US and more specifically, at the world-renowned transplant center where two of the authors work.

A.8.1 Two Splitting Methods for SLT

There are two splitting methods and a liver can only be split once, according to the OPTN white paper (OPTN & UNOS, 2016):

- An adult-child split. In this splitting method, a small child or very small-statured adult receives the smaller left lobe, and an adult receives the

extended right lobe.

- An adult-adult (or adult-big child) split where an adult receives the right lobe, and an adult (or big child) receives the left lobe.

Current SLT practice indicates that the adult-child split is consistently favorable. Nevertheless, recent reports indicated good results could be achieved in relatively healthier recipients and advanced techniques (OPTN & UNOS, 2016).

A.8.2 SLT Expertise

In the US, after graduating from medical schools and having chosen their specialization areas, surgeons complete their residency programs to obtain an unrestricted license to practice medicine and a board certificate for their chosen surgical specialty, in our case, the liver transplant. It is during residency that surgeons may learn SLTs at selected TCs, such as the one at University of California, San Francisco.

A.8.3 The Liver Allocation Procedure and SLT Use as Exceptional Cases

In practice, successful SLTs involve a complicated process, including registration, procurement, allocation, logistics, surgical operations, and post-surgery recovery. To start, eligible ESLD patients choose transplant centers and register for the national liver transplant waitlists. When a deceased-donor liver becomes available and is being evaluated to determine whether it is medically splittable (based on donor age, body mass index, size, etc.), UNOS generates a ranked list (known as the *match-run*), based on computerized algorithms. The organ is offered to the match-run candidates sequentially, until a candidate/candidate pair accepts it. The

longer it takes between the removal of blood supply from the deceased-donor organ and the transplantation into the recipient(s) (the *cold ischemia time*), the more the organ's quality deteriorates. An organ is discarded if the cold ischemia time is determined to be too long (exceeding 12 - 18 hours).

Once a liver is accepted, the organ is harvested (and split if to be used in SLTs) by a trained team at the donor hospital. The matching of transplant surgeons and candidates is finalized after a candidate has accepted an offer and right before the surgery. After being harvested, the procured whole liver (two split liver grafts) is transported to the WLT patient TC (SLT patient TCs), where the WLT surgery (SLT surgeries) is performed by the patient's transplant surgeon, and patient recovery occurs.

Currently, most SLTs are performed in few major transplant centers; thus, the primary recipients (usually children) and secondary recipients are usually within the same TC (Ge et al., 2020). However, because of UNOS's new acuity circles policy that took effect in 2019 (UNOS system notice: Liver and intestinal organ distribution based on acuity circles implemented February 4, 2020, https://unos.org/news/system-implementation-notice-liver-and-intestinal-organ-distribution-based-on-acuity-circles-implemented-feb-4/), patients from different TCs within the acuity circles may receive halves of the same donor liver more frequently. Researchers are also exploring continuous distribution that do not rely on geographical boundaries (Bertsimas et al., 2020; Kasiske et al., 2020).

A.9 Numerical Experiments

A.9.1 Numerical Setup

In the first set of experiments, we compare the objective values of the optimal solution (the “fluid optimal") with objective values of (2.15) under alternative policies. This experiment focuses on the structural properties of the “fluid optimal" policy and captures only the first-order dynamics of the liver wait lists.

In our second experiment, we set up a discrete simulation model for a 1-year time horizon. Liver and patient arrivals in each discrete time step follow Poisson Distribution; and patient deaths, transitions, and removals follow a Binomial distribution. All parameters are calibrated to available data. Each simulation run is a sample path, and we generate Figure 2.6 using five runs for each box. The simulation model mimics the liver allocation system and incorporates higher-order dynamics.

For specific parameter settings and the codes, please see the source code. Code can be accessed using this link: https://drive.google.com/drive/folders/1WU90jtm9A_fTj1oVzHM0mrghPS_pyGph?usp=sharing.

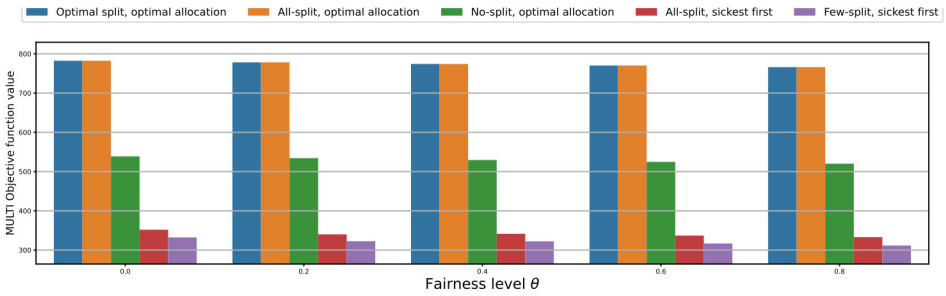


Figure A.1: Comparisons of five policies under the maximizing multi objective where $\kappa = 0.01$. We compare the same five policies discussed in Section 2.6: The "all-split, optimal allocation" policy seems to perform as well as "optimal-split, optimal allocation" and dominates other policies. "All-split, sickest first" consistently outperforms "few-split, sickest first." The benefits of wider use of SLT appear to be more significant in "optimal allocation" policies.

A.9.2 Additional Numerical Experiment Results

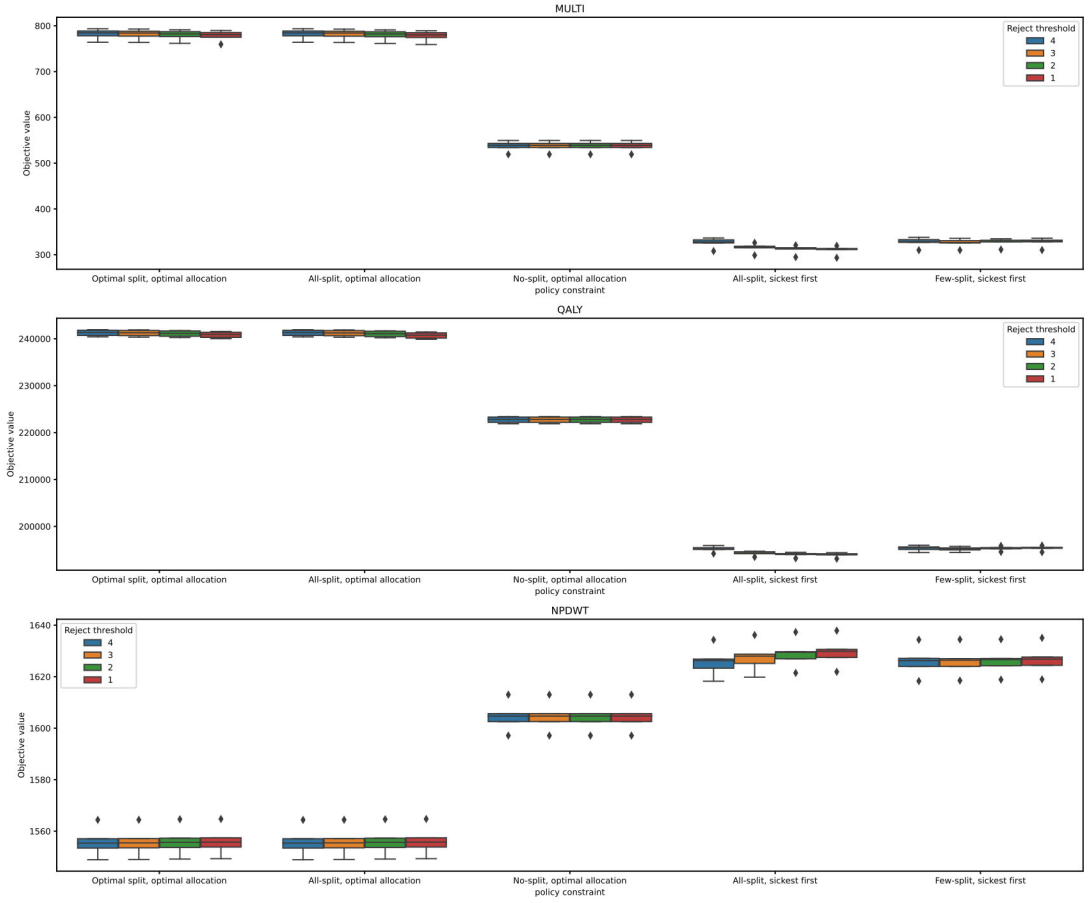


Figure A.2: Simulation results based on OPTN data: We experiment with smaller reject thresholds in this experiment. Smaller reject thresholds indicate worse objective values. The "all-split, sickest first" policy seems the most sensitive to strategic behaviors.

A.9.3 Additional Discussions on Our Numerical Experiments

The improvement of our "optimal split, optimal allocation" policy over other policies that do not utilize SLT likely gives a conservative estimate of the potential of SLTs to achieve multiple liver transplantation goals, because in reality $> 10\%$ of donated livers are splittable (OPTN & UNOS, 2016), yet we assume $\bar{\boldsymbol{\mu}} = 0.1\boldsymbol{\mu}$ as there is no consensus on the exact percentage of splittable livers.

As already mentioned, "all-split, optimal allocation" performs nearly as well at maximizing QALY and MULTI, and minimizing NPDWT; this similarity in performance is driven by the fact that the "optimal split, optimal allocation" strategy splits more than 99% of medically-splittable livers in the simulation.

A.10 Future Directions

This paper is the first in its kind that studies fluid models in SLT. We hope to provide insights, recommend policy modifications, generate discussion, and inspire more detailed analyses regarding implementation in the operations research and transplantation communities. For instance, we estimate the SLT outcomes using the data available, but one could do sensitivity analysis regarding outcomes, factoring in selection biases, heterogeneous medical expertise, medical learning, and geographical distributions. For example, Y. S. Tang et al., 2021 study the donated liver allocation problem in a setting where surgeons with different potential abilities may learn SLT, becoming skilled over time. They formulate a multi-armed bandit that could incorporate first-order queueing dynamics using our fluid limit decomposition.

Appendix B

Appendix for Chapter 3

B.1 Proofs for Theoretical Results in Section 3.4

B.1.1 Alternative Statement of Theorem 3.4.1 and Proof

While Theorem 3.4.1 is a canonical statement of the regret upper bounds, Theorem B.1.1 is a stronger statement mathematically.

Theorem B.1.1. Let $\hat{\alpha}_{a,n}$ be the estimator for the aptitude of arm a, i.e. α_a, after it has been chosen n times. Suppose $\hat{\alpha}_{a,n}$ has a per-coordinate difference bound with parameter $C^w_{a,n}$ and bias $b_{a,n}$. Define $\delta_{a,\tau,n} := \sqrt{\frac{2\log\tau}{nC^w_{a,n}}}$ For any sub-optimal arm a, if there exists a $u_{a,t} \in [1,t]$ such that $\Delta_a \geq 2\delta_{a,\tau,n}$ and $|b_{a,n}| \leq \frac{1}{10}\delta_{a,\tau,n}$ hold for any $t \geq \tau \geq n \geq u_{a,t}$, then arm a is pulled on average at most

$$\mathbb{E}[T_a(t)] \leq u_{a,t} + 2\zeta(1.24)$$

times, where ζ is the Riemann zeta function, i.e. $\zeta(s) = \sum_{n=1}^{+\infty} n^{-s}$, and $\zeta(1.24)$ is approximately 4.76. If such a $u_{a,t}$ exists for any sub-optimal arm, then the expected cumulative regret is bounded by

$$\mathbb{E}[R_t] \leq \sum_{a \neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\,(u_{a,t} + 2\zeta(1.24))$$

Remark: Before we prove this theorem, we show its application in some simple cases.

First, when $\hat{\alpha}_{a,n}$ is the empirical mean of n independent Bernoulli random variables or any random variables on $[0,1]$, we have $C^w_{a,n} = 1$ and $b_{a,n} = 0$. We may choose $u_{a,t} = \frac{8\log t}{\Delta_a^2}$, indicating that this theorem recovers the bound of the vanilla UCB yet with the larger constant $2\zeta(1.24) \approx 9.52$ compared to $\frac{\pi^2}{3} \approx 3.29$. The larger constant results from the loose inequality dealing with the bias, i.e. as we decrease the $\frac{1}{10}$ in $|b_{a,n}| \leq \frac{1}{10}\delta_{a,\tau,n}$ towards 0, the constant will approach $\frac{\pi^2}{3}$.

Second, if we scale the value of $\hat{\alpha}_{a,n}$ and the sub-optimal gap by ℓ, then $C^w_{a,n}$ becomes $\frac{1}{\ell^2}$, and thus $u_{a,n}$ is unchanged. This indicates the bound is scale-free.

Third, when $\hat{\alpha}_{a,n}$s have smaller and/or different $C^w_{a,n}$s and still zero bias, and when $C^w_a := \inf_n\ C^w_{a,n} > 0$, i.e. when $C^w_{a,n}$ is uniformly bounded by a constant from below, we know the minimal $u_{a,t}$ is at most $\frac{2\log t}{C^w_a \Delta_a^2}$ (because we proved $\frac{2\log t}{C^w_a \Delta_a^2}$ is a valid choice for $u_{a,t}$ in Theorem 3.4.1) and therefore $\mathbb{E}[T_a(t)]$ is still in $O(\log t)$ scale, although the coefficient is larger.

Fourth, when $C^w_a := \inf_n\ C^w_{a,n} > 0$ and $C^b_a := \sup_n\ \sqrt{n}|b_{a,n}| < +\infty$, i.e. $|b_{a,n}| = O\left(\frac{1}{\sqrt{n}}\right)$, we may let $u_{a,t} = \max\left\{\exp\left(C^w_a (C^b_a)^2/200\right), \frac{2\log t}{C^w_a \Delta_a^2}\right\}$, and then $\mathbb{E}[T_a(t-1)]$ is still in $O(\log t)$ scale.

Fifth, similarly, as long as $C^w_a := \lim_{n\to+\infty} \frac{nC^w_{a,n}}{\log n} \geq \frac{8}{\Delta_a^2}$, i.e. either $C^w_{a,n} = \Omega(\frac{\log n}{n})$ or $C^w_{a,n} = \Theta(\frac{\log n}{n})$ but $C^w_{a,n} \leq \frac{8\log n}{n\Delta_a^2}, \forall n$, such a $u_{a,t}$ exists, but $u_{a,t}$ might be in $\Omega(\log t)$. Again, the exact value of $u_{a,t}$ is beyond our concern, because we aim to provide a bound for a general scenario. When $C^w_{a,n} \to 0$, $|b_{a,n}| = O\left(\sqrt{\frac{\log n}{n}}\right)$ is a sufficient condition of the existence of such a $u_{a,t}$.

Sixth, in contrast, when $C^w_{a,n}$ diminishes too fast, i.e. $C^w_{a,n} = o(\frac{\log n}{n})$, $\delta_{a,t,n}$ is no

longer a decreasing function of n. This implies $\delta_{a,t,t}$ might be greater than Δ_a for any arbitrarily large t. Hence, no feasible $u_{a,t}$ exists for large t and this theorem is not applicable to these cases. Again, the exact or approximate threshold of the arbitrarily large value of t is not related to this theorem which focuses on what we can bound for an estimator with good properties, i.e. large $C^w_{a,n}$ and small $|b_{a,n}|$.

Below, we show the full proof for Theorem B.1.1 and Theorem 3.4.1.

Proof. Let $a \in \mathcal{A}$, $\tau \in \mathcal{T}$, and $n := T_a(\tau - 1)$. And we derive probabilistic bounds for $\hat{\alpha}_{a,n}$,

$$\begin{aligned}
P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \varepsilon\right) &= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \geq \varepsilon\right) \\
&\leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \geq \varepsilon\right) \\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \varepsilon - |b_{a,n}|\right) \\
P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon\right) &= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \leq -\varepsilon\right) \\
&\leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] - |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \leq -\varepsilon\right) \\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\varepsilon + |b_{a,n}|\right)
\end{aligned}$$

Let $\bar{\varepsilon} := \varepsilon - |b_{a,n}|$. When $\bar{\varepsilon} > 0$, using the bounded difference inequality McDiarmid, 1989, we have

$$\begin{aligned}
P(\hat{\alpha}_{a,n} - \alpha_a \geq \varepsilon) \leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C^w_{a,n}\right) \\
P(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon) \leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C^w_{a,n}\right)
\end{aligned}$$

Set $\varepsilon = \delta_{a,\tau,n} = \sqrt{\frac{2\log\tau}{nC^w_{a,n}}}$, and thus $\bar{\varepsilon} = \sqrt{\frac{2\log\tau}{nC^w_{a,n}}} - |b_{a,n}|$. When $\bar{\varepsilon} > 0$, the above two

inequalities can be rewritten as

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right) \tag{B.1}$$

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right) \tag{B.2}$$

If a sub-optimal arm a is pulled at time τ, i.e. $\sigma_\tau = a$, we know that $B_{a,\tau,T_a(\tau-1)} \geq B_{a^*,\tau,T_{a^*}(\tau-1)}$, where a^* denotes the arm with maximum aptitude. This indicates either $B_{a,\tau,T_a(\tau-1)}$ is at least α_a or $B_{a^*,\tau,T_{a^*}(\tau-1)}$ underestimates α_{a^*} (or both), i.e. either $B_{a,\tau,T_a(\tau-1)} \geq \alpha_a$ or $B_{a^*,\tau,T_{a^*}(\tau-1)} \leq \alpha_{a^*}$ (or both). If arm a has been chosen at least $u_{a,t}$ times prior to this time, i.e. $T_a(\tau-1) \geq u_{a,t} = \frac{8\log\tau}{C_a^w\Delta_a^2}$, then $\Delta_a \geq 2\delta_{a,\tau,T_a(\tau-1)}$, which implies, if $B_{a,\tau,T_a(\tau-1)} \geq \alpha_a$, then $\hat{\alpha}_{a,\tau} - \delta_{a,\tau,T_a(\tau-1)} \geq \alpha_a$, i.e. even the 'lower bound' of arm a overestimates α_a. Therefore, if $\sigma_\tau = a$ and $T_a(\tau-1) \geq u_{a,t}$ for some τ, at least one of the following two inequalities holds

$$\hat{\alpha}_{a,T_a(\tau-1)} - \delta_{a,\tau,T_a(\tau-1)} \geq \alpha_a$$

$$\hat{\alpha}_{a^*,T_{a^*}(\tau-1)} + \delta_{a^*,\tau,T_{a^*(\tau-1)}} \leq \alpha_{a^*}$$

Now, by definition and the above results, the following inequalities hold for any real number $u > 1$

$$\begin{aligned}
T_a(t) &\leq u + \sum_{\tau=\lfloor u\rfloor+1}^{t} \mathbb{1}\left\{\sigma_\tau = a \wedge T_a(\tau-1) \geq u\right\} \\
&\leq u + \sum_{\tau=\lfloor u\rfloor+1}^{t} \mathbb{1}\left\{B_{a,\tau,T_a(\tau-1)} \geq B_{a^*,\tau,T_{a^*}(\tau-1)} \wedge T_a(\tau-1) \geq u\right\} \\
&\leq u + \sum_{\tau=\lfloor u\rfloor+1}^{t} \mathbb{1}\left\{\exists v \in \{\lfloor u\rfloor,\ldots,\tau-1\}, v^* \in \{1,\ldots,\tau-1\} : B_{a,\tau,v} \geq B_{a^*,\tau,v^*}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \sum_{v=\lfloor u \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \mathbb{1}\left\{B_{a,\tau,v} \geq B_{a^*,\tau,v^*}\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \sum_{v=\lfloor u \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \mathbb{1}\left\{\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a \vee \hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \sum_{v=\lfloor u \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\mathbb{1}\left\{\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right\} + \mathbb{1}\left\{\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right\}\right)
\end{aligned}
$$

Set $u = u_{a,t}$, take the expectation on both side and we have

$$
\begin{aligned}
\mathbb{E}[T_a(t)] &\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} \sum_{v=\lfloor u_{a,t} \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(P\left(\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right) + P\left(\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} \sum_{v=\lfloor u_{a,t} \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,v}|\sqrt{vC^w_{a,v}}\right)^2\right)\right. \\
&\qquad \left. + \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a^*,v^*}|\sqrt{v^*C^w_{a^*,v^*}}\right)^2\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} \sum_{v=\lfloor u_{a,t} \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} 2\exp\left(-2\left(\frac{9}{10}\sqrt{2\log\tau}\right)^2\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} 2\tau^2 \exp\left(-\frac{324}{100}\log\tau\right) \\
&\leq u_{a,t} + 2\sum_{\tau=1}^{+\infty} \tau^{-\frac{124}{100}} \\
&= u_{a,t} + 2\zeta(1.24)
\end{aligned}
$$

The third inequality holds because $b_{a^*,v^*} \leq \frac{1}{10}\sqrt{\frac{2\log\tau}{v^*C^\omega_{a^*,v^*}}}$.

Once we have the bounds of $\mathbb{E}[T_a(t-1)]$, we can directly derive the bounds for total regret. Let $\bar{r}_a := \sup_s r_{a,s}$ and $\underline{r}_a := \inf_s r_{a,s}$, then

$$
\mathbb{E}[R(t)] \leq \sum_{a \neq a^*} \mathbb{E}[(\bar{r}_{a^*} - \underline{r}_a)T_a(t-1)] \leq \sum_{a \neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(u_{a,t} + 2\zeta(1.24)\right) \tag{B.3}
$$

□

B.1.2 Proof of Proposition 3.4.1

Proof. Let $\omega_i^* = \inf w_i$. By definition, we know $\omega_i^* \leq 1$, as the image set of φ is $[0,1]$, thus $C_n^* = \frac{1}{n\sum_{i=1}^n \omega_i^{*2}} \geq \frac{1}{n\sum_{i=1}^n 1^2} = \frac{1}{n^2}$, proving the left inequality. Before we proceed to prove the right inequality, we briefly introduce Chebyshev's sum inequality (Hardy et al., 1952):

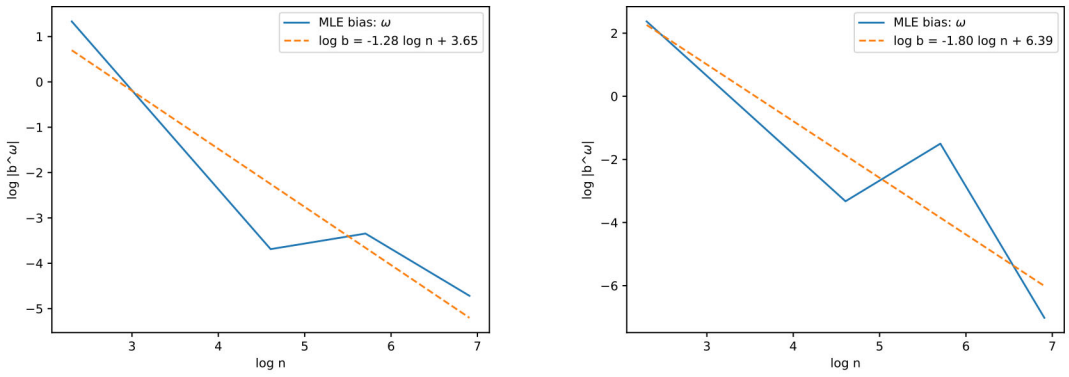
Lemma B.1.1 (Chebyshev's sum inequality)**.** Suppose $c_1, \ldots, c_n, b_1, \ldots, b_n \in \mathbb{R}$ such that $c_1 \geq c_2 \geq \ldots c_n$ and $b_1 \geq b_2 \geq \ldots b_n$, and then $\frac{1}{n}\sum_{i=1}^n c_i b_i \geq \left(\frac{1}{n}\sum_{i=1}^n c_i\right)\left(\frac{1}{n}\sum_{i=1}^n b_i\right)$.

By Chebyshev's sum inequality, $\sum_{i=1}^n w_i^* \leq \sqrt{n\sum_{i=1}^n w_i^{*2}} = \sqrt{\frac{1}{C_n^*}}$. Suppose by contradiction that $C_n^* > 1$, that is $\sum_{i=1}^n w_i^* \leq \sqrt{n\sum_{i=1}^n w_i^{*2}} = \sqrt{\frac{1}{C_n^*}} < 1$ Using Chebyshev's sum inequality, for any two points $x, x' \in \mathcal{X}^n$, $|\varphi(x) - \varphi(x')| \leq \sum_{i=1}^n w_i^* \leq \sqrt{\frac{1}{C_n^*}} < 1$. This indicates that the image set of φ has a length at most C^* that is strictly less than 1, which contradicts the assumption that φ has an image set of length 1. Thus, $C_n^* \leq 1$, the right inequality holds. □

B.2 More on Bias Conditions in Example 3.4.3

Figure B.1 shows the bias decay rates of $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$. We might be interested in $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$'s bias decay rates for general dynamic learning problems with unknown vector parameters. For Theorem 3.4.1 to hold in the SLT problem that focuses on the long-term, full potentials of arms, we only need to verify the bias

conditions for $\alpha_{1,n}^{MLE}$ and $\alpha_{2,n}^{MLE}$.



(a) The bias of $\omega_{1,n}^{MLE}$ is $O(n^{-1.28})$ **(b)** The bias of $\omega_{2,n}^{MLE}$ is $O(n^{-1.80})$

Figure B.1: Verifying bias scales of $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$. The bias scales are both $o\left(\sqrt{\frac{\log n}{n}}\right)$; although not needed, we can see that the bias decay rates of $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$ satisfy the bias condition in Theorem 3.4.1.

B.3 Proof of Bandits with Delayed Feedback in Section 3.6.1

Let $\hat{\hat{\alpha}}_{a,n}$ denote our point estimate of α_a when $T_a(t) = n$ and up to k_a true rewards have not been revealed but reward estimates are available.

Proof. Let $a \in \mathcal{A}$, $\tau \in \mathcal{T}$, and $n := T_a(\tau - 1)$. Below we derive probabilistic bounds for $\hat{\hat{\alpha}}_{a,n}$,

$$
\begin{aligned}
P\left(\hat{\hat{\alpha}}_{a,n} - \alpha_a \geq \varepsilon\right) &= P\left(\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n} + \hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \geq \varepsilon\right) \\
&\leq P\left(|\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n}| + (\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}]) + |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \geq \varepsilon\right) \\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \varepsilon - |b_{a,n}| - |e_{a,n}|\right) \\
P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon\right) &= P\left(\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n} + \hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \leq -\varepsilon\right) \\
&\leq P\left(-|\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n}| + (\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}]) - |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \leq -\varepsilon\right)
\end{aligned}
$$

$$= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\varepsilon + |b_{a,n}| + |e_{a,n}|\right)$$

Let $\bar{\varepsilon} := \varepsilon - |b_{a,n}| - |e_{a,n}|$. When $\bar{\varepsilon} > 0$, using the bounded difference inequality McDiarmid, 1989, we have

$$P(\hat{\alpha}_{a,n} - \alpha_a \geq \varepsilon) \leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C_{a,n}^w\right)$$
$$P(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon) \leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C_{a,n}^w\right)$$

Set $\varepsilon = \delta_{a,\tau,n} = \sqrt{\frac{2\log\tau}{nC_{a,n}^w}}$, and thus $\bar{\varepsilon} = \sqrt{\frac{2\log\tau}{nC_{a,n}^w}} - |b_{a,n}| - |e_{a,n}|$. When $\bar{\varepsilon} > 0$, the above two inequalities can be rewritten as

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w} - |e_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right) \tag{B.4}$$

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w} - |e_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right) \tag{B.5}$$

The rest of the proof follows that of Theorem 3.4.1 in Section B.1.1. The only minor changes are needed after we set $u = u_{a,t}$ and take the expectation on both side; we have

$$\begin{aligned}
\mathbb{E}[T_a(t)] &\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(P\left(\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right) + P\left(\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,v}|\sqrt{vC_{a,v}^w} - |e_{a,v}|\sqrt{vC_{a,v}^w}\right)^2\right)\right. \\
&\qquad \left. + \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a^*,v^*}|\sqrt{v^*C_{a^*,v^*}^w} - |e_{a^*,v^*}|\sqrt{v^*C_{a^*,v^*}^w}\right)^2\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} 2\exp\left(-2\left(\left(1-\frac{1}{10}-\frac{1}{40}\right)\sqrt{2\log\tau}\right)^2\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} 2\tau^2 \exp\left(-\frac{1225}{400}\log\tau\right) \\
&\leq u_{a,t} + 2\sum_{\tau=1}^{+\infty} \tau^{-\frac{425}{400}} \\
&= u_{a,t} + 2\zeta(1.063)
\end{aligned}
$$

The third inequality holds because $b_{a,v} \leq \frac{1}{10}\sqrt{\frac{2\log v}{vC^{\omega}_{a,v}}} \leq \frac{1}{10}\sqrt{\frac{2\log\tau}{vC^{\omega}_{a,v}}}$ and $b_{a^*,v^*} \leq \frac{1}{10}\sqrt{\frac{2\log v^*}{v^*C^{\omega}_{a^*,v^*}}} \leq \frac{1}{10}\sqrt{\frac{2\log\tau}{v^*C^{\omega}_{a^*,v^*}}}$. Similarly, $e_{a,v} \leq \frac{1}{40}\sqrt{\frac{2\log v}{vC^{\omega}_{a,v}}} \leq \frac{1}{40}\sqrt{\frac{2\log\tau}{vC^{\omega}_{a,v}}}$ and $e_{a^*,v^*} \leq \frac{1}{40}\sqrt{\frac{2\log v^*}{v^*C^{\omega}_{a^*,v^*}}} \leq \frac{1}{40}\sqrt{\frac{2\log\tau}{v^*C^{\omega}_{a^*,v^*}}}$.

Once we have the bounds of $\mathbb{E}[T_a(t-1)]$, we can directly derive the bounds for total regret. Let $\bar{r}_a := \sup_s r_{a,s}$ and $\underline{r}_a := \inf_s r_{a,s}$, then

$$
\mathbb{E}[R(t)] \leq \sum_{a\neq a^*} \mathbb{E}[(\bar{r}_{a^*} - \underline{r}_a)T_a(t-1)] \leq \sum_{a\neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(u_{a,t} + 2\zeta(1.063)\right) \tag{B.6}
$$

□

B.4 Proof of Theorem 3.5.1: FL-UCB Regret Bounds

Proof. Proof of FL-UCB regret bounds: First, we consider the LP defined by (3.11) ∼ (3.15). For the sake of notational simplicity and generality, we write it in the standard form

$$
\max_{z} \quad f(z) \tag{B.7}
$$

$$
s.t. \quad z \in C_{\text{set}} \tag{B.8}
$$

where $C_{\text{set}} \in \mathbb{R}^{|\mathcal{A}|}$ is a nonempty convex set and $f : \mathbb{R}^{|\mathcal{A}_|} \mapsto \mathbb{R}^{|\mathcal{A}|}$ is a convex function. A point z^* is optimal for the convex optimization problem (B.7) $\sim$ (B.8) if

$$\exists \xi \in \mathbb{R}^{|\mathcal{A}|} \backslash \{\mathbf{0}\} \qquad s.t. \quad \xi(z - z^*) \leq 0, \quad \forall z \in C_{\text{set}} \tag{B.9}$$

Before we further analyze the optimality criterion (B.9), we define the concept of *normal cones*.

Definition B.4.1. The normal cone of a closed, convex set $C_{\text{set}} \in \mathbb{R}^n$is

$$N_C(z^*) = \begin{cases} \{\xi \in \mathbb{R}^n | (\forall z \in C_{\text{set}}) \xi^T (z - z^*) \leq 0\} & \text{if } z^* \in C_{\text{set}} \\ \emptyset & \text{if } z^* \notin C_{\text{set}} \end{cases} \tag{B.10}$$

(B.9) is equivalent to requiring that $\xi \in N_C(z^*) \backslash \{\mathbf{0}\}$. To find the normal cone of the feasible region defined in (3.12) $\sim$ (3.15), we need to use the following lemma

Lemma B.4.1. Let $A \in \mathbb{R}^{m \times n}$ and let $b \in \mathbb{R}^m$. Consider the polyhedron $Q(A, b) = \{x \mid Ax \leq b\}$. Suppose $x \in Q(A, b)$, then $N_{Q(A,b)}(x) = \{A^T y | y \in \mathbb{R}^m \text{ such that } y \geq 0 \text{ and } y^T(b - Ax) = 0\}$.

The optimal solution to (3.11) $\sim$ (3.15) is:

$$z_a^* = \begin{cases} \theta_a^A, & a \in \mathcal{A}_A \setminus \mathcal{A}_{BK} \\ \theta_a^{BK}, & a \in \mathcal{A}_{BK} \setminus \mathcal{A}_A \setminus \{a^*\} \\ \max\{\theta_a^{BK}, \theta_a^A\} & a \in \mathcal{A}_{BK} \cap \mathcal{A}_A \setminus \{a^*\} \\ 0, & a \notin \mathcal{A}_{BK} \cup \mathcal{A}_A \\ 1 - \sum_{a \in \mathcal{A} \setminus \{a^*\}} z_a^*, & a = a^* \end{cases} \tag{B.11}$$

Therefore, the normal cone at an optimal solution z^* for the convex set Q_{FUCB} as defined by (3.12) $\sim$ (3.15) is a convex cone defined by the following inequalities, assuming $\mathcal{A}_{BK}$ is known (we will estimate it later):

$$B_{a^*,T_{a^*}(t-1)} \geq B_{a,T_a(t-1)} \qquad \forall a \in \mathcal{A} \tag{B.12}$$

$$B_{a,T_a(t-1)} \geq 0 \qquad \forall a \in \mathcal{A} \tag{B.13}$$

If we replace α with $B_{s_{t-1}} := B_{\mathcal{A},T_{\mathcal{A}}(t-1),s_{\mathcal{A},t-1}} = [B_{a,T_a(t-1),s_{a,t-1}}]_{a=1}^{|\mathcal{A}|}$, as long as $B_{s_{t-1}} \in N_C(z^*)$, the optimal basis stays optimal for the new LP problem with the objective defined by (3.16). (B.12) $\sim$ (B.13) show that as long as we are able to identify a^* (the unique solution of the new LP is the optimal) and other top-K arns using the UCB indexes soon enough and only explore the other arms rarely afterwards.

Moreover, we need to bound the regret incurred while estimating the members of $\mathcal{A}_{BK}$ and the ordering. Specifically, we want to distinguish the difference between the k-th best arm $a^{(k)}$ and the $k+i$-th best arm $a^{(k+i)}$, $\forall i \in \{1, \ldots, |\mathcal{A}| - k\}$. The

proof of regret bound in distinguishing the k-th and the $(k+i)$-th best arm is analogous to that of proving L-UCB regret upper bounds: The difference is that we are not only interested in a^* or $a^{(1)}$, but also $a^{(k)}$ for $k \in \{2, \ldots, K\}$. Specifically, to compute the expected number of pulls of $a^{(k+i)}$ when we actually want to pull $a^{(k)}$, we choose

$$u_{a^{(k+i)}}^{a^{(k)}} = \bar{T} := \frac{8 \log t}{C_a^w \Delta^2_{a^{(k)}, a^{(k+i)}}}$$

where $\Delta_{a^{(i)}, a^{(j)}} = \alpha_{(i)} - \alpha_{(j)}$, $\forall i, j \in \{1, \ldots, |\mathcal{A}|\}$. The number of times that an arm $a^{(k+i)}$ is mistaken in the $\mathcal{A}_{BK}$ as $a^{(k)}$ set is bounded by

$$\frac{8 \log t}{C_a^w \Delta^2_{a^{(k)}, a^{(k+i)}}} + 2\zeta(1.24)$$

Therefore, the expected number of times we pull "worse" arms (whose true parameters are worse than those of the ones we intend to pull) when imposing BK-fairness, is bounded by

$$\sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} \left(\frac{8 \log t}{C_a^w \Delta^2_{a^{(k)}, a^{(k+i)}}} + 2\zeta(1.24) \right)$$

And thus the expected regret when imposing BK-fairness is bounded by

$$\begin{aligned} \mathbb{E}[R^{BK}(t)] &= \sum_t r_{a^*, T_{a^*}, T_{a^*}(t-1)} - r_{a_t, T_a(t-1)} \leq \sum_a (\bar{r}_{a^*} - \underline{r}_a)\, \mathbb{E}[T_a] \\ &\leq \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} (\bar{r}_{a^*} - \underline{r}_{a^{(i)}}) \left(\frac{8 \log t}{C_a^w \Delta^2_{a^{(k)}, a^{(k+i)}}} + 2\zeta(1.24) \right) \end{aligned}$$

Note that imposing BK- or AA-fairness incurs linear PoF as long as $\theta^A + \theta^{BK} \neq \mathbf{0}$, which is not counted as part of *F-regret* or regret. Moreover, $\mathcal{A}_A$ is assumed to be known based on inherent, known arm features and thus does not require estimation.

□

B.5 Extension: Arm Correlation

In this section we study bandit problems where the learning processes of arms are correlated. Specifically, we study bandits where arm experience is correlated in a linear fashion: Linear correlation is among the most common dependence patterns in the literature, and its mathematical simplicity enables us to derive clean analytical results that shed light on the influence of arm dependence on bandits. We have proved that the regret bounds of bandits with mutually independent learning arms are $O(\log t)$. We will show that similar results hold when arms are correlated.

In our SLT problem setting, the arms are patient-TC-surgery tuples. Arm correlation may arise from the fact that the skill sets required to perform successful SLT surgeries of various types typically overlap. This translates to a bandit problem where an arm's hidden parameter might change along with its the learning curve, even if that particular arm is not chosen.

B.5.1 Experience-Correlated Bandits

As discussed in Section 3.6, the skills learned from different surgeries could be partially transferable as the skill sets required for similar surgeries may overlap. We consider linear correlation based on experience in bandit contexts; we explicitly

define bandits with this particular form of arm dependence.

Definition B.5.1. (Experience-Correlated Bandit) A bandit problem is experience-correlated if the experience score $s_{a,t-1}$ of an arm $a \in \mathcal{A}$ can be written as

$$s_a(t) = s_a(t-1) + \sum_{j \in \mathcal{A}^a} \beta_{a,j,t} \mathbb{1}(a_t = j) \qquad \beta_{a,j,t} \geq 0, \forall t \geq 1, \tag{B.14}$$

where $\mathcal{A}^a \neq \emptyset$ is the set of arms that are correlated with a.

When arms are uncorrelated/independent, $\mathcal{A}^a = \{a\}$ and $\beta_{a,a,t} = 1, \forall t$. When $\exists j \neq a, j \in \mathcal{A}^a$, s.t. $\beta_{a,j,t} > 0$, we say arm a is dependent on arm j. In this case, $s_a(t) \geq T_a(t)$, with this inequality being strict for at least one a: $T_a(t)$ is the number of times that an arm has been pulled (affecting the total regret), while $s_a(t)$ affects the current proficiency parameter, $\theta_a(\alpha_a, s_{a,t})$. Here, correlation affects learning. And we demonstrate through the examples below that such problems can be challenging, in general.

Unlike the vanilla bandit problem, the optimal policy of an experience-correlated bandit is not a straightforward stationary policy, i.e., always pulling the arm with the highest aptitude α^* may turn out to be a sub-optimal strategy in both large-t the long-term regime and small-t the short-term regime. Consider the following example:

Example B.5.1. Consider an experience-correlated bandit with two arms: arm 1

and arm 2. The learning curves and correlations are explicitly known:

$$\theta_{1,t} := l_1(s_1(t)) = 0.5 + \min\left\{\frac{0.5 s_1}{100}, 0.5\right\} \quad t \geq 1 \tag{B.15}$$

$$\theta_{2,t} := l_2(s_2(t)) = \min\left\{\frac{s_2}{100}, 0.9\right\} \quad t \geq 1 \tag{B.16}$$

$$s_1(t) = s_1(t-1) + \mathbb{1}(a_t = 1) + 100 \cdot \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{B.17}$$

$$s_2(t) = s_2(t-1) + \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{B.18}$$

where $s_1(0) = s_2(0) = 0$.

The optimal policy is to pull arm 2 at $t = 1$, and choose arm 1 when $t \geq 2$.

Proof. Under π^*, the total expected reward is
$\mathbb{E}\left[\sum_t r_{a_t, s_{a_t,t-1}}\right] = 0 + 1 + \cdots + 1 = T - 1$.

First, we show that if arm 2 has been pulled once, we should always pull arm 1 in later rounds. Because pulling arm 2 once will guarantee that $s_1(t) \geq 100$ and $\theta_1(t) = 1$; thus, the expected reward of pulling arm 1 in a later time will yield the highest possible expected single period reward ($= 1$), while pulling arm 2 will give no more than 0.9 expected reward. As a result, in an optimal policy, once arm 2 is pulled, arm 1 should always be chosen in later rounds.

Now, we show that we will pull arm 2 at least once. If we never pull arm 2, then we always pull arm 1; the expected total reward under this policy is
$0.5 + 0.505 + 0.51 + 0.515 + \cdots + 0.995 + 1 \times (T - 99) < T - 1$. Therefore, π^* has higher expected total rewards compared to the policy that pulls arm 1 throughout the time horizon.

Finally, we show that we will pull arm 2 precisely at $t = 1$. If we follow a policy π' that first pulls arm 2 at $t = k, k \in \{2, \cdots, t\}$, then the expected reward at $t = 1$, $\mathbb{E}r_{1,0}(1) = 0.5$. The expected total reward of the policy π',
$\mathbb{E}\left[\sum_t r^{\pi'}(t)\right] \leq r^{\pi'}(1) + r^{\pi'}(k) + 1 \times (T-2) = 0.5 + 0 + T - 2 = T - 1.5 < T - 1$.
Therefore, π^* has higher expected total reward compared to any policy that dictates pulling arm 2 at time $t = 1$.

The arguments above show that π^* is the optimal policy.

In many applications the solution to the offline experience-based bandits can be found using dynamic programming. The following example shows that the optimal policy for our problem may require switching arms more than once and revisiting an arm.

Example B.5.2. Consider an experience-correlated bandit with two arms, arm 1 and arm 2. The learning curves and correlations are explicitly known:

$$\theta_1(t) := l_1(s_1(t-1)) = 0.5 + \min\{\frac{0.5 s_1}{100}, 0.5\} \quad t \geq 1 \tag{B.19}$$

$$\theta_2(t) := l_2(s_2(t-1)) = \min\{\frac{s_2}{100}, 0.9\} \quad t \geq 1 \tag{B.20}$$

$$s_1(t) = s_1(t-1) + \mathbb{1}(a_t = 1) + 100 \cdot \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{B.21}$$

$$s_2(t) = s_2(t-1) + 100 \cdot \mathbb{1}(a_t = 1) + \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{B.22}$$

where $s_1(0) = s_2(0) = 0$.

The optimal policy is to choose arm 1 at $t = 1$, choose arm 2 at $t = 2$, and then choose arm 1 when $t \geq 3$.

Proof. First, we prove that in the optimal policy, we pull arm 1 and 2 each at least once. The expected reward of always choosing arm 1 is $0.5 + 0.505 + \cdots + 0.995 + 1 \times (T - 100) < T - 1$; similarly, the expected reward of always choosing arm 2 is $0 + 0.01 + \cdots + 0.89 + 0.9 \times (T - 90) < T - 1$. However, the expected total reward of π^* is $0.5 + 0.9 + 1 \times (T - 2) = T - 0.6 > T - 1$. Therefore, both stationary policies cannot be optimal.

Next, we show that we pull arm 2 at most once. Now we know that both arms are chosen at least once in the optimal policy. Suppose we have pulled arm 2 at time t', then pulling arm 2 at any time $t'' > t'$ will not increase $\theta_1(t'')$, but will yield a lower immediate reward; thus, the marginal benefit of choosing arm 2 at t'' is strictly negative. As a result, we choose arm 2 exactly once.

Finally, we prove that we choose arm 2 at $t = 2$. If we choose arm 2 at $t = 1$, then the expected total reward is $0 + 1 \times (T - 1) = T - 1 < T - 0.6$, thus choosing arm 2 at $t = 2$ is better than pulling arm 2 at $t = 1$. If a policy π' pulls arm 2 at $t = k, k \in \{3, \cdots, T\}$, then the expected total reward $\mathbb{E}[\sum_t r^{\pi'}(t)] < r^{\pi'}(1) + r^{\pi'}(2) + r^{\pi'}(k) + 1 \times (T - 3) \leq 0.5 + 0.505 + 0.9 + T - 3 = T - 1.05 < T - 0.6$. In summary, the optimal policy is to pull arm 2 at $t = 2$. □

Example B.5.2 shows that in the optimal policy, a low immediate-reward, high contributed-experience arm (arm 2) may be chosen after higher-reward, low experience arms in an optimal strategy. An explanation for the fluidity and complexities is that the hidden parameters of arms may change when any arms are pulled, and depending on the specific structure of learning curves and correlation patterns, an arm that is useless at one time may be incredibly useful in later rounds.

This being said, imposing conditions on the correlation would potentially yield structural results on the optimal policy. Such an exploration is deferred for future work. □

B.5.2 Heterogeneous livers

We can incorporate liver heterogeneity in two ways. The most straightforward approach is to formulate parallel MABs, one for each liver type; within each MAB, livers are viewed as homogeneous. This formulation is practical and realistic, as livers are often allocated to TCs and patients within the geographical region in which they are acquired (see Section B.7 for more information). Surgical experience and expertise at one TC rarely transfer to another that is geographically remote.

Alternatively, we can incorporate liver heterogeneity in one MAB, i.e., $|\mathcal{L}| > 1$, implying that for each $\ell \in \mathcal{L}$, all arms in $\mathcal{A}$ can be pulled, i.e., a liver of type ℓ can be allocated to any TC and any patient type. This formulation can be helpful when we are interested in a more granular classification of liver types within the same geographical region, as experience gained from operating with different liver types is carried forward. Our FL-UCB algorithms apply to the case $|\mathcal{L}| > 1$, except that we estimate $\hat{\alpha}^{\ell}_{a,n}$ for each $\ell \in \mathcal{L}$. All theoretical regret bounds hold (the upper bounds for heterogeneous livers are $|\mathcal{L}|$ times the original bounds for the homogeneous case). The actual regrets might be much lower, as surgical experience transfers and accumulates faster.

B.6 More Details about the Numerical Study in Section 3.7

B.6.1 Details about the SLT Simulation Setup

Below we detail how we estimate α's from the STAR files. For each medically-splittable liver, it can save two patients' lives. In current SLT practice, the smaller left lobe is usually allocated to a sick child. The other half, depending on its size and the patient waitlists, may be allocated to a small adult/big child or a medium adult. There is a liver-splitting technique that allows a more even splitting of a donor's liver and thus can save two small or medium adults' lives. The two partial livers can be used for two recipients at two different transplant centers; thus, we view each partial liver arrival as an independent time step. A partial liver may be shared across a large geographical area; see (UNOS, 2021) for detail about the acuity circles policy.

Currently, a splittable liver may be shared across a large geographical area; see the acuity circles policy UNOS, 2021 for detail. We consider a 500 nautical mile circle that includes OPTN regions 2, 9, 10, 11, and Wisconsin and Illinois (URL: https://optn.transplant.hrsa.gov/about/regions/). In 2022, there were around 8000 donated livers and 10 big transplant centers in the 500NM Circle. (See https://optn.transplant.hrsa.gov/data/view-data-reports/regional-data/ for more detail.)

Each (partial) liver graft can be allocated to a patient within one of the five health condition groups. Patients' health conditions are described by the Model for End-Stage Liver Disease (MELD) score (for adults) and Pediatric End-Stage Liver

Disease (PELD) score (for children), which are indicators of medical urgency. MELD and PELD scores take integer values in $[6, 40]$; for critically sick patients, there are 1A, 1B, 2A, and 2B special urgent categories. We divide the patients into five score buckets: ≥ 40 (including MELD/PELD $= 40$ and critically sick patients), $35 \sim 39$, $30 \sim 34$, $20 \sim 29$, $6 \sim 19$. The current OPTN system allocates (whole) livers preferentially to eligible patients with the highest scores (the sickest patients) (Emre & Umman, 2011); SLT surgeries are rarely performed, but the current SLT patient matching does not strictly follow the "sickest-first" rule, due to lack of policy clarity in matching the secondary recipient, and the primary recipient is often a child. Since SLT is a challenging medical procedure and saves twice as many lives, it makes sense to consider allocating partial livers to healthier patients to maximize overall survival and welfare.

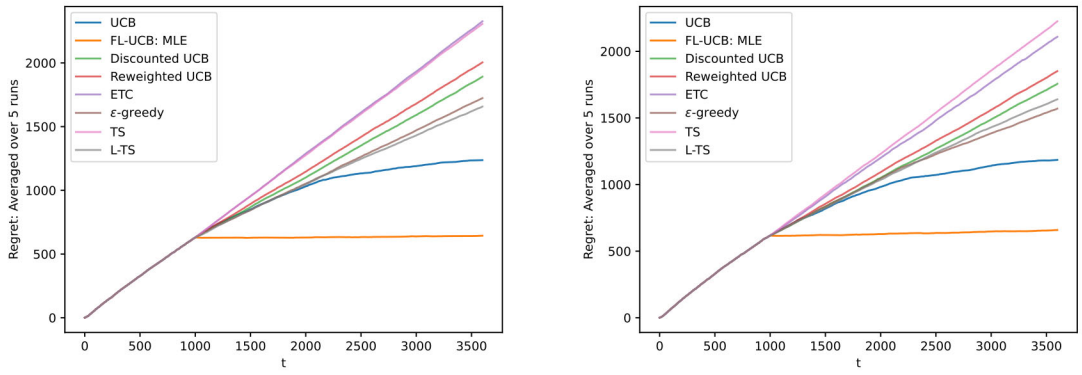
Therefore, in total, we have $10 \times 5 = 50$ arms for the livers splittable in the geographical region of interest. Recall that 10% of livers are medically safe to split (OPTN & UNOS, 2016), so at least 800 livers can be used for SLT a year in the 500NM Circle, with each liver supporting two SLT surgeries. A total of 1600 SLT surgeries are possible. Livers are heterogeneous; among the medically safe livers, it is estimated that $\sim 63\%$ (Perito et al., 2019), or around 1008 of them, satisfy the strictest medical criteria and thus are of the highest quality. In our simulation, we consider allocating these high-quality livers to patients and TCs in the 500NM geographical circle. See Section B.6 for more details about the allocation of high-quality livers acquired in different geographical regions (i.e., heterogeneity) and please refer to Section B.7 for more facts about current SLT practice in the US.

The α 's are drawn from $(0.3, 0.95)$, where the upper and lower bounds of the range are estimated directly from the STAR files: We compute the 1-year graft survival for different surgery technique types in each geographical region; these statistics are then used for simulate the distribution and range of SLT's 1-year survival outcomes. These statistics of past surgeries (WLTs and a small number of SLTs) show that 1-year graft survival range from 0.33 to 1. Retrospective reviews and anecdotal accounts report that SLT outcomes can be comparable and as good as WLT outcomes in few, proficient TCs that have gained SLT mastery through a good amount of experience (Duke Health, 2021; Hackl et al., 2018). Since SLT is a more complex surgery by nature, we adjust the lower limit of the 1-year graft survival rate to 0.3.

B.6.2 Outcome Prediction Accuracy and Uncertainty Quantification

In Section 3.7 we assume the prediction accuracy in SLT is 60%; Figure B.2 shows results assuming the prediction accuracy is 85%.

Similar to the case where the prediction accuracy is 60%, FL-UCB with MLE estimation has the lowest regrets and converges fast when an 85%-accurate estimate is available. However, with a higher accuracy level, the UCB performance is significantly improved and is second only to FL-UCB; its regrets also show signs of convergence at $t = 3600$.



(a) Regret w/o fairness

(b) F-regret w/ fairness

Figure B.2: Comparing FL-UCB regret against benchmarks when medical learning exists and assuming there is a 1-year delay in observing true rewards (the rollout policy is described in Section 3.7.1). Estimates based on demographics and perioperative clinical metrics are available and are 85% accurate.

B.6.3 More about the Bandit Algorithms Used for Comparison

Procedure 3: L-TS Algorithm Pseudo Code

1: **Initialization:** Choose prior distributions $Beta(\tilde{\alpha}_{a,0}, \tilde{\beta}_{a,0})$, $\forall a \in \mathcal{A}$.

Select each arm a m_a times. Update posterior distribution as in Step 3.

2: **Select arm:** Sample $a_t \sim Beta(\tilde{\alpha}_{a,n}, \tilde{\beta}_{a,n})$

3: **Update distribution:** $\tilde{\alpha}_{a_t,T_{a_t}(t-1)}, \tilde{\beta}_{a_t,T_{a_t}(t-1)}) \leftarrow$

$\Big((\tilde{\alpha}_{a,T_a(t-1)} + r^t(1+\exp(\omega_{a_t} - T_{a,t-1})), \tilde{\beta}_{a,n-1} + (1 - r^t(1+\exp(\omega_{a_t} - T_{a,t-1}))\Big)$

4: **Increment** t, $T_{a_t,t} = T_{a_t,t-1} + 1$ and **Go to Step 2**

In TS, we update the posterior distribution using

$(\tilde{\alpha}_{a_t,T_{a_t}(t-1)}, \tilde{\beta}_{a_t,T_{a_t}(t-1)}) \leftarrow (\tilde{\alpha}_{a,T_a(t-1)} + r^t, \tilde{\beta}_{a,n-1} + (1 - r^t))$. Recall that r^t is the random reward (or the estimated reward) at time t. In our numerical study, we choose $m_a = 20$ and $(\tilde{\alpha}_{a,0}, \tilde{\beta}_{a,0}) = (2, 2)$ for all a in both L-TS and TS. For ETC algorithm implemented, the exploitation starts once the 500 rounds of round robin

conclude. In ϵ-greedy, with probability 0.95 we greedily choose the arm with the highest estimated reward (breaking ties arbitrarily) and we explore arms with equal probability when not exploiting. In our discounted UCB, we use $\delta = 0.9$.

B.7 Current SLT Practice in the US

To better understand how SLTs are practiced, we consulted with senior surgeons from the globally renowned transplant center affiliated with the University of California, San Francisco (UCSF). In the US, after graduating from medical schools and having chosen their specialization areas, surgeons complete their residency programs to obtain an unrestricted license to practice medicine and a board certificate for their chosen surgical specialty, in our case, the liver transplant. It is during residency that prospective surgeons may learn SLTs at selected TCs, such as the one at UCSF. Such residency programs involve assisting with actual SLT surgeries. Besides graduation requirements that enable some residents to learn SLT, young physicians may be intrinsically interested in saving more lives, expanding their skill sets, and mastering the techniques to perform complex surgeries; extrinsic motivations such as recognition from the surgical community and income increase brought by more transplants can also incentivize medical learning and overcome risk aversion.

In practice, successful SLTs involve a complicated process, including registration, procurement, allocation, logistics, surgical operations, and post-surgery recovery. To start, eligible ESLD patients choose transplant centers and register for the national liver transplant waitlists. When a deceased-donor liver becomes available and is being evaluated to determine whether it is medically splittable (based on donor age,

body mass index, size, etc.), OPTN (which is administered by United Networks for Organ Sharing, short for UNOS) generates a ranked list (known as the *match-run*), based on computerized algorithms. The organ is offered to the match-run candidates sequentially until a candidate/candidate pair accepts it. The longer it takes between the removal of blood supply from the deceased-donor organ and the transplantation into the recipient(s) (the *cold ischemia time*), the more the organ's quality deteriorates. An organ is discarded if the cold ischemia time is determined to be too long (exceeding 12 - 18 hours). From our discussions with UCSF transplant physicians and transplant software professionals, we learned that transplant centers could also make provisional offers to more than one patient to reduce organ waste and maximize societal welfare. Therefore, practically speaking, UNOS/OPTN essentially assigns livers to transplant centers and certain patient health groups; at the center level, the medical teams on call perform the surgery with the assigned recipients or make adjustments under OPTN guidelines when necessary.

Once a liver is accepted, the organ is harvested (and split if to be used in SLTs) by a trained team at the donor hospital. The matching of transplant surgeons and candidates is finalized after a candidate has accepted an offer and right before the surgery. After being harvested, the procured split liver grafts are transported to the SLT recipient TCs, where the two transplant teams perform the SLT surgeries. After that, patient recoveries occur. Currently, most SLTs are performed in few major transplant centers; thus, the primary recipients (usually children) and secondary recipients are usually within the same TC. However, because of UNOS's new acuity circles policy that took effect in 2019 (UNOS, 2021) , patients from different TCs within the acuity circles may receive halves of the same donor liver more frequently.

The acuity circles policy aims to enable broader organ sharing but has been controversial due to challenging logistics, incentive misalignment, and increased organ waste. There has also been debate over the transplant objective itself: Should we allocate livers to the sickest patient(s) within the 500NM circle? Is the current "sickest-first" principle simply preventing more immediate deaths but not optimizing societal welfare (e.g., survival outcomes, quality-adjust life years, equity)?

B.8 More on Related Work

UCB variants for nonstationary environments: Garivier and Moulines, 2011 consider abrupt changing environments where the reward distributions may remain constant for epochs and change at unknown breakpoints. The authors proposed D-UCB and SW-UCB policies to overcome environment nonstationarity. Similar to the idea of our discounted UCB, D-UCB averages past rewards with a discount factor which gives more weight to recent observations. The difference between our discounted UCB and D-UCB is in the padding function (i.e., the term added to our estimate of the arm parameter). Their SW-UCB relies on a local empirical average of the last few plays, leaving out earlier observations. Alternatively, we proposed reweighted UCB to discount the past observations less aggressively, as in our SLT problem the reward distribution changes gradually.

Bandits and queueing: Our MAB formulation can further be enhanced by adding a constant (computed by a separate queueing module) to each arm's full potential to capture the endogenizing queueing effects (e.g., number of patient deaths while waiting for transplants) while keeping the problem stateless. Previous attempts to incorporate queueing or non-stationarity into bandit problems have utilized stateful

formulations, for instance, *restless bandits* (Bertsimas & Niño-Mora, 2000; Jacko, 2010; Krishnasamy et al., 2016; Whittle, 1988). Such explicit modeling would likely render our SLT learning problem intractable; thus, analyzing queueing dynamics via a separate module sounds more viable: We may incorporate them into a subroutine of our proposed algorithm to maintain a stateless bandit. Compared to these previous works, the objective in the SLT problem is also markedly different: For example, Krishnasamy et al., 2016 combined MAB with queueing by designating the MAB's rewards as queue lengths; while Whittle, 1988 and Bertsimas and Niño-Mora, 2000 did not incorporate queueing behaviors in their analysis. In our problem, the reward functions can be written as convex combinations of immediate rewards and queueing metrics, both of which are functions of the bandit decisions.

Experience-based learning: Human learning describes how human individuals acquire and possess knowledge or skills under cognitive and environmental influences, taking into account prior experience (Illeris, 2002; Jarvis, 2006; Lefrancois, 2019).

For the new franchisee problem, Darr et al., 1995 studied the transfer of knowledge acquired through learning by doing empirically—they found evidence of learning based on weekly data collected from 36 pizza stores. To our best knowledge, there has not been analytical modeling work that studies both experience-based learning and queueing dynamics.