Empirical Modeling and Applications in Financial Economics and Healthcare Management

Yiwen Shen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

ABSTRACT

Empirical Modeling and Applications in Financial Economics and Healthcare Management

Yiwen Shen

With increased availability of data in various fields, researchers often need to combine efficient empirical methods with innovative analytical modeling techniques to make data-driven decisions and gain managerial insights from the large-scale raw data. In light of this, my thesis combines empirical methods and analytical modeling to study several data-related problems in the fields of financial economics and healthcare management. The first two parts of the thesis focus on two topics in financial economics: the role of dynamic information in asset pricing and the link between index-based investment and intraday stock dynamics. The last two parts of the thesis study the ICU admission decisions and cardiac surgery scheduling using data from different hospital units.

The first part of the thesis focuses on the role of information in financial market. As a fundamental topic in asset pricing, information is known to play an important role in determining asset prices and market volatility. In most of the existing literature, the information environment, i.e., the amount of knowable information, is assumed to be fixed and independent of investor's choice. However, in a dynamic market, the level of available information can vary substantially due to changes in technology and regulations. On the other hand, rational news producers may respond to investors' demand for information. Such effects are commonly seen in the reality, but are less studied in the literature. To bridge this gap, we develop a model of investor information choices and asset prices where the availability of information about fundamentals is time-varying. A competitive research sector produces more information when more investors are willing to pay for that research. This feedback, from investor willingness to pay for information to more information production, generates two regimes in equilibrium, one having high prices and low volatility, the other the opposite. Information dynamics move the market between regimes, creating large price drops even with no change in fundamentals. In our calibration, the model suggests an important role for information dynamics in financial crises.

In the second part of this thesis, we investigate how the growth of index-based investing impacts the intraday stock dynamics using a large high-frequency dataset, which consists of 1-second level trade data for all S&P 500 constituents from 2004 to 2018 (500GB). We estimate intraday trading volume, volatility, correlation, and beta using estimators that are statistically efficient under market microstructure noise and observation asynchronicity. We find the intraday patterns indeed change substantially over time. For example, in the recent decade, the trading volume and correlation significantly increase at the end of trading session; the betas of different stocks start dispersed in the morning, but generally move towards one during the day. Besides, the daily dispersion in trading volume is high at the market open and low near the market close. These intraday patterns demonstrate the implication of the growth of index-based strategies and the active-open, passive-close intraday trading profile. We theoretically support our interpretation via a market impact model with time-varying liquidity provision from both single-stock and index-fund investors.

In the third part of the thesis, we study the intensive care units (ICUs) admission decisions in a large hospital system. In the case of ICUs, which provide the highest level of care for the most severe patients, it is known that admission rates of some patients decrease as occupancy increases. It is also known that, for at least some conditions, ICU admission is not just a function of patients' illness, and that a significant proportion of the variation in ICU admission rates is due to hospital, not patient, factors. To understand such variation, we employ two years of data from patients admitted to 21 Kaiser Permanente Northern California ICUs from the ED. We quantify the variation in ICU admission from the ED under varying degrees of ICU and ED occupancy. We find that substantial heterogeneity in admission rates is present, and that it cannot be explained either by patient factors or occupancy levels alone. We use a structural model to understand the extent that intertemporal externalities could account for some of this variation. Using counterfactual simulations, we find that, if hospitals had more information regarding their behaviors, and if it were possible to alter hospital admission processes to incorporate such information, hospitals could reduce their ICU congestion in a safe way.

The last part of the thesis focuses on the impact of system workload on service time and quality in the context of cardiac surgeries. Using a detailed data set of more than

5,600 cardiac surgeries in a large hospital, we quantify how surgeon's daily workload level (e.g., number of surgeries) affects surgery duration and patient outcomes. To handle the endogeneity of surgeon's daily workload, we construct instrument variables using hospital operational factors, including the block schedule of surgeons. We find high daily workload of surgeons is associated with longer incision times and worse patient outcomes. Specifically, increased daily workload of surgeons leads to longer post-surgery length-of-stay in ICU and hospital, as well as higher likelihoods of reoperation and readmission for their patients. These results highlight the potential negative impact of surgeon's fatigue under long working hours. We then develop a surgery scheduling model that incorporates the effects of surgeon's daily workload levels.

Contents

List of Figures

List of Tables

Acknowledgments

My PhD study at Decision, Risk, and Operations Division, Columbia Business School has been an unforgettable journey. I feel fortunate and grateful to study and work in a friendly, inclusive, and encouraging environment like here. During the five years of my PhD study, I had the opportunity to learn how to become a qualified researcher and teacher from my excellent advisors. Their academic achievements, dedication to research, and passion for teaching greatly motivate me. I am also thankful to my colleagues, who are always supportive and ready to help when I experience hard times.

I would like to express my sincerest thanks to my advisors Prof. Paul Glasserman, Prof. Carri Chan, Prof. Fanyin Zheng, and Prof. Harry Mamaysky, for advising me through my PhD study. My discussions with them every one or two weeks helped me enter the exciting world of research and academia. Prof. Glasserman is the role model in my academia pursuit. He is innovative, elegant, and passionate in research. With great breadth and depth of knowledge, he maintains high curiosity to new ideas and keeps exploring the latest advances in the field. I also learnt a lot from his capacity to translate complex math to insightful intuitions. Prof. Chan and Prof. Zheng led me into the field of healthcare management at the third year of my PhD study. In this new field, I was excited by the numerous interesting research questions and the huge potential to deliver practical impact to people that need help the most. Through our collaborations on multiple projects, I developed my first understanding of the challenges and opportunities in different healthcare systems. In addition, I learnt from them the dedication and perseverance needed for young researchers, as well as the courage and determination when facing obstacles. I am especially grateful to their generous help in my job seeking process. Prof. Mamaysky has rich experience in both academia and the finance industry. He taught me how to interpret the modeling results in a way that is economically insightful and practically important. He is also a nice friend and

we are happy to share interesting stories in our daily lives.

I am thankful to my other coauthors, including Prof. Olivier Scaillet at Swiss Finance Institute, Prof. Chenxu Li at Peking University, and Meiqi Shi currently working at Morgan Stanley. I am lucky to have the opportunities to collaborate with them. They offered me great help in our research projects and always inspired the best of me. Meiqi Shi has the magic to solve all my computing-related problems, ranging from building the powerful server I need to implementing complex estimators on a large dataset in the most efficient way. This is a perfect complement to my capacities. I want to especially thank Prof. Li for his guidance and encouragement since my junior year at Peking University, which motivated me to apply to the PhD program at DRO and pursue my academia career. Besides, I want to thank Prof. Agostino Capponi for being on my defense committee and his valuable suggestions on my thesis. I am also grateful to Prof. Yash Kanoria, Prof. Jing Dong, Prof. Awi Federgruen, and Prof. Mark Broadie for their helpful comments on my research projects and seminars.

During my PhD study, I feel fortunate to work with a group of most talented colleagues in a collaborative and supportive environment. I benefited a lot from my discussions with senior PhD students including Pengyu Qian, Seungki Min, Pu He, and Zhe Liu. In the first year of my PhD study, I received great support from my same year colleagues Sharon Huang and Muye Wang. I am thankful to them and wish them all the best as they start their careers. I also want to thank Clara Magram and Winnie Leung at DRO, as well as Elam Elizabeth and Dan Spacher at PhD office for their patient assistance during my PhD.

Most importantly, I owe my deepest thanks to my families for their solid support. They have made great, heroic, and painful sacrifice for my growth and academia career. They are the best families a person could have. Without their support, it would be impossible for me to go this far and have this unbelievable journey from an ignorant child to a PhD from one of the best institutions in the world. I wish I would make them proud by becoming a man with integrity, diligence, and tenacity. Finally, I would thank my cat Kasli for carrying me through so many difficult battles with her unique and surprising abilities.

Dedication

To my parents Junhua Shen and Dongmei Lin

Chapter 1

Dynamic Information Regimes in Financial Markets

1.1 Introduction

Most research linking investor information acquisition and asset prices assumes a constant information environment. But why should the level of potentially available information remain constant in a market that is perpetually in flux? Changes in technology and regulation can generate persistent shocks to what an investor can learn about company fundamentals; and changes in what can be learned should influence investors as they decide whether to acquire costly information. Pushing this idea a step further, we investigate what happens when the information environment itself changes in response to investor demand for information. In other words, we posit that the news media, financial intermediaries, company executives, regulators, and prominent investors are not simply passive streams of information: the level of information they provide depends on investor demand. We then find that asset prices can change dramatically in response to changes in the supply and demand for information.

To capture these ideas, we develop a dynamic model of information and asset prices in which the level of available information changes in response to exogenous shocks and endogenous investor demand. In our model, an information shock changes the precision of information about fundamentals. An information shock is neither good news nor bad news – it is simply a change in the amount of knowable information. It is possible that information shocks occur in the absence of any shocks to fundamentals. Surprisingly we find that even such pure information shocks can have very large price impacts. This is a novel result and it has important ramifications. When increased information production accompanies a negative shock to fundamentals, our model suggests that the information shock greatly exacerbates the shock to fundamentals.

Examples of pure information shocks that are independent of fundamentals include regulatory changes (e.g., the Sarbanes-Oxley Act or Regulation Fair Disclosure), changes in accounting standards, or voluntary disclosure decisions by firms or governments. Such exogenous shocks trigger an endogenous response in the number of investors who choose to become informed. As more investors become informed, more information about fundamentals becomes available. This happens because a competitive information production sector, with a zero marginal cost of transmitting information once it has been discovered, will produce more information when more investors are willing to pay for it. This mechanism magnifies the asymmetry between informed and uninformed investors, it tends to increase price volatility, and it can amplify small information shocks into large price drops. In the model, such price drops result from an endogenous transition of the economy from a low- to a high-information regime.

Information shocks – that is changes in the amount of information being produced in the economy – often coincide with shocks to fundamentals. For example, in 2009 the Greek government revised its estimated budget deficit. This revision triggered a large increase in investor demand for information about Greek debt, as reflected for example in media attention and internet searches. Figure 1.1 shows the large and persistent increase in the number of Bloomberg articles mentioning Greece starting in 2009, and it provides at least circumstantial evidence that greater demand for information was met with greater supply. More information followed in the form of further revisions to official statistics, revelations about falsified data, stories of investment banks complicit in masking true conditions, research reports by industry analysts and non-governmental organizations, and a downgrade to junk by Standard & Poor's followed by Moody's. In turn, this increase in information invited further investor scrutiny, which just begot further information production.

Contemporaneous with these events, the price of Greece's debt dropped sharply as the volatility of its sovereign credit default swap spreads rose. In our model, the feedback between the demand and supply of information can lead to large price drops and increased volatility. Such price drops can occur even without a change in fundamentals; but they are amplified when an increase in information precision and a decline in fundamentals occur together. The model suggests that had the Greece shock not captured quite as much in-

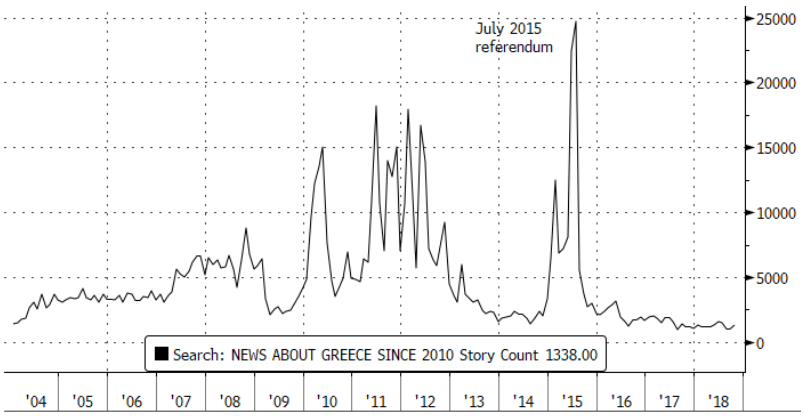


Figure 1.1: Number of Bloomberg articles per month that mention Greece.

vestor and media attention, the associated financial crisis would have been much smaller in magnitude and of considerably shorter duration.

Two more examples of increased information production that accompanied fundamentals shocks are worth mentioning. In June 2007, Bear Stearns disclosed that two of its hedge funds were on the brink of failure, fueling investor demand for information about the type of subprime mortgages in which the funds had invested. Indeed, Gorton and Ordonez (2014) and Dang et al. (2012, 2020) have argued that the demand for information about "safe" collateral triggered the ensuing crisis. In our narrative, as more investors chose to incur the cost of becoming informed, more information became available — through revised credit ratings, academic and industry research, media scrutiny and regulatory reports — as research producers, including rating agencies, news media, and sell-side research shops, began to produce more information as more investors began to demand it. The less informed investors, fearing an informational disadvantage, fled to safer assets.

Mamaysky (2020) argues that a portion of the volatility and price drops observed during the early phase of the COVID-19 crisis is attributable to exactly this dynamic. In the language of our model, the large fundamental shock of the coronavirus pandemic also triggered a positive information shock in the form of increased information production, which in turn caused the economy to temporarily transition to a high-information, low-price regime. In our model, information dynamics on their own can produce crisis-like effects, with low prices and high risk premia. Furthermore, fundamentals shocks that are accompanied by information shocks lead to considerably more severe market disruptions than do fundamentals

shocks in isolation.

Our model combines exogenous shocks to the quality of available information, an endogenous response by investors who may choose to become informed at a cost, and feedback from investor information choices to information producers, and ultimately to the amount of information that is produced. Most of our analysis uses a reduced-form representation of the feedback mechanism, but an appendix provides a microfoundation for the mechanism through a competitive information production sector that supplies investor demand for information. Information about fundamentals falls in three categories: publicly known, privately knowable at a cost, and completely unknowable.[1] In the interest of clarity, we only treat the case in which the fraction of knowable information varies, while the portion of knowable information that is publicly known is fixed.

In more detail, we develop an overlapping generations (OLG) model with a single risky asset, which pays a dividend each period, and a riskless asset. In each period, a new generation of investors observes the information environment, i.e. the current precision of the signal about the end-of-period dividend, decides whether to become informed at a cost, sets optimal demands, and trades to clear an exogenous net supply of shares. Market clearing determines the price. At the end of the period, these investors receive their dividend and sell their shares at the new price. The notion of "generations" should not be taken literally in our setting; the OLG framework simply provides a tractable dynamic setting to model changes in information, and it ensures that investors care about future prices as well as the next dividend.

Crucially, in making their information choices at the start of the period, investors take into account the distribution of exogenous shocks to information precision and the feedback from information choices in the current period to future precision. The future precision will affect the end-of-period asset price and thus investors' capital gains. Incorporating such time variation in information precision into a rational expectations setting is a technical challenge, and we develop a new solution methodology to address it.

[1] Publicly known information includes a product release that is covered in the New York Times. Privately knowable but costly information includes the performance of a firm's supplier network, which can be analyzed with painstaking analysis of public information. And information that is unknowable includes the outcome of a future medical trial relative to expectations.

Using this framework, we show that information shocks can lead to large and persistent drops in prices and increases in volatility. This is the main contribution of this study. We show that information shocks alone can produce prolonged periods of depressed prices and elevated volatility; we know of no other model in the literature that exhibits this behavior. Why does greater information precision have these adverse effects? Most of the paper is devoted to explaining this pattern, but a key part of the answer is time-varying information asymmetry: greater information precision for informed investors puts the uninformed at a greater disadvantage; persistence in the information state amplifies this effect. Within our framework, feedback from the demand for information to the amount of information available is essential to producing this behavior. When we shut off the feedback, the effects of information shocks become much more transient.

In an extension of our model, we allow for information shocks that are correlated with shocks to dividends, as in the historical scenarios discussed above. The effects of greater information precision are amplified when they are accompanied by negative shocks to dividends. We believe information shocks are typical of crises — demand for information about government deficits, subprime mortgages, or the spread of a virus grows with concerns about negative effects on fundamentals. Fairly minor negative dividend shocks, when accompanied by positive information shocks, can result in very large market disruptions — much larger than what would have occurred without an increase in information production. This suggests that while some market crises are precipitated by particularly large fundamentals shocks, others can result from relatively minor shocks to fundamentals that are associated with large increases in information production. More generally, if positive information shocks coincide with adverse shocks to fundamentals, then information production may exhibit countercyclical behavior and be a first order contributor to the market fluctuations that are observed across the business cycle.

To examine the magnitude of price effects arising from information shocks, we calibrate our model to stock market data. The equilibrium dynamics of the calibrated model fluctuate between two regimes, one with low volatility and high prices, and one with high volatility and low prices. The model can spend long intervals in each regime. A transition from one to the other can be sudden and result in a price move of over 10%, with no change in

fundamentals. The two regimes emerge from investor information choices; we do not impose them in setting up the model. When information shocks co-occur with fundamentals shocks, the former reinforce the latter, leading to much larger and longer lasting price moves. These information effects are present even though our investors are fully rational: they understand that the economy can transition from one regime to the other.

1.1.1 Contribution to the Literature

The interplay between information and asset prices is often studied through single-period models of the type in Grossman and Stiglitz (1980), Hellwig (1980), Admati (1985), and a large subsequent literature. But there are four important features available in a dynamic model that are inaccessible in a single-period model, and these merit discussion. First, persistent exogenous information shocks are necessary, though it turns out not sufficient, to generate endogenous low- and high-information regimes. For that, the feedback generated by our microfoundation is also needed; more investors have to induce more information production. Third, in a single-period model, exogenous shocks are often approximated by changes in model parameters, but such changes are necessarily outside the model and, in particular, not contemplated by the agents in the model. In contrast, our agents' beliefs take into account that the economy can transition between different information regimes; such transitions are therefore a feature of the model itself.

Finally, a dynamic model captures two distinct aspects of an increase in available information: greater information reduces uncertainty about the next dividend but can increase volatility in future prices and thus in capital gains. The first of these effects is clear — the information we model is information about dividends. To appreciate the second effect, note that in the absence of dividend information, price volatility is driven entirely by supply volatility; but when some investors have dividend information, this information is partly reflected in the price, so a persistent increase in signal precision leads to persistent price volatility. In a single-period model, the price merely determines the cost of a claim to an end-of-period dividend. With overlapping generations, investors earn the change in price over the period as well as a dividend, so the variance in this return affects their investment decisions at the beginning of the period. The two information effects, on dividends and on

end-of-period prices, are potentially offsetting and lead to more complex tradeoffs than can be captured in a single-period setting.[2] We will see that this dual role of information in dynamic models can lead to starkly different conclusions than those of static models.

To the best of our knowledge, our model is the first to capture a stochastic information environment, endogenous investor information choices, and feedback from these choices to available information. Spiegel (1998) and Dutta and Nezlobin (2017) develop overlapping generations models in which all investors have the same information. Watanabe (2008) extends Spiegel (1998) model by introducing asymmetric information. Biais et al. (2010) also model asymmetric information in an OLG setting. In their model, as in Watanabe (2008), the fraction of informed investors and the precision of their signals are fixed and exogenous. Wang (1993) develops a continuous-time model of trading among differentially informed investors with a fixed fraction of informed investors and a fixed information environment; Wang (1994) is a discrete-time version of the model that investigates trading volume. In Avdis (2016), the fraction informed is endogenous but does not affect the information environment. The model of Veldkamp (2006) includes a dynamic information market, but its investors are indifferent to end-of-period prices, leading to starkly different implications than our model. The OLG model of Farboodi and Veldkamp (2020) incorporates a changing information environment, but the change is limited to a deterministic increase in investor information processing capacity over time. Signal precision also changes deterministically over time in Brennan and Cao (1997).

Information revelation is at the center of the crisis explanation of Gorton and Ordonez (2014). In their account, a crisis results when lenders choose to acquire information about borrowers' collateral; with less information available, borrowers with poor collateral have access to credit, and the increased supply of credit sustains higher growth. We work in an entirely different framework, but one contrast is particularly noteworthy. In Gorton and Ordonez (2014), the information revealed is bad news; following an aggregate shock, some

[2]This dual role of information is also highlighted in the multiperiod models of Avdis (2016) and Dutta and Nezlobin (2017), but those models do not include feedback effects. In Avdis (2016), serial correlation in asset supply allows investors who acquire information about the current dividend to make inferences about future discount rates. In Dutta and Nezlobin (2017), the tradeoff between current information and future volatility is examined through a firm's growth rate. These features are very different from the considerations that drive our model.

unobservable amount of collateral becomes bad, thus inducing more information acquisition. In our setting, it suffices for the precision of information to change — a shock may bring more news or less news without being specifically good or bad news. An increase in precision leads to a price drop when it magnifies the information asymmetry between informed and uninformed investors, leading the uninformed to reduce their demand for the risky asset. As noted previously, negative correlation between information precision and fundamentals amplifies the price drop.

We present our model in Section 1.2. Section 1.3 outlines the model solution and our main theoretical results. Section 1.4 studies changes in the level of knowable information and shows that feedback can lead to two information regimes, using parameters calibrated to market data. To isolate the effect of information dynamics, and to emphasize the potential effect of information alone, we focus most of the analysis in this section on the case of information shocks which are uncorrelated with dividends. In Section 1.4.4 we extend the model to handle the case of correlated information and dividend shocks. Section 1.5 explores the mechanisms leading to large price changes across regimes and considers information asymmetry, the cost of information production, and strategic complementarity in information acquisition. Appendix A.1 microfounds our feedback mechanism, and subsequent appendices provide proofs of our theoretical results. A Supplementary Appendix provides some additional proofs and details of our calibration and numerical calculations.

1.2 Model

1.2.1 Dividends and Timing

A single infinitely-lived security pays a dividend in each period. The dividend paid at the end of period t is given by

$$D_{t+1} = \bar{D} + \rho(D_t - \bar{D}) + M_{t+1} = \underbrace{(1-\rho)\bar{D}}_{\mu_D} + \rho D_t + M_{t+1}. \tag{1.1}$$

The innovation M_{t+1} decomposes as

$$M_{t+1} = \underbrace{m_t + \theta_t}_{\tilde{m}_t} + \epsilon_{t+1}, \tag{1.2}$$

with the following interpretation: m_t is known to informed investors; θ_t is public information; $\tilde{m}_t$ is the knowable portion of the innovation; and ϵ_{t+1} is unknowable at the beginning of period t. These are mean zero, normally distributed random variables, independent across time,[3] with variances given by

$$\mathsf{var}(\tilde{m}_t) = f_t \mathsf{var}(M) \qquad \text{and} \qquad \mathsf{var}(\epsilon_{t+1}) = (1 - f_t)\mathsf{var}(M). \tag{1.3}$$

and

$$\mathsf{var}(m_t) = \phi \mathsf{var}(\tilde{m}_t) \qquad \text{and} \qquad \mathsf{var}(\theta_t) = (1 - \phi)\mathsf{var}(\tilde{m}_t). \tag{1.4}$$

Thus,

$$\begin{aligned} f_t &= \text{fraction of dividend innovation that is knowable;} \\ 1 - \phi &= \text{fraction of knowable part of dividend innovation that is public.} \end{aligned}$$

The parameter ϕ will control the degree of asymmetry between informed and uninformed investors. Higher ϕ corresponds to higher asymmetry.

The economy contains overlapping generations of agents. The new generation is in the market for two periods, t and $t + 1$. Before making investment decisions in period t, all agents observe f_t, θ_t, D_t, (and ϕ), and the time-t informed agents observe m_t.[4] A fraction $\lambda_t \in [0, 1]$ of agents are informed at time t. Becoming informed entails paying a fixed cost c_I; a fixed portion of this cost, c_M, goes to pay a news production sector to discover new information. Under our model parameterization, informed agents find it optimal to pay both c_I and c_M. The time-t uninformed agents, representing $1 - \lambda_t$ of the population, in addition to observing θ_t and D_t, also observe the market clearing price P_t. Since the market-clearing price contains information about m_t through the demands of the informed traders, the uninformed also make rational inferences from the price about the innovation m_t. The price is not fully revealing about m_t because of the presence of unobservable supply shocks. In this respect, for a given f_t and ϕ, our information environment is the same as in Grossman and Stiglitz (1980). After observing all available (public or private) information, investors set their demands as functions of the price, which determines the price through

[3] More precisely, they are conditionally independent given all f_t.

[4] We have solved the model with time-varying ϕ_t but, for clarity, we assume it is constant in this chapter.

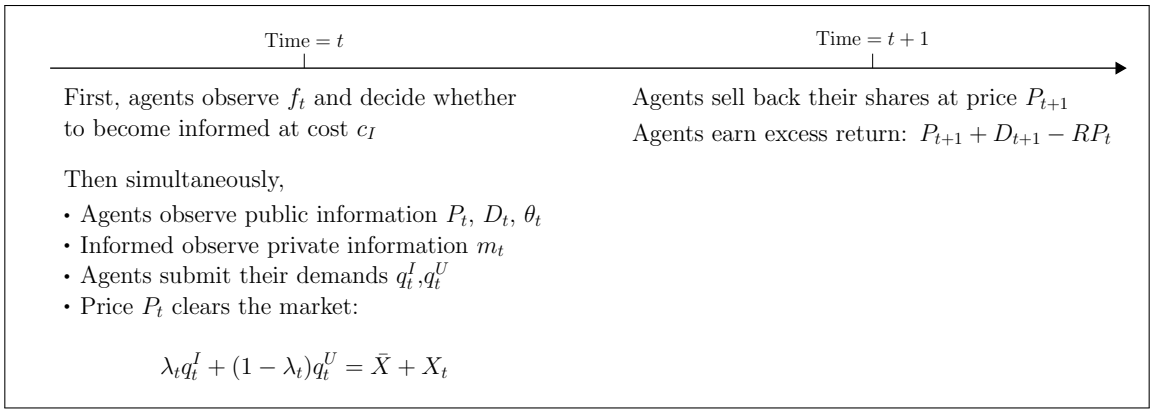


Figure 1.2: Sequence of events in each period.

market clearing. At time $t+1$, investors receive the dividend, sell their shares at the time $t+1$ price, and the process repeats. Figure 1.2 summarizes the timing of the model; the agent demands q and asset supply X are discussed in Section 1.2.4.

1.2.2 Information Environment

The innovation of this study is to allow the information environment, as represented by f_t, to evolve over time in response to exogenous shocks and in response to information decisions made by past generations of investors. We show below that a straightforward microfoundation leads to simple dynamics of information precision: f_t follows an AR(1) process combined with a feedback effect from today's fraction informed λ_t to tomorrow's precision, or

$$f_{t+1} = a_f + \kappa_f (f_t - a_f) + b_f \lambda_t + \epsilon_{f,t+1}. \tag{1.5}$$

for constants a_f, κ_f and b_f, as well as a noise term $\epsilon_{f,t+1}$. We assume (for now) that the information shocks $\epsilon_{f,t+1}$ and fundamental shocks M_{t+1} are independent. This allows us to cleanly separate the effect of changing information precision from the effect of changing fundamentals; in Section 1.4.4, we introduce correlation in the shocks. To be consistent with the interpretation of f_t as a measure of signal precision in (1.3), we need to restrict f_t to values between 0 and 1. We therefore apply a mapping $\Pi_{\mathcal{D}}$ to the right side of this equation, where $\Pi_{\mathcal{D}}$ maps the real line to a set $\mathcal{D} \subseteq [0,1]$.[5] We thus arrive at our model of

[5] In the simplest case, $\Pi_{\mathcal{D}}(x) = \min(1, \max(0, x))$ projects x to $[0,1]$. For some of our theoretical results in Section 1.3 and for our numerical results, we will discretize f_t to finite subsets of the unit interval, but

the information environment:

$$f_{t+1} = \Pi_{\mathcal{D}} \left(a_f + b_f \lambda_t + \kappa_f \left(f_t - a_f \right) + \epsilon_{f,t+1} \right), \tag{1.6}$$

This specification provides the simplest model that captures persistent, stochastic time variation in the information environment and, most importantly, feedback from the fraction informed λ_t to the available information.

To generate the f_t dynamics in (1.5) and (1.6), we assume that the dividend innovation M_{t+1} consists of a large number of i.i.d. pieces of information. This information can be about local economic conditions that affect a firm's profitability or the economy's output, technological innovation across different product lines, consumer demand, managerial talent, competitor performance, relevant industry and macro trends, and so on. Each unit of information can be in one of two states: observable or unobservable. The state of being observable or unobservable is persistent. For example, informed investors may push a company or government to disclose a certain piece of information, and once the company or government agrees, it is likely to continue to disclose this information, thus making it observable. However, at some point the disclosure policy may change, and previously disclosed information may become undisclosed, and thus unobservable. Observability does not depend only on disclosure. For example, technological innovation may make certain characteristics of an oil well observable, even if they were unobservable in the past. Similarly, a company may build a canopy over its distribution facility rendering satellite imagery no longer informative. In both of these examples, the change in observability is persistent. We assume any observable unit of information has a ϕ probability of being only privately observable and a $1 - \phi$ probability of being publicly observable.

A profit maximizing, competitive information production sector can discover, at a per unit cost c_P, previously unobservable units of information, and then reveal these to its clients. Once a unit of information is discovered, the marginal cost of revealing it to investors is zero. Furthermore, we assume that discovered units become observable. We refer to this sector as the news producers, though in addition to financial journalists, it can also contain stand-alone, sell-side, or buy-side research firms, ratings agencies, or bloggers on outlets

for now we keep the discussion general.

like Seeking Alpha who are compensated for the number of views their posts receive. In our model calibration, informed investors optimally choose to spend a portion c_M of their cost of becoming informed c_I to purchase the information produced by the news sector. Investors who purchase information from the news producers are legally obligated not to share information they obtain from news outlets with one another; thus the only way to obtain information is to buy it from the producers. The zero marginal cost condition and the legal obligation not to disclose purchased information mirror the assumptions of Veldkamp (2006), and provide a strong incentive to produce more information when more investors demand it.

As we show in Appendix A.1, these assumptions yield the f_t process in (1.5). The AR coefficient κ_f, which determines the degree of persistence of the information state, equals one minus the sum of two probabilities: (1) the probability that a given unit of information transitions from unobservable to observable, and (2) the probability that a unit transitions from observable to unobservable. The coefficient a_f, which would be the steady-state level of information precision in the absence of the feedback effect, $b_f \lambda_t$, equals the probability of transitioning from unobservable to observable, conditional on a transition taking place. Finally, we show in the appendix that $b_f = c_M / c_P$, and therefore the feedback effect $b_f \lambda_t$ reflects the incentive of news producers to discover more information: b_f is positive and increasing in the amount spent by investors to buy news, c_M; it is decreasing in the cost of producing a new unit of information, c_P; and the overall effect is increasing in the number of informed λ_t.

1.2.3 Illustration of Dynamics

To illustrate the dynamics of the information state f_t and the implications for the equilibrium asset price P_t, we first consider an example. As detailed in subsequent sections, our dynamic equilibrium includes market clearing and utility maximizing decisions by agents in setting their demands and deciding whether to become informed, taking into account feedback from the fraction informed to the information available. An example should help explain where we are headed. Figure 1.3 plots an equilibrium path in our economy, subject to a particular set of information and dividend shocks. For this example, we assume

information and dividend shocks are uncorrelated, and supply shocks (discussed in the next section) are turned off. Each of the three series in the figure is subject to the same set of shocks, but represents the equilbrium path across three different sets of parameter values.

The top panel of Figure 1.3 shows the evolution of f_t in (1.6) in blue with baseline parameters $a_f = 0.175$, $b_f = 0.384$, and $\kappa_f = 0.91$ from the calibration in Appendix A.2. The other model parameters are detailed in Table A.1. The red, dash-dotted line turns off the feedback by setting $b_f = 0$. Early in the plotted history, the information state experiences several consecutive positive shocks. Without feedback, it quickly mean-reverts toward $a_f = 0.175$. With feedback, that is with an increase in λ_{t+1} due to a positive information shock $\epsilon_{f,t+1}$ feeding into a higher value of f_{t+2}, f_t remains elevated much longer. (We discuss the behavior of λ_t as a function of f_t in Section 1.4.1.) The bottom panel shows the consequences for the equilibrium price P_t. During the protracted period of elevated f_t, the model with feedback produces substantially lower prices than the model without feedback. (Dividends are identical in the two cases.) In the calibration we interpret each period as a month, so the lower figure shows roughly a 5-year period of depressed prices resulting entirely from information dynamics. Prices do not revert back to the high-price regime until the information state experiences several consecutive negative shocks.

For comparison, the figures show a dashed black line corresponding to no feedback but greater persistence, with $\kappa_f = 0.98$. As expected, greater persistence slows the mean-reversion in f_t. But the price in this case is nearly identical to the case $b_f = 0$, $\kappa_f = 0.91$. In other words, the effect of feedback is qualitatively different from ordinary persistence. Indeed, we will see that in the model with feedback f_t is drawn toward a level of 0.88 as well as to the point $a_f = 0.175$. Feedback endogenously introduces two regimes associated with high and low levels of f_t. That several f_t shocks need to occur in rapid succession in order to induce transitions from one regime to the other makes each regime in the feedback model highly persistent. A transition from the low information (i.e. low f_t) regime to the high one is accompanied by a large drop in price and, we will see later, an increase in volatility.

The vast majority of the asymmetric information literature (see the discussion in Section 1.1.1) assumes a constant f_t. In this case, the three equilibrium paths in Figure 1.3 would be identical, since the only difference between the paths is in the behavior of f_t. Even if one

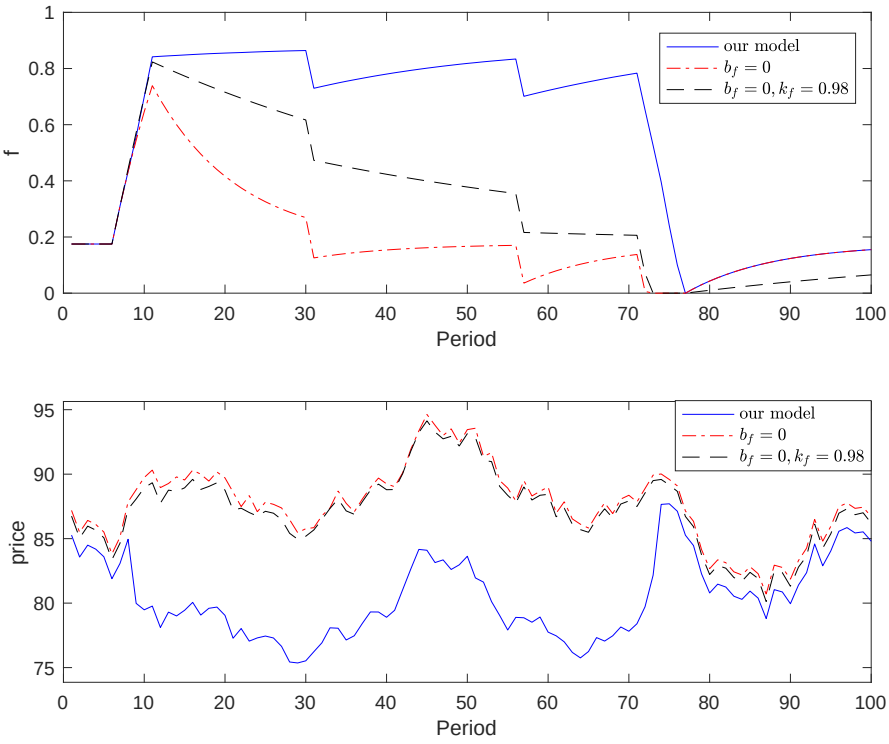


Figure 1.3: Simulated paths under different model settings

Top: Information state f_t. Bottom: Price P_t. Feedback in f_t creates a prolonged period of depressed prices.

extends the standard models in the literature to the case of a stochastic and highly persistent f_t, but with no feedback, the difference in equilibrium price paths would be only of second order (as is the difference between the dash-dotted red and dashed black lines in the bottom panel of the figure). The introduction of feedback, from the number of informed to the information state, leads to a first order difference in the behavior of prices (and volatilities as we discuss later). And this large difference in prices occurs even in the absence of any differences in fundamentals, as the dividend and supply shocks across the three equilibrium paths are identical.

1.2.4 Investor Optimization Problem

We return to the model formulation. At the beginning of period t, a unit mass of new (young) investors enter the market, each endowed with wealth W_t, known at time t. For an investor who buys q shares of the risky asset at price P_t at the beginning of the period and

sells the shares at the end of the period at price P_{t+1}, terminal wealth is given by

$$\begin{aligned} W_{t+1} &= R(W_t - qP_t) + q(D_{t+1} + P_{t+1}) \\ &= RW_t + q(D_{t+1} + P_{t+1} - RP_t), \end{aligned} \tag{1.7}$$

where $R > 1$ is the gross return on a riskless asset. It will be convenient to define the per period net profit from owning a single share of the stock as

$$\pi_{t+1} \equiv D_{t+1} + P_{t+1} - RP_t, \tag{1.8}$$

in which case the budget constraint becomes $W_{t+1} = RW_t + q\pi_{t+1}$. Agents who enter at time t consume their wealth at $t+1$ and leave the market. These agents set their demands for shares of the risky asset at time t by solving

$$J_t^\iota \equiv \max_q \mathsf{E}\left[\mathsf{E}[W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}] - \frac{\gamma}{2}\mathsf{var}(W_{t+1}|\mathcal{I}_t^\iota, f_{t+1})\middle|\mathcal{I}_t^\iota\right], \quad \iota \in \{I, U\}, \tag{1.9}$$

where $\mathcal{I}_t^U = \{f_t, \lambda_t, D_t, \theta_t, P_t, W_t\}$ is the uninformed agents' information set at time t, $\mathcal{I}_t^I = \mathcal{I}_t^U \cup \{m_t\}$ is the informed agents' information set, and $\gamma > 0$ is a risk aversion parameter. Similar objectives are used in Peress (2010), Van Nieuwerburgh and Veldkamp (2010), and Mondria (2010), and can be interpreted as expressing a preference for early resolution of uncertainty, in the sense of Kreps and Porteus (1978). Maximizing (1.9) is equivalent to maximizing

$$\mathsf{E}\left[v\left\{\mathsf{E}\left[-\exp(-\gamma W_{t+1})|\mathcal{I}_t^\iota, f_{t+1},\right]\right\}|\mathcal{I}_t^\iota\right],$$

with $v(u) = -\frac{1}{\gamma}\log(-u)$, if W_{t+1} is conditionally normal, as it will be in our equilibrium. We could allow investors to condition on past values of variables in their information sets in (1.9), but past information will be irrelevant, given our independence assumptions.

In addition to investor demands for shares of the risky asset, we need to specify the supply. As in the OLG model of Allen et al. (2006), we assume that X_t, the stochastic part of the supply of the risky asset, is independent and identically distributed from one period to the next. As explained in Allen et al. (2006), i.i.d. supply can be interpreted as the result of trading by price-insensitive noise traders who reverse their trades at the end of each period. New investors each period thus only clear a new exogenous supply shock.[6]

[6]Our model extends easily to allow persistent supply shocks, at the expense of adding an additional state variable, which complicates our numerical examples. See Avdis (2016) for a model in which supply persistence influences investors' decisions to become informed.

We assume each X_t is normally distributed with mean zero and variance σ_X^2. Furthermore, we assume that there exists a positive net supply $\bar{X}$ of the risky asset, and that this fixed supply is constant over time.

1.2.5 Equilibrium

Given a function $\lambda : [0,1] \mapsto [0,1]$, yielding the fraction informed $\lambda(f_t)$, a market equilibrium is defined by a price process P_t and demands q_t^I and q_t^U, depending on the price and other time-t information $\mathcal{I}_t^I$ and $\mathcal{I}_t^U$, that clear the market,

$$\lambda_t q_t^I + (1-\lambda_t) q_t^U = \bar{X} + X_t, \tag{1.10}$$

and for which q_t^ι solve (1.9), $\iota \in \{I, U\}$, for all t.

Market clearing and investor optimality define a market equilibrium, given a function λ that determines the fraction of investors who are informed. Next we define what it means for this fraction to be determined endogenously. As in our discussion of Figure 1.2, we suppose that investors at the beginning of the period can choose to become informed at a cost c_I, incurred at the beginning of the period but after observing the current information state f_t. Investors' decisions to become informed or remain uninformed thus define a mapping from the information state to the fraction informed, which is precisely λ. We will use the following:

Definition 1 (Endogenous fraction informed)**.** *Given the f_t dynamics in (1.6), we call λ the endogenous fraction informed if it satisfies the following conditions for each $f \in [0,1]$:*

(i) $\lambda(f) = 0$ *and* $\mathsf{E}[J_t^I - Rc_I | f_t = f] < \mathsf{E}[J_t^U | f_t = f]$*; or*

(ii) $0 \le \lambda(f) \le 1$ *and* $\mathsf{E}[J_t^I - Rc_I | f_t = f] = \mathsf{E}[J_t^U | f_t = f]$*; or*

(iii) $\lambda(f) = 1$ *and* $\mathsf{E}[J_t^I - Rc_I | f_t = f] > \mathsf{E}[J_t^U | f_t = f]$.

Note the expectations in Definition 1 are taken prior to the agents receipt of their signals. In case (ii), the fraction $\lambda(f)$ is the point at which the marginal investor is indifferent between becoming informed and remaining uninformed. Cases (i) and (iii) cover the possibility that one choice dominates the other and is therefore selected by all investors.

1.3 Model Solution

The main challenge in our model is to combine a time varying information environment with agents' rational expectations. In this section, we outline our model solution, leaving the statements and proofs of our main results to the appendix. We first take an arbitrary fixed fraction informed $\lambda(f)$, for each information state f, and find a market equilibrium consistent with that function $\lambda(\cdot)$. The role of $\lambda(\cdot)$ is to fully specify the f_t process in (1.5), though we do not yet require that $\lambda(\cdot)$ corresponds to optimizing behavior by agents in the model. Then, from a market equilibrium, we give conditions for the fraction informed $\lambda(\cdot)$ to be optimal, in the sense that optimizing behavior by agents yields exactly the number of informed that $\lambda(\cdot)$ specifies. We then combine the results to give conditions for an information equilibrium, in which conditions for a market equilibrium and an endogenous fraction informed are jointly satisfied. The end result is a rational expectations equilibrium (REE) in a model with a dynamic information environment; the solution for an REE in this setting is an important technical contribution of this study.

1.3.1 Market Equilibrium

Proceeding with the first of these statements, we show that, for any choice of λ, the model admits a market equilibrium in which the price process takes the form

$$P_t = a_t + b_t m_t + g\theta_t - c_t X_t + dD_t, \tag{1.11}$$

where g and d are constants, and a_t, b_t, c_t are functions of the information state f_t but do not otherwise depend on t.

To characterize investor demands, we need to find the utility of terminal wealth. If prices are given by (1.11), we can write terminal wealth W_{t+1} in (1.7) as

$$W_{t+1} = RW_t + q(1+d)D_{t+1} + q(P_{t+1} - dD_{t+1} - RP_t) \tag{1.12}$$

$$= RW_t + q\Big[(1+d)D_{t+1} + a_{t+1} + b_{t+1}m_{t+1} + g\theta_{t+1} - c_{t+1}X_{t+1} - RP_t\Big]. \tag{1.13}$$

Note that m_{t+1}, θ_{t+1} and X_{t+1} are independent of D_{t+1}, and of any time t information. With

a view to solving (1.9), we evaluate the conditional mean of terminal wealth as

$$\mathsf{E}[W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}] = q\left[(1+d)(\mu_D+\rho D_t+\theta_t+\mathsf{E}[m_t|\mathcal{I}_t^\iota])+a(f_{t+1})-RP_t\right]+RW_t, \quad \iota \in \{I, U\}. \tag{1.14}$$

In the above, we write a_t from (1.11) as $a(f)$ to make explicit its dependence on the state variable, and discuss it further below. For the conditional variance, we use (1.12)–(1.13) to write

$$\mathsf{var}(W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}) = q^2(1+d)^2\mathsf{var}[D_{t+1}|\mathcal{I}_t^\iota, f_{t+1}] + q^2\mathsf{var}[P_{t+1} - dD_{t+1}|\mathcal{I}_t^\iota, f_{t+1}]. \tag{1.15}$$

The problem with this expression is that the $\mathsf{var}[P_{t+1} - dD_{t+1}|\mathcal{I}_t^\iota, f_{t+1}]$ term depends on the coefficients of the price function in (1.11) and these are not yet determined. To overcome this difficulty, we introduce a *conjectured variance* function $V_B(f)$, which allows us to rewrite (1.15) as

$$\mathsf{var}(W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}) = q^2(1+d)^2\left[\mathsf{var}(m_t|\mathcal{I}_t^\iota) + (1-f_t)\sigma_M^2\right] + q^2V_B(f_{t+1}). \tag{1.16}$$

We are ultimately interested in a rational expectations equilibrium, which means that the conjectured variance function must be "correct." Using the dividend process in (1.1)–(1.3) and the price function in (1.11), the correctness condition requires that

$$V_B(f) = b(f)^2\phi f\sigma_M^2 + g^2(1-\phi)f\sigma_M^2 + c(f)^2\sigma_X^2 \qquad \forall f, \tag{1.17}$$

as can be seen by comparing the last term in (1.15) and (1.16). If (1.17) holds, then investors' conjectures about how the variance of $P_{t+1} - dD_{t+1}$ depends on f_{t+1} are consistent with the equilibrium price process. However, we initially allow investors to have an arbitrary, strictly positive variance conjecture V_B, which is shared by all investors. In other words, we do not initially assume that investors know the coefficient functions b and c.

With arbitrary V_B, we do not have equality in (1.16). That is the true conditional variance of wealth (on the left-hand side) is not equal to the conjectured variance of wealth (on the right-hand side). Instead, we posit that investors solve their optimization problems (1.9) as though (1.16) held. In other words, investors solve (1.9) but with the conditional variance replaced by the right side of (1.16). A market equilibrium with conjectured variance V_B is then a price process and investor demand functions that clear the market and solve (1.9) with this modification. Proposition 3 ensures the existence of such an equilibrium.

This market equilibrium is not in general a rational expectations equilibrium because the correctness condition in (1.17) does not hold. But we can think of agents in the model as learning over time. Starting from an initial conjecture, investors set their demands and clear the market at a price of the form in (1.11). They (or the next generation) then observe the realized variance given by the right side of (1.17). They update their expectations by setting V_B equal to this realized variance (which is fully specified for all f), and the process repeats. This in fact is how we solve our model numerically. A rational expectations market equilibrium is characterized as a fixed point of this iterative process, and Proposition 4 gives modest parameter restrictions under which such a fixed point exists. The fixed point determines $b(f)$ and $c(f)$, and these coefficients determine $a(f)$ through market clearing (as shown in Appendixes A.4.1 and A.4.2). Going forward, we use the term market equilibrium to refer to a rational expectations market equilibrium.

1.3.2 Information Equilibrium

Propositions 3 and 4 show the existence of a market equilibrium with an exogenous $\lambda(\cdot)$. We now need to show that our notion of the endogenous fraction informed in Definition 1 is meaningful. Given a variance conjecture $V_B(f)$ we need to find, for each f, a $\lambda(f)$ that makes investors indifferent between paying the cost c_I of becoming informed or staying uninformed; if no such λ exists, we set λ equal to zero or one according to Definition 1. Proposition 5 ensures the existence of an endogenous λ. The $V_B(\cdot)$ function for which we calculate an endogenous fraction informed will not, in general, be consistent with condition (1.17).

We have argued that, given λ, we can find a market equilibrium and in particular a correct conjectured variance V_B; and given a variance conjecture V_B, we can find an endogenous λ. The remaining step combines these results to arrive simultaneously at a market equilibrium and an endogenous λ. The precise statement of our combined result is in Proposition 6. For our numerical examples, we discretize the state space and approximate a $V_B(f)$ consistent with rational expectations and an endogenous fraction informed $\lambda(f)$ as follows:
This iterative procedure generates an approximate solution V_B to (1.17) and yields solutions to all the equilibrium quantities and price coefficients, except $a(f)$. The last step of the

discretize the state space $f \in \mathcal{D}$ via $\mathcal{D} = [0, 1/(n-1), 2/(n-1), \ldots, 1]$;
start with an conjectured belief $V_B^0 = (V_B^0(0), V_B^0(1/(n-1)), \ldots, V_B^0(1))$;
repeat
  set $V_B^1 \leftarrow V_B^0$;
  for each f, solve for optimal $\lambda^*(f)$, $b(f)$, and $c(f)$ given V_B^1;
  set V_B^0 equal to the right-hand side of (1.17) given $\lambda^*(f)$;
until $||V_B^1 - V_B^0|| < \varepsilon$;
solve for $a(f)$ as the fixed point to (A.18);

Algorithm 1: Solution algorithm for the model.

algorithm solves for the discretized $a(f)$ function. For more details see Section A.9 of the Supplementary Appendix.

1.4 Analysis of the Model

Our model is motivated by the idea that as more investors become informed, more information may become available. This type of feedback can arise at the onset of market stress in response to heightened investor attention. In this section, we will show that this dynamic can lead to periods of low and high volatility and high and low prices driven purely by changes in the information state, with no change in fundamentals, as illustrated in Figure 1.3. In other words, we can generate transitions similar to business cycles or even financial crises through changes in the level of information, without necessarily the release of negative information. In Section 1.4.4, we analyze the behavior of the model in the case of correlated information and fundamentals shocks.

1.4.1 Dynamics of Information Precision

To provide insight into the model, we develop the numerical example of Figure 1.3. A single period in our model is one month, and the model parameter values are given in Table A.1. The details of the calibration are given in Appendix A.2. The solid line in Figure 1.4 shows λ as a function of f. We calculate this curve by starting from a flat variance conjecture V_B and iteratively updating V_B and λ as discussed in Section 1.3. This iterative process converges very quickly in our numerical experiments.

At low levels of information precision f, the figure shows a flat section where $\lambda(f) = 0$; with little information available, no investor chooses to bear the cost of becoming informed. Once f increases to just above 0.4, we have a positive fraction of investors informed, and this fraction generally increases with the precision f.[7]

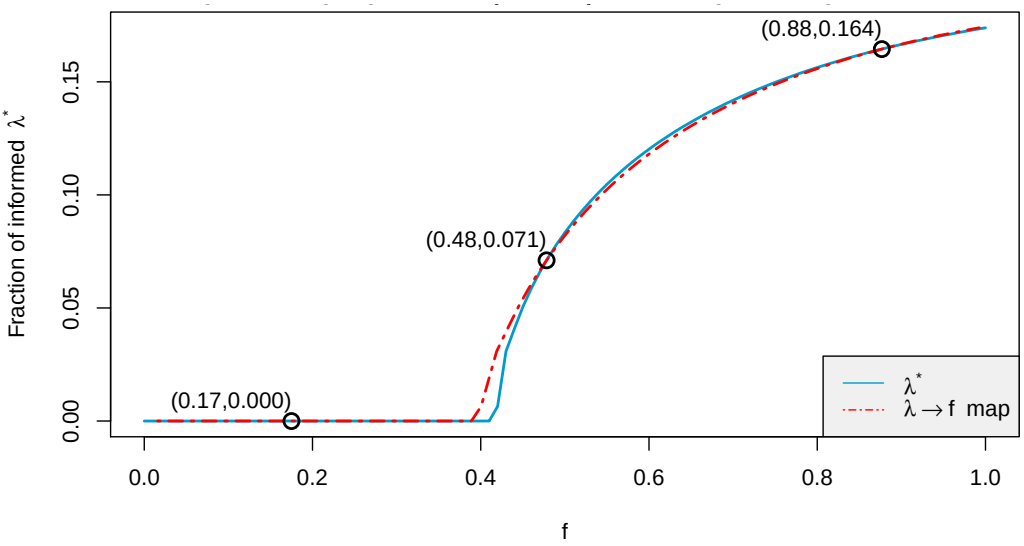


Figure 1.4: Endogenous fraction informed as a function of f

Note: The solid line shows the fraction informed $\lambda(f_t)$ in information state f_t, and the dashed line shows the mapping from λ to f_{t+1} without exogenous shocks. Each circle shows a point where $f_t = f_{t+1}$ when the shocks in (1.6) are zero, labeled with its (f, λ) value. The figure uses parameter values from Table A.1.

To interpret the dashed line in Figure 1.4, we shut off the exogenous shocks in the evolution of f_t by setting $\epsilon_{f,t+1} \equiv 0$ in (1.6). The dashed line then shows the mapping from λ to the next value of f. That is, starting from any $f_t = f$ on the horizontal axis, reading up to the solid line then across to the dashed line and back down to the horizontal axis yields f_{t+1}. Points where the two lines cross are fixed-point combinations of $(f, \lambda(f))$ in a model without exogenous shocks. In other words, the three circled points in the figure are cases where $f_{t+1} = f_t$ when $\epsilon_{f,t+1} = 0$.

Consider, for example, the circled point near $f = 0.48$, $\lambda(f) = 0.071$. Starting at that

[7]For some parameter values, at f near 1 we have a small decline in $\lambda(f)$. The possibility of a decline in $\lambda(f)$ as f increases reflects the dual roles of information in a multiperiod model. Becoming informed benefits an investor by reducing uncertainty about the end-of-period dividend. However, as more investors become informed, the variance of the end-of-period asset price increases, so the net effect on the variance of an investor's end-of-period wealth is indeterminate.

f, the endogenous fraction informed $\lambda(f)$ is precisely the value that keeps the information state at f under the evolution in (1.6) without endogenous shocks. The model still has feedback from λ to f (and f to λ), but f_t remains fixed. The same argument applies to the intersection near $f = 0.88$. In the lower left, the curves intersect throughout an interval where $\lambda(f) = 0$, and we have a fixed point at $(a_f, 0)$ because the dynamics in (1.6) drive f_t to a_f when $\lambda_t = \epsilon_{f,t+1} = 0$.

If we keep $\epsilon_{f,t} = 0$ and start the evolution of f_t near 0.88, it will move toward 0.88; and if we start the evolution near 0.175, f_t will move toward 0.175. In contrast, the point $f = 0.48$ is an unstable fixed point: starting to the left of this point will drive f_t to 0.175, and starting to the right will drive f_t to 0.88. When we reintroduce the shocks $\epsilon_{f,t}$, we therefore expect f_t to spend long periods near 0.175 and long periods near 0.88.

This behavior explains the pattern we saw in Figure 1.3. An initial set of positive shocks increase f_t. With feedback dynamics, f_t stays near 0.88 for a long time: once the fraction informed $\lambda(f_t)$ is high, the demand for information keeps f_t high. Eventually, exogenous negative shocks decrease f_t sufficiently that it moves toward 0.175. The effect of feedback is therefore to endogenously create two regimes (corresponding to the two stable fixed points in Figure 1.4). We have not yet explained why the high f_t regime is associated with low prices and, as we will see, with high volatility. That explanation will come in Section 1.4.3.

Figure 1.5 provides additional information on the stochastic dynamics of f_t. The left panel shows the steady-state distribution of f_t (indicated by the blue circles in the left panel of the figure), calculated using a Markov chain representation.[8] The distribution is bimodal, showing that the economy spends the majority of its time in the vicinity of the two stable fixed points from Figure 1.4, and confirming the presence of two regimes. If we fix λ at its mean value of 0.0731, which effectively turns off the feedback effect, the steady-state distribution (shown by red triangles) becomes unimodal — we no longer get two regimes.[9]

The right panel shows that the two regimes in the feedback model are highly persistent, in the sense that the cumulative probability of transitioning from one to the other remains low, even after many periods. The probability of transitioning within 240 periods is only

[8] See the Supplementary Appendix for details.

[9] The steady-state distribution of f_t in an economy when $b_f = 0$ is also unimodal, and centered at 0.175.

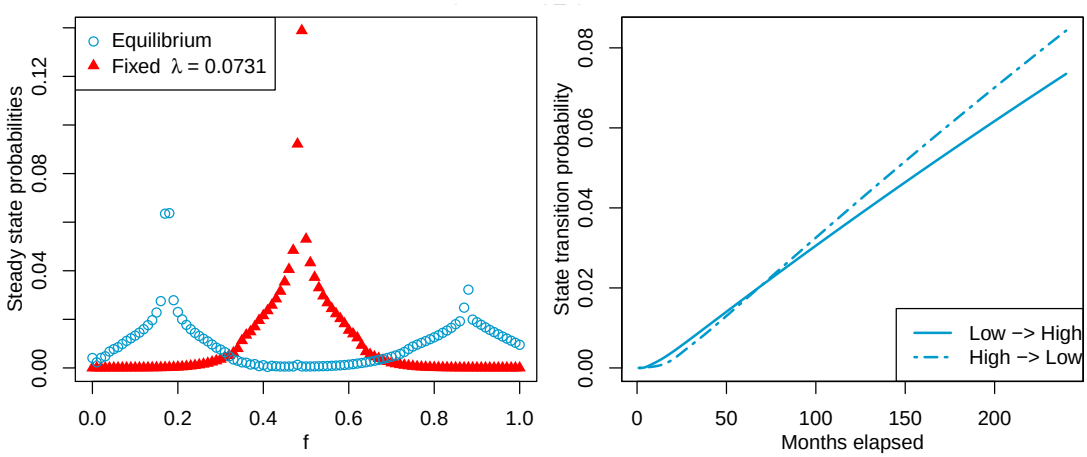


Figure 1.5: Steady-state distribution and transition probabilities of f_t

Note: The left panel shows the equilibrium steady state distribution of f_t. The right panel shows the low-to-high regime transition probability $\mathsf{P}[f_{t+i} > 0.5 | f_t = .17]$ (solid line) and the high-to-low regime transition probability $\mathsf{P}[f_{t+i} < 0.5 | f_t = .88]$ (dashed line) as a function of i (measured in months). The figure uses parameter values from Table A.1.

about 6-8%.

1.4.2 Price Drops and Volatility Spikes

Figure 1.6 shows model quantities calculated using the parameters in Table A.1. The first three panels show the price coefficient functions a, b, and c from (1.11). The lower right-hand panel shows the expected net profit from owning one share of the stock. When no investor is informed, no dividend information is reflected in the price, and $b = 0$. As f increases to the point where some investors become informed, b and c both increase, which drive up the price variance.[10] The increase in c reflects a higher compensation for accommodating supply shocks and is attributable to higher price variance and a growing informational disadvantage of the uninformed relative to the informed.

The upper left panel of Figure 1.6 shows that $a(f)$ drops sharply as f increases. The left panel of Figure 1.7 shows the resulting effect on the expected stock price $P_0 \equiv a(f) + d\bar{D}$. The price response is dramatic: a small increase in f leads to a price drop of 10%. We will

[10] As b measures the sensitivity of the price to dividend information, the monotonicity of b parallels an empirical finding in Brancati and Macchiavelli (2019) that prices become more information-sensitive when information precision increases.

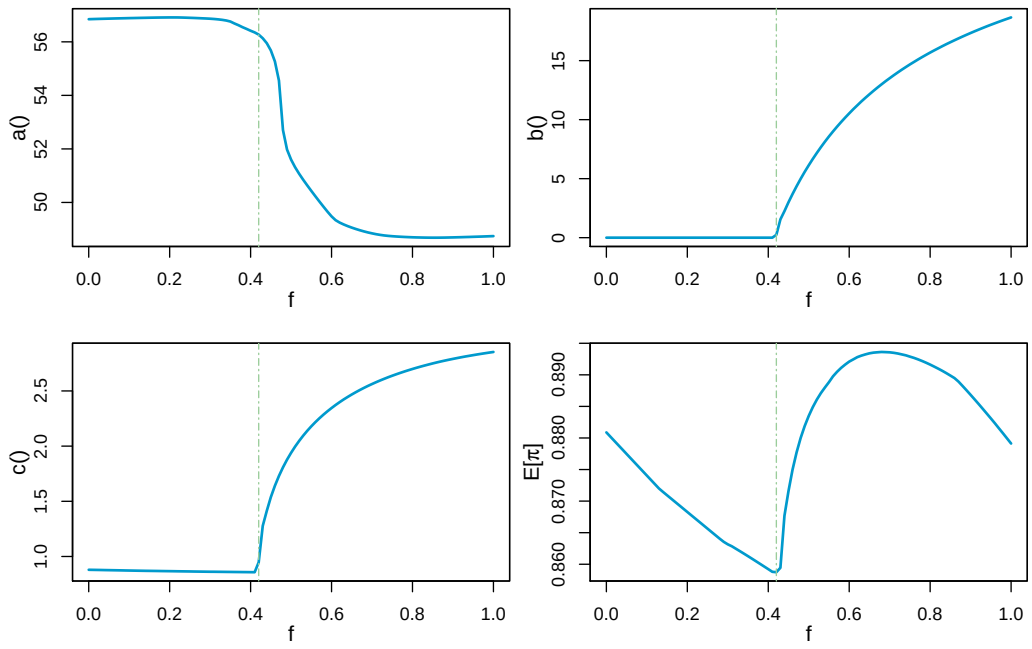


Figure 1.6: Price coefficients and expected profit as functions of f

Note: Equilibrium price coefficients $a(f)$, $b(f)$, and $c(f)$ and the expected investment net profit $\mathsf{E}[\pi]$ per period. The vertical, dashed green line shows the point at which λ becomes positive. The figure uses parameter values from Table A.1.

explain the price drop in Section 1.4.3, using the conditional variances of net profit π in the right panel of the figure.

The price variance $V_B(f)$ in (1.17), shown in the middle panel of Figure 1.7, increases monotonically in f, together with $b(f)$ and $c(f)$. It follows that the price drop associated with an increase in f is accompanied by a spike in volatility. Indeed, in the low-f region where $b(f) = c(f) = 0$, $V_B(f)$ in (1.17) is below 0.5; but for f near 0.88, $V_B(f)$ exceeds 1.5, so the change in information regime produces more than a three-fold increase in price variance.

It is customary to associate large declines in market values with the arrival of bad news. Following a 10% decline (the price drop in Figure 1.7) in an individual stock price or the overall market, one would expect media and expert accounts of what bit of bad news — a product failure, a CEO scandal, a change in government policy — triggered the fall. But in our setting it is simply more news — in the form of increased precision f_t — that

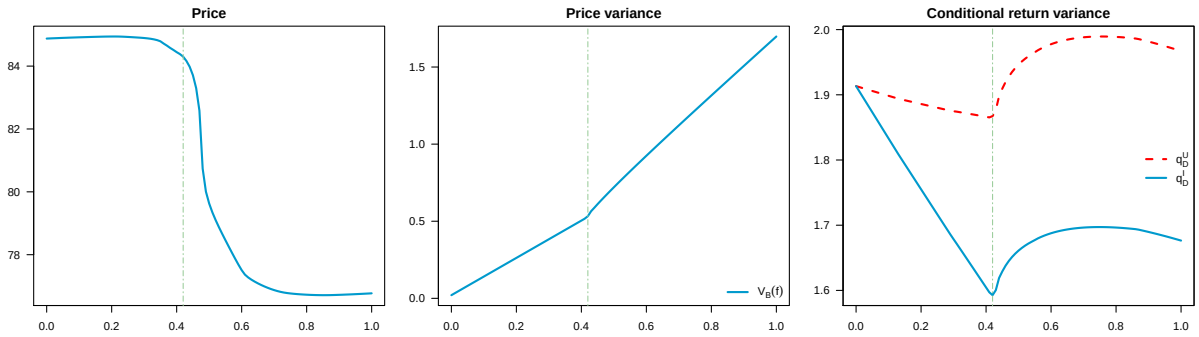


Figure 1.7: Price and variances as functions of f

Note: Equilibrium expected price $P_0 \equiv a(f) + d\bar{D}$ and conditional expected return variance for informed, q_D^I, and uninformed, q_D^U, as functions of information state f. The figure uses parameter values from Table A.1.

drives investors, not necessarily good or bad news.[11] In Section 1.4.4 we show that a negative correlation between information shocks and shocks to fundamentals can significantly amplify the price drop.

The potential for increased volatility from increased information has policy implications. A regulatory change that leads to persistently higher information precision for informed investors is potentially destabilizing in times of market stress.[12] Interestingly, in their analysis of disclosure of the results of regulatory stress tests for banks, Goldstein and Leitner (2018) conclude that disclosure is valuable only under adverse conditions. Our results do not conflict but rather reflect different considerations, as the objective in Goldstein and Leitner (2018) is optimal risk sharing among banks, and the information disclosed separates weak and strong banks.

[11] In a macro context, CESA-BIANCHI and FERNANDEZ-CORUGEDO (2018) find that an increase in economic uncertainty results in a decrease in the risk premium, which is consistent with our results.

[12] The information disclosed about regulatory stress tests is disclosed publicly, but the design of scenarios and the interpretation of the results are technical matters that are arguably accessible only to informed investors who have acquired the necessary expertise.

1.4.3 Decomposing Price and Volatility

The price drop in Figure 1.7 is driven by the drop in the $a()$ curve in Figure 1.6. In Appendix A.4.2 we show that a_t can be decomposed into two components,

$$a_t = \frac{(1+d)\mu_D}{R-1} - \sum_{i=1}^{\infty} \frac{1}{R^i} \mathsf{E}_t[\pi_{t+i}], \tag{1.18}$$

where, π_{t+i} is the net profit from holding one share of the stock from $t+i-1$ to $t+i$, as in (1.8), and E_t denotes conditional expectation given f_t. From this expression, we see that the $a_t + dD_t$ component of P_t is the present value of all future expected dividend payments minus a discount reflecting the expected present value of all future net profits.[13] In the context of the Campbell (1990) and Vuolteenaho (2002) return variance decomposition, the second term in a_t represents the effect of a time-varying discount rate on the stock price. To understand how a change in information precision f_t creates a price drop, we need to understand the effect of f_t on the second term in a_t.

The stock's expected net profit over a single period is given by[14]

$$\mathsf{E}_t[\pi_{t+1}] = \underbrace{\gamma}_{\text{risk aversion}} \times \underbrace{\bar{X}}_{\text{asset supply}} \times \underbrace{\left(\lambda \frac{1}{q_D^I} + (1-\lambda)\frac{1}{q_D^U}\right)^{-1}}_{\text{average uncertainty}}, \tag{1.19}$$

where q_D^I and q_D^U are the expected conditional variances of the net profit for informed and uninformed investors,

$$q_D^{I/U} = \mathsf{E}\left[\mathsf{var}(\pi_{t+1}|\mathcal{I}_t^{I/U}, f_{t+1})\middle|\mathcal{I}_t^{I/U}\right].$$

Equation (1.19) thus reflects the average return uncertainty faced by investors, weighted by the fractions of informed and uninformed in the economy, and scaled by $\gamma\bar{X}$.

The right panel of Figure 1.7 shows q_D^I (solid line) and q_D^U (dashed). The shape of these curves reflects the tradeoff engendered by increased information precision. When f is low, an increase in f decreases the expected variance of net profits for informed and uninformed investors because more is known about next period's dividend, the D_{t+1} term in (1.8). As

[13] If the second term in (1.18) is zero, we get $a_t + dD_t = \sum_{i=1}^{\infty} \frac{1}{R^i}\mathsf{E}[D_{t+i}|D_t]$ from (1.1).

[14] This is shown in equation (A.27) of the appendix. This expression generalizes the corresponding quantity derived from equation (A10) in Grossman and Stiglitz (1980).

long as f is low enough so that $\lambda(f) = 0$ (to the left of the vertical, dashed line in the graphs) this is the only effect, and higher information precision lowers uncertainty. However, past the no-informed point, with f large enough that $\lambda(f) > 0$, the uncertainty of next period's net profit starts to increase, due to the increasing variance of P_{t+1} in the expression for π_{t+1} in (1.8). This effect outweighs the decrease in the variance of next period's dividend, and thus increases the conditional variances of the net profit. For high enough f the increased information about next period's dividend begins to dominate, and the expected conditional variance begins to fall again. This pattern depends crucially on the dynamic structure of our model: in a single-period setting, where investors care about the next dividend but not future prices, more precise information always reduces investment uncertainty.

Through (1.19), the common shape of q_D^I and q_D^U is inherited by $\mathsf{E}_t[\pi_{t+1}]$, as illustrated in the bottom-right panel of Figure 1.6. Recall from our discussion of Figures 1.4 and 1.5 that f_t spends most of its time near $f = 0.175$ or near $f = 0.88$. From the bottom-right panel of Figure 1.6, we see that $\mathsf{E}_t[\pi_{t+1}]$ is greater near $f = 0.88$ than it is near $f = 0.175$, indicating an increase in the expected profit from holding the stock as we move from the low-information regime to the high-information regime. This increase in expected profit is associated with a decrease in the current price of the stock, and it contributes to the price drop we see in Figure 1.7.

Notice, however, that the change in expected net profit across regimes is quite small, as indicated by the vertical scale in the lower-right panel of Figure 1.6. How does a small change in expected profit get amplified into a 10% price drop? The answer lies in the combination of the price discount reflected in (1.18) and the persistence of the two f_t regimes.

We saw in the right panel of Figure 1.5 that transitions between $f_t \approx 0.175$ and $f_t \approx 0.88$ are rare. We observed the inequality $\mathsf{E}_t[\pi_{t+1}|f_t = 0.88] > \mathsf{E}_t[\pi_{t+1}|f_t = 0.175]$ in Figure 1.6. As a consequence of the persistence in regimes, we expect this inequality to extend to $\mathsf{E}_t[\pi_{t+i}|f_t = 0.88] > \mathsf{E}_t[\pi_{t+i}|f_t = 0.175]$, for large i. The present value of such terms is subtracted from the price P_t through the a_t coefficient in (1.18). Thus, even a relatively small single-period difference in expected profits around $f = 0.175$ and $f = 0.88$ is amplified to a large change in the price because the f_t process spends long periods in each of the two

regimes before moving towards the other.[15]

Are such infrequent regime transitions plausible? Barro (2009) estimates country level crises occur with a 1.7% per year probability. Assuming independence across time, a given country has a 29% (i.e., $1-(1-0.017)^{20}$) probability of experiencing at least one crisis over a 20-year period. As we saw in Figure 1.5, the probability of a low to high state transition in our model is approximately 7% over a 20-year period. Our calibration therefore suggests that one out of four country-level crises may be accompanied by the information-driven price drop of our model. If crises are typically associated with positive shocks to f_t, the actual ratio may be higher. We treat this correlated case next.

1.4.4 Correlated Information and Dividend Shocks

We have thus far kept the dividend shocks and information shocks independent of each other. This separation has allowed us to isolate the impact of a more precise signal for informed investors from the impact of specifically positive or negative signals. In practice, events that fuel investor demand for greater information are often accompanied by adverse effects on fundamentals. We therefore extend our model to allow negative correlation in shocks to f_t and shocks to dividends. As expected, this correlation amplifies the resulting price drop.

We introduce correlation by replacing the dividend innovation M_{t+1} in (1.2) with

$$M_{t+1} = m_t + \theta_t + \epsilon_{t+1} - \epsilon_{f,t+1}, \tag{1.20}$$

where $\epsilon_{f,t+1}$ is the shock to f_t in (1.6). Our previous solution method goes through essentially unchanged because of the form of the utility functions in (1.9). The conditional mean and variance there are conditioned on f_{t+1} and thus on $\epsilon_{f,t+1}$. See Section A.8 of the Supplementary Appendix for details.[16]

Figure 1.8 illustrates the effect of correlation through impulse response functions. We start with f_t at 0.4, a level from which it can move toward its high or low regimes with a single

[15]The same argument predicts a sharp decline in the $a()$ curve around the unstable fixed point near $f = 0.48$ in Figure 1.4. Starting to the right of 0.48, f_t will tend to move toward 0.88, whereas starting to the left of 0.48, f_t will tend to move toward 0.175.

[16]The case of $\epsilon_{f,t+1} = \epsilon_{f1,t+1} + \epsilon_{f2,t+1}$ and $M_{t+1} = m_t + \theta_t + \epsilon_{t+1} - h \times \epsilon_{f1,t+1}$ also leaves our solution method unchanged, assuming the inner expectation and variance in (1.9) condition on $\epsilon_{f1,t+1}$ and $\epsilon_{f2,t+1}$.

shock. The solid blue curves show the response to a pure dividend shock of $\epsilon_{t+1} = -0.135$. This shock has no effect on f_t so, as expected, f_t mean reverts toward 0.175. In the lower panel, the expected price given f_t drops because of the negative dividend shock, and then gradually recovers.

For the dashed red curves, we set $\epsilon_{f,t+1} = 0.135$; this is the size of a positive information shock in Table A.1, which is why we chose this magnitude. Through (1.20), the dividend shock is again -0.135, but now the shock affects both D_{t+1} and f_{t+1}. In the top panel of Figure 1.8, the red curve shows that following a positive shock, f_t is pulled toward 0.88. The lower panel shows that the price drop is now much greater, because it reflects the combined effect of higher f_t and a lower dividend.

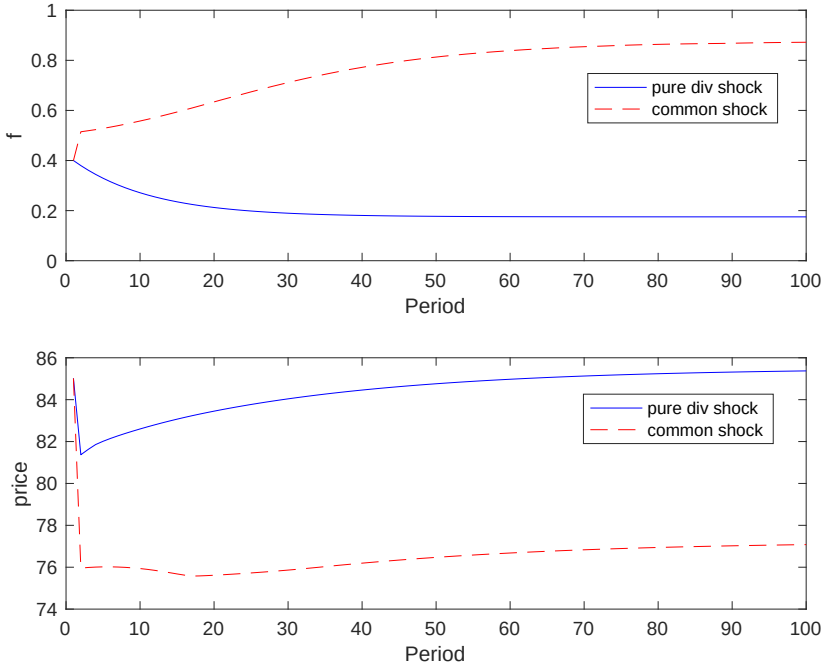


Figure 1.8: An example path under correlated shocks

Note: Response of f_t (top) and price (bottom) to a dividend shock without correlation (blue) and with correlation (red).

At the onset of a crisis, we expect both a decline in fundamentals and, in response to investor demand, an increase in information production. The contrast between the pure dividend shock and the joint dividend-information shock in Figure 1.8 indicates that the demand for information can significantly amplify a shock to fundamentals, leading to a

deeper and more prolonged crisis. Our model suggests that when fundamentals shocks happen in an already elevated information environment — had f_t been lower than 0.4 in the simulation, the positive information shock would not have induced a transition to the high-information regime — the effect of adverse fundamentals shocks can be greatly exacerbated. For example, had the 2009 Greek budget deficit revision happened in a less media-heavy climate than what prevailed in the aftermath of the Global Financial Crisis, the impact on Greek (and other European sovereign) markets may have been much smaller.

1.5 Exploration of the Mechanism

This section further investigates the features of our model that drive its behavior. Sections 1.5.1 and 1.5.2 connect the price drop with the degree of information asymmetry and the cost of information production. Section 1.5.3 contrasts our model with models of strategic complementarity in information acquisition.

1.5.1 The Role of Time-Varying Information Asymmetry

Information asymmetry plays an important role in generating price and volatility cycles in our model. Large price drops occur when the economy transitions from low- to high-information asymmetry states. Whether this can happen is dictated by ϕ, the fraction of knowable information that is private. Figure 1.9 compares equilibrium $a()$ curves for different values of ϕ; the case $\phi = 0.35$ is the one we have analyzed thus far.

When $\phi = 0$ and all knowable information is public, the economy is characterized by no information asymmetry — the knowable information is equally known to all agents. The $a()$ curve corresponding to this no-asymmetry case is the highest one (shown as a solid line), indicating the smallest price discount relative to the present value of future dividends. The $a()$ curve in this case is quite insensitive to f_t. The $\phi = 1$ case represents the highest informational asymmetry possible in the model, and corresponds to the lowest $a()$ curve, representing a large price discount needed to induce the informationally disadvantaged uninformed agents to participate in risk sharing, regardless of the information state f_t.

Only for intermediate values of ϕ can the economy transition from low- to high-information asymmetry states. Such regime shifts are accompanied by large price changes.

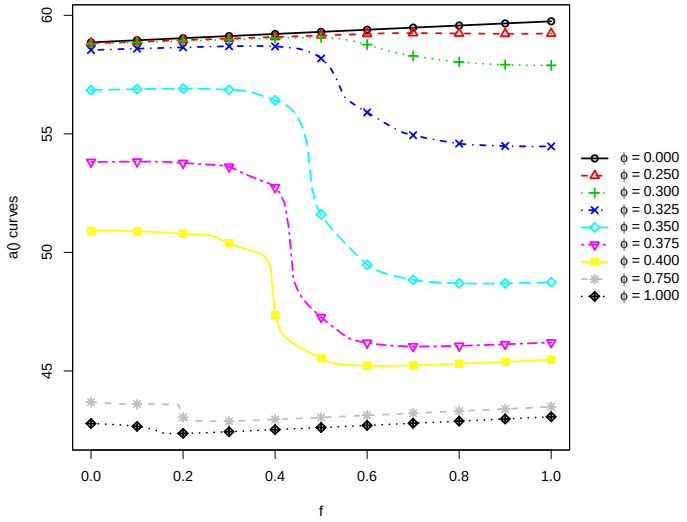


Figure 1.9: $a()$ curves as functions of f for different values of ϕ.

Note: The figure uses parameter values from Table A.1.

The reason that prices in the cases $\phi = 0$ and $\phi = 1$ do not change much across different values of f can be seen from Figure 1.10, which shows the steady-state distribution of f in the different ϕ models. Changing ϕ changes the steady-state distribution because it changes the endogenous $\lambda(f_t)$. When $\phi = 0$, there are no informed investors since all knowable information is public. With $\lambda = 0$ in (1.6), any positive $\epsilon_{f,t+1}$ shock quickly decays, pulling f_t back to its low-information fixed point of 0.175. This dynamic is seen in the unimodal

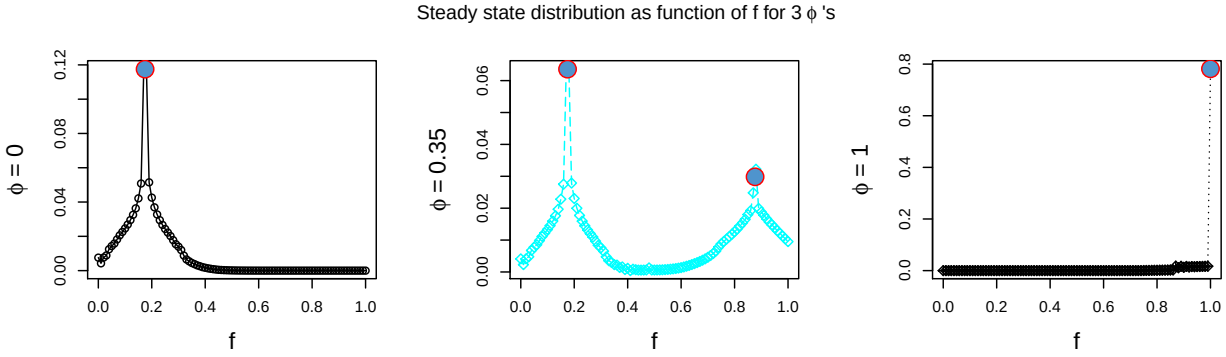


Figure 1.10: Steady-state distribution of f_t at different ϕ's

Steady-state distribution of f_t across different levels of ϕ. The figure uses parameter values from Table A.1. The solid circles indicate fixed points of the $f \to \lambda^*$ mapping.

distribution, with the peak centered at a_f, when $\phi = 0$. Similarly, when $\phi = 1$, all knowable information is private, and $\lambda(f)$ is relatively large. Via the b_f term in the dynamics of f_{t+1} in (1.6), a relatively high λ produces a steady state distribution that is unimodal at $f = 1$. Any negative $\epsilon_{f,t+1}$ quickly dissipates as f is pulled back to one. In these cases, $\mathsf{E}_t[\pi_{t+i}|f_t = f]$ will be close to either $\mathsf{E}_t[\pi_{t+i}|f_t = 0.175]$ or $\mathsf{E}_t[\pi_{t+i}|f_t = 1]$ for any f, and a_t in (1.18) is consequently insensitive to f.

For intermediate values of ϕ (0.35 in our calibration), the equilibrium λ curve is in an intermediate range, and the tendencies of f_t towards a_f and towards the high-information fixed point are balanced. The steady state f_t distribution becomes bimodal as can be seen in Figure 1.10. Therefore, a sequence of shocks can occasionally push the economy from one information regime to the other. Yet both regimes are very persistent. As in Section 1.4.3, this persistence amplifies differences in expected net profit $\mathsf{E}_t[\pi_{t+1}]$ at different values of f to produce large price changes.

This effect results from an increase in information asymmetry, rather than just from an increase in information precision. When f_t is low, there is little information but also no information asymmetry because all agents are uninformed. In this case, prices are high. But as f_t increases, private information becomes more revealing and some investors start to acquire it at a cost. The uninformed then find themselves at a growing informational disadvantage and the price falls.

1.5.2 The Effects of Cheaper Information

Connecting the previous discussion of information asymmetry to our news production sector from Section 1.2.2, we ask what happens in the economy when information gets easier to produce. We can proxy for this by assuming that the per unit cost of news production c_P falls. Dropping the cost of news production c_P, while keeping the expenditure on news producers c_M fixed, results in an increasing b_f (equation A.4 in the appendix shows that $b_f = c_M/c_P$). That is, the feedback $b_f \lambda_t$ from this period's informed to next period's signal precision f_{t+1} becomes larger. Figure 1.11 shows that cheaper information pushes the economy towards the low-price-high-volatility regime. In fact, with a sufficiently low cost of information production, i.e. a very high b_f, the entire weight of the steady-state

distribution gravitates to the low-price-high-volatility state.[17] In contrast, when information production is expensive, there is no feedback, and the economy is always in the high price, low volatility regime.

To analyze the effect of cheaper information on welfare, we look at the utility of the uninformed J^U, which is the utility of all agents in equilibrium because it is always either higher than that of the informed (when $\lambda = 0$) or equal to that of the informed.[18] In the left panel of Figure 1.12, we see that higher b_f (cheaper information production) increases $J^U(f)$ at each f. However, higher b_f also pushes probability mass into the high f region, where $J^U(f)$ is lower. When we take the expectation of $J_U(f)$ over the steady-state distribution of f, the net effect, shown in Figure 1.12, is to lower expected utility. In particular, then, a lower cost of information production, i.e., a higher b_f, reduces welfare. As before, this conclusion is a consequence of the degree of information asymmetry determined by ϕ. Greater information production is welfare-reducing when a substantial fraction of that information remains private.

1.5.3 The Value of Becoming Informed

A key property of the Grossman and Stiglitz (1980) setting is that the value of becoming informed decreases as the number of informed investors increases. Subsequent work has investigated conditions in which the value of becoming informed increases as more investors become informed. Sources of this type of strategic complementarity identified in the literature include high fixed costs and low marginal costs in information production (Veldkamp (2006)); certain deviations from normally distributed uncertainty (Chamley (2007)); settings in which investors learn about supply as well as cash flows (Ganguli and Yang (2009) and Avdis (2016)); other settings with multiple sources of information (Manzano and Vives (2011) and Goldstein and Yang (2015)); and settings in which information acquisition affects cash flows (Dow et al. (2017). With few exceptions, these are static models, but they often result in multiple equilibria, with different asset prices in different equilibria.

[17] The steady-state distributions of f_t with $b_f = 0$, $b_f = 0.384$, and $b_f = 1.536$ are very similar or identical to the three distributions in Figure 1.10 at three values of ϕ: increasing b_f shifts mass to the right.

[18] This would not be true if $\lambda = 1$, which doesn't happen in our calibration.

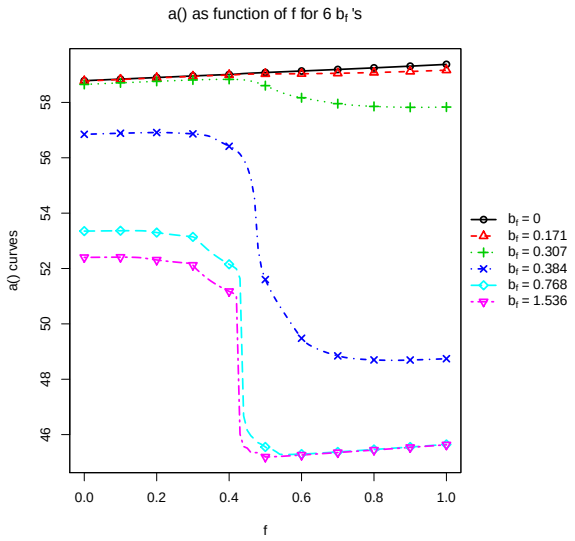


Figure 1.11: $a()$ curves at different b_f's

Note: The figure shows $a()$ curves as function of f_t across different media regimes. The figure uses parameter values from Table A.1.

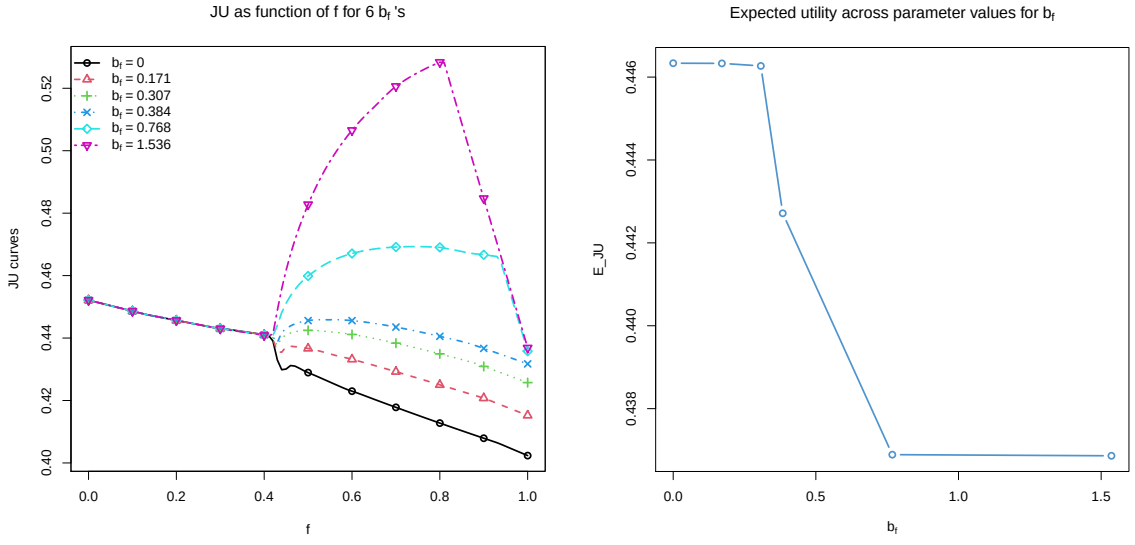


Figure 1.12: Investor utility at different b_f's

Note: The left panel shows that J_U increases with b_f at each level of f. But the right panel shows that expected utility decreases with b_f. This happens because increasing b_f shifts the distribution of f_t to the right. The figure uses parameter values from Table A.1.

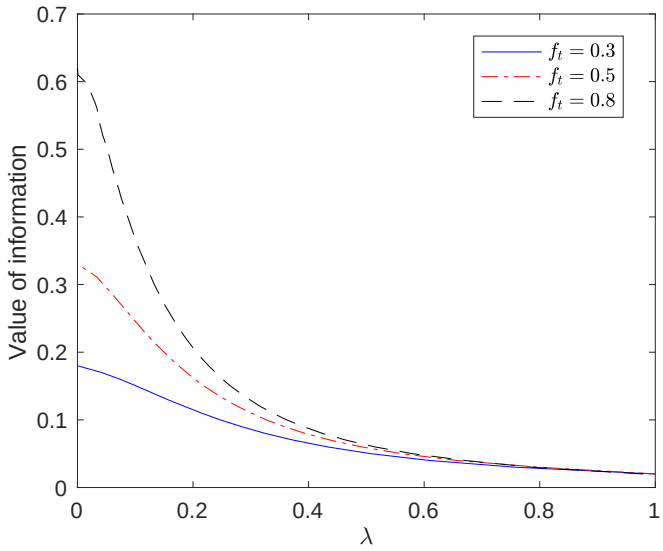


Figure 1.13: Value of information by λ

Note: The curves show how the equilibrium value of becoming informed varies with the fraction informed $\lambda(f)$, at different levels of f. As more investors become informed, the value of becoming informed decreases.

In our dynamic setting, large price changes occur within the model, rather than through a change of equilibrium selection in a static model. But we will see that the contrast with earlier work goes beyond this feature. In order to explore strategic complementarity, we vary the cost c_I of becoming informed and recalculate the equilibrium. For each c_I, we get a new $\lambda(\cdot)$ function; we can fix a value of f and see how $\lambda(f)$ varies with c_I. Because $\lambda(f)$ is the equilibrium fraction informed, c_I is precisely the value of becoming informed when $f_t = f$.

Figure 1.13 shows the results of these calculations at three levels of f. The curves are calculated by varying c_I and recomputing $\lambda(f)$. But we read them in reverse, i.e. with $\lambda(f)$ on the horizontal axis, to see how the equilibrium value of becoming informed varies with the fraction informed $\lambda(f)$. The curves are decreasing, indicating that the value of becoming informed decreases as more investors become informed. In this respect, our model shows the behavior in Grossman and Stiglitz (1980) and differs from the large class of subsequent models that exhibit strategic complementarity in the value of becoming informed.

The underlying source of the feedback effect in our f_t dynamics (1.6), as developed

through the microfoundation in Appendix A.1, is the response of a competitive information production sector to increased demand when the marginal cost of transmitting already discovered information is zero. A similar cost structure of information production is highlighted by Veldkamp (2006) as a source of strategic complementarity, so it is interesting to contrast the implications of our models. The models have different objectives and differ in many respects; but, most notably, in Veldkamp (2006) higher prices are associated with a larger fraction of informed investors, whereas we saw in Section 1.4.2 that in our model an increase in λ_t can precipitate a large price drop.

Complementarity in Veldkamp (2006) arises because the price of information in equilibrium is declining in the number of informed investors; with more informed investors, there is the same amount of information, but at a lower per capita cost. In our case, the amount paid by each investor for information is fixed at c_M; therefore with more informed investors, more information gets produced. We show in Section A.5.3 that each informed investor optimally chooses to pay c_M in equilibrium. This difference leads to information cycles in our model, and these are absent in Veldkamp (2006).

Another key feature driving the difference is the dual role of information in our dynamic model discussed in Section 1.4.3: more precise information decreases dividend uncertainty but can increase future price variance. The second effect is absent in Veldkamp (2006), where investors earn dividends but do not earn capital gains from reselling their shares, making them indifferent to price variance. With no dependence on the next period's prices, the analysis reduces to a sequence of single-period problems.

To see that end-of-period prices drive the difference in our conclusions we can work backwards as follows. The price drop in our model is driven by the drop in $a(f)$ in (1.18), which, through (1.19), is driven by the increase in the conditional variances q_D^U and q_D^I of π_{t+1} shown in Figure 1.7. If the return π_{t+1} in (1.8) did not include P_{t+1}, the conditional variances q_D^U and q_D^I would instead decrease with f as the uncertainty in D_{t+1} decreases, eliminating the price drop we see with increasing f.

1.6 Conclusions

We have developed a model of a financial market in a dynamic information environment and shown that information dynamics can have a profound effect on prices. The model combines exogenous shocks to the level of potentially available information, an endogenous response by investors, and feedback from investor information choices to the information environment through information production by a competitive research sector. The dynamic structure of the model leads to a dual role for information, in which greater information reduces uncertainty about the next dividend but may increase price variance.

We show that the equilibrium dynamics of our model, calibrated to market data, are characterized by two regimes, one with high prices and low volatility, and one with low prices and high volatility. A transition from the first regime to the second is reminiscent of a financial crisis but with no change in fundamentals — the price drop is driven by the dynamics of information and an increase in information asymmetry. In the case of correlated information and fundamentals shocks, an increase in information production can meaningfully increase the price impact of adverse dividend shocks.

Furthermore, we show that in our calibration, the effect of an increased feedback from today's informed to future information is welfare decreasing in the steady-state of the economy. This is true despite the fact that for any given level of the information state, more feedback makes the current set of investors better off.

Our model points to an important role for information dynamics in financial crises. At the onset of a crisis, growing investor demand for information met with increasing production of non-public information can drive down prices and increase volatility. The effect can be counteracted by making costly information public to reduce the asymmetry between informed and uninformed investors.

Chapter 2

Index-based Investing and Intraday Stock Dynamics

2.1 Introduction

Understanding the pattern of intraday stock dynamics is an important topic with various practical applications. Portfolio managers executing large orders can reduce the transaction costs by trading in the hours with abundant market liquidity; intraday traders can better exploit price comovement of different stocks during the periods when correlation is high; risk managers can reduce intraday risk by avoiding times with large price fluctuations. In this chapter, we will show that the intraday patterns of US stocks have changed in important ways since 2004. For example, we find in the recent decade, the trading volume and correlation increase significantly near the market close; the betas of different stocks are dispersed in the morning, but generally move towards one throughout the day. These patterns demonstrate the substantial implications from passive investment, and more specifically, the index-based strategies.

In the recent decade, the growth of passive investment and index-based strategies have drawn great attention from both industry and academia (Appel et al., 2016). The index-based strategies tend to make investment decisions based on portfolio-level approaches instead of selecting individual stocks in a discretionary way. For example, index-based strategies include buying or selling all S&P 500 constituents according to their market capitalization; investing in stocks with high or low beta; buying stocks in specific sectors, etc. These strategies trade multiple stocks in a systematic manner, and make stocks more likely to move in same directions. On the other hand, it has been widely noticed by financial press that the index-based strategies from passive investors tend to concentrate their trading near the market close (Strumpf (2015) and Driebusch et al. (2018)). This behavior can be explained by, among other reasons, minimizing the tracking errors of orders benchmarked to

the closing price, efficiently buying or selling large number of stocks with market-on-close orders, and reducing the inventory risk for redemption and creation settlement (see relevant discussion in Cushing and Madhavan (2000), Foucault et al. (2005), and Wu (2019)).

In this chapter, we first use two empirical studies to provide evidence for the growth of passive investment and the intraday trading pattern of index-based strategies. In the first one, we find the stocks with high (resp. low) passive ownership have higher trading volume near the market close (resp. open). In the second one, we show the trading volume tends to drop dramatically near the market close after a stock is removed from the S&P 500 Index, thus less tracked by the index-based strategies. These findings suggest an active-open, passive-close intraday trading profile, i.e., more discretionary (resp. index-based) trading at the market open (resp. close). Such trading profile can have substantial impact on the intraday patterns of stock dynamics. Accordingly, we propose four hypotheses that motivate our subsequent studies. First, we expect the correlation between stocks to be low at the market open, and high at the end of trading session. This is because the index-based strategies tend to drive multiple stocks to move in same directions. Similarly, we expect the betas of different stocks to be more dispersed in the morning, but move towards one near the market close due to index-based orders. Next, the daily dispersion in trading volume is supposed to be lower at the end of trading session, as the trading from institutional investors, who execute most of the index-based strategies, is shown to be highly persistent across days (Campbell et al., 2009). Finally, we expect the volatility to be higher at the market open and close, but low during the day.

The impact of institutional and passive investment on stock trading, at both the intraday and overnight level, has been a popular topic in recent years. Karolyi et al. (2012) and Koch et al. (2016) provide empirical evidence showing that trading from ETFs and index funds contributes to the commonality in daily trading volume. Heston et al. (2010) suggest that institutional fund flows and trading algorithms can generate periodicity in intraday returns and volumes. Subsequent work by Bogousslavsky (2016) and Gao et al. (2018) shows, both theoretically and empirically, that delayed portfolio rebalancing from institutional investors leads to positive correlation between the returns in the last and first half hours of adjacent trading sessions. Our work complements and extends this line of research by revealing the

implication of passive investment from other important aspects in high-frequency setting, including intraday correlation, beta, and volume dispersion. We show the intraday patterns indeed changed substantially over years. In particular, the four hypotheses hold in our large dataset, especially in the recent decade during which the passive investment has become more prevalent.

To estimate intraday patterns, high-frequency data plays an indispensable role. With the development of financial technologies, there are growing applications of high-frequency data in various fields of finance. We list some examples below among many others in this rich area. For intraday volatility, Andersen and Bollerslev (1997) analyze the intraday periodicity in volatility and its impact on return dynamics; Andersen et al. (2001) test potential pattern shift in intraday volatility with variance-ratio statistics. Some literature on volatility forecasting with high-frequency data and its applications can be found in Andersen et al. (2003), Hansen et al. (2012), Stroud and Johannes (2014), and Liu et al. (2018). Intraday trading volume is also widely studied. For example, Kappou et al. (2010) analyze how the addition of a stock to an index impacts the trading volume and return on adjacent days; Min et al. (2018) develop a time-varying liquidity model based on intraday trading volume pattern and study its impact on optimal portfolio execution. Some recent work studies the estimation of covariance with high-frequency data and demonstrates its benefit in portfolio allocation. Boudt and Zhang (2015) show that an equal-risk portfolio constructed from jump-robust intraday covariance estimation delivers higher return and lower risk than traditional equal-weight portfolio; Bibinger et al. (2019) reveal that intraday covariances follow periodicity patterns, and increase strongly with the arrival of new information; Bollerslev et al. (2019) show the factor-based covariance estimates can improve the performance of risk minimization portfolio in the high-dimensional setting.

While high-frequency data can provide valuable information, the estimators based on high-frequency data are often contaminated by two undesired issues, i.e., market microstructure noise and asynchronicity in price observations. Ait-Sahalia et al. (2005) show that as sampling frequency decreases to zero, the return variance becomes fully induced by microstructure noise instead of the underlying price process. The well-known "Epps" effect (Epps, 1979) states that the asynchronicity in price observations tends to attenuate the

correlation between stocks. To handle these difficulties in high-frequency setting, there is a vast literature on efficient high-frequency estimators. Consistent estimators for realized variance in the presence of market microstructure noise include the multi-scale sub-sampling method of Zhang et al. (2005) and Aït-Sahalia et al. (2011), the realized kernel estimator of Barndorff-Nielsen et al. (2008), and the pre-averaging approach of Jacod et al. (2009). For realized covariance, estimators accounting for both microstructure noise and asynchronicity include, among others, the quasi maximum likelihood estimator of Aït-Sahalia et al. (2010), the multivariate realized kernel approach of Barndorff-Nielsen et al. (2011), the two-scale method of Zhang (2011), and the factor-based method in Bollerslev et al. (2019).

In this chapter we estimate and analyze the intraday patterns of S&P 500 constituents with a large high-frequency dataset. The dataset consists of 1-second level trade data from the Trade and Quote (TAQ) database for all S&P 500 constituents over 15 years (2004 – 2018). This large sample, both cross-sectional and over time, allows us to obtain robust and general patterns for S&P 500 constituents and examine how the patterns change over years. Specifically, we estimate the intraday correlation, beta, volatility, and trading volume for all stocks or stock pairs in the S&P 500 Index. This establishes a comprehensive picture of various aspects of intraday stock dynamics. We employ the two-scale based estimators developed in Zhang et al. (2005) and Zhang (2011), which are unbiased under market microstructure noise and asynchronicity. Besides, the estimators make full use of the large sample, thus avoid the information loss suffered by traditional estimators based on sparse sampling. Furthermore, the nonparametric two-scale based estimators can be efficiently implemented on a large group of stocks, which is essential for our study.

We find informative intraday patterns for S&P 500 stocks. For realized correlation, We show it exhibits certain intraday patterns that evolve over time. First, in the recent decade, the realized correlation starts low and increases in the morning, stays flat in the middle of the day, and further increases near the market close. The magnitude of the intraday increase in realized correlation is on average larger than 0.2, which is relatively substantial. Second, in 2016 to 2018, the realized correlation for the stock pairs with low daily correlations starts even lower at the market open, and increases rapidly throughout the entire trading session. For example, in 2018, the average realized correlation for the stock pairs with bottom 1%

daily correlations (1140 pairs) increases from −0.1 at the market open to 0.2 at the end of trading session. Similar patterns are also observed for realized correlation between sector pairs. On the other hand, we find the realized beta of different stocks are more dispersed in the morning, but moves towards one near the market close. In 2018, the average realized beta for the high-beta stocks (top 10% daily betas) decreases from 1.85 to 1.23 during the day, while that for the low-beta stocks (bottom 10% daily betas) increases from 0.17 to 0.65. These patterns confirm our hypotheses on the implication of index-based strategies, and reveal the substantial impact of the active-open, passive-close trading profile on various aspects of intraday stock dynamics. As an additional theoretical support, we develop a market impact model with single-stock and index-fund investors, and show the time-varying liquidity provision indeed produces the observed intraday patterns of realized correlation and beta.

Next, we find the intraday trading volume shows a U-shape pattern, with higher volume near the market open and close. Furthermore, the U-shape pattern becomes more skewed to the right in the recent decade, as the trading volume near the market close increases significantly. In particular, the proportion of trading volume in the last half hour of trading session increases from 15% in 2004 to 22% in 2018. Such shift in trading volume, as a consequence of the growth of passive investing, has been widely noticed in recent research (see, e.g., Min et al. (2018) and Wu (2019)). Moreover, we show the daily variation in trading volume is high in the morning, but low at the end of trading session. This can be attributed, in part, to the persistent trading from institutional investors who execute most of the index-based strategies near the market close. Finally, we find the intraday realized volatility shows a U-shape pattern skewed to the left, i.e., starts relatively high at the market open and drops subsequently. Besides, we observe the realized volatility near the market close further decreases after 2012, making the intraday curves flatter at the end. While the intraday volatility and volume have been studied in the literature (see, e.g., Wood et al. (1985) and Pagano et al. (2008)), our large dataset and estimators that are efficient under market microstructure noise enable us to obtain robust intraday patterns and examine how they evolve over time.

The rest of this chapter is organized as follows. In Section 2.2, we show the implication

of index-based investment via two empirical studies, which motivate our estimation and analysis of intraday stock dynamics. Section 2.3 establishes the estimators used in our high-frequency setting. In Section 2.4, we describe the data and implementation details. We provide the main estimation results of intraday patterns in Section 2.5, including realized correlation, beta, volume, and volatility. In Section 2.6, we develop a theoretical market impact model with time-varying liquidity provision. Section 4.6 concludes the study and provides further discussions.

2.2 Implication of Index-based Investment on Intraday Trading

In this section, we use two empirical studies to show the growth of index-based investment indeed impacts the intraday trading activities. We propose four hypotheses on the intraday patterns of stock dynamics, which motivate our study with high-frequency data in subsequent sections.

2.2.1 Evidence from Passive Fund Ownership

First, we demonstrate the growth of index-based investment strategies using the degree of passive fund ownership of S&P 500 constituents. We calculate the passive and active mutual fund ownership following the classification method in Appel et al. (2016)[1]. Specifically, a fund is classified as either passive or active by searching for certain strings in its name that identify index funds and the supplementary information on the index fund indicator from CRSP. Figure 2.1 plots the equal-weight average of the fractions of shares owned by either passive or active funds (left), as well as the ratio of shares owned by passive funds to the total shares owned by both types of mutual funds[2] (right). Note the passive and active shares on the left do not sum to 1 as not all shares are owned by mutual funds. We see the average proportion owned by passive mutual funds significantly increases in the recent

[1] A detailed description can also be found in Appendix A.3 of Glasserman et al. (2019).

[2] The patterns in Figure 2.1 match well with the results in Figure 2 of Glasserman et al. (2019); see also Figure 2.8 in 2018 Investment Company Institute Fact Book.

decade. Thus, we expect more trading from index-based strategies for stocks in the S&P 500 Index.

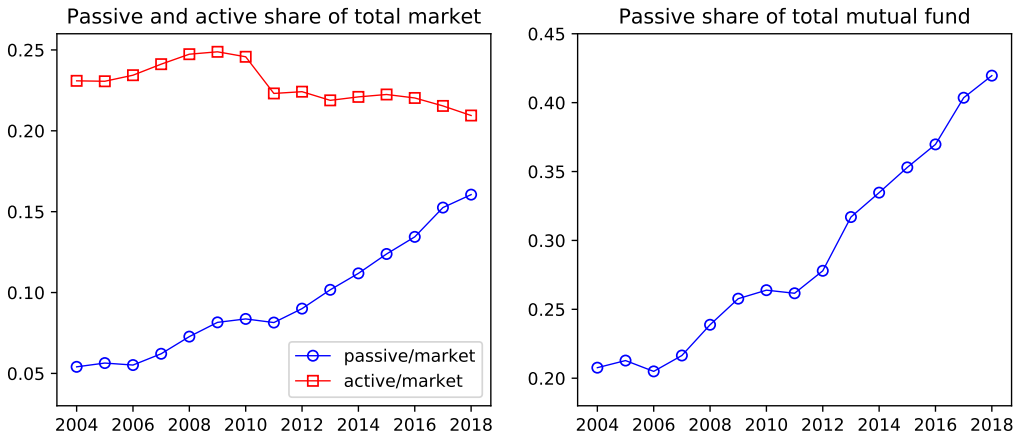


Figure 2.1: Fraction of shares owned by active and passive funds

Note: The left panel shows the fractions of shares owned by either active and passive funds. The right panel shows the ratio of shares owned by passive fund to that owned by mutual funds.

The growth of index-based investment strategies leads to substantial change in the intraday pattern of trading activities. To show this, we check how the intraday distribution of trading volume differs across stocks with high and low passive ownership. We define the scaled trading volume for stock i in time interval t of day d as

$$\text{SVolm}_{idt} = \frac{\text{Volm}_{idt}}{\text{TotVolm}_{id}}, \tag{2.1}$$

where Volm_{idt} is the trading volume in the interval, i.e., the number of shares traded, and TotVolm_{id} denotes the total trading volume of stock i on day d. The scaled trading volume allows us to compare across different stocks, which can have very different shares outstanding and trading volume.

We construct two bins of stocks with low and high passive ownership in each year as follows. For each year, we select the stocks that are in the S&P 500 Index for the entire year. We define their degree of passive ownership as the percent of shares outstanding held by passive mutual funds (averaged over the four quarters). The two bins consist of the stocks with degree of passive ownership below the fifth percentile and above the 95th percentile,

respectively. Thus, each bin has approximately 24 stocks in a given year. The average passive ownership for the two bins is reported in Table B.1 in Appendix C.2.3.

In Figure 2.2, we plot the scaled trading volume of the two bins in the first and last half hours of the trading session, i.e, 9:30 – 10:00 and 15:30 – 16:00. The results are computed as the equal weight average across the corresponding stocks and trading days in the given year. By the left panel, we find the scaled trading volume at the market open is higher for the low passive ownership bin than that for the high passive one since 2008. Moreover, the gap keeps increasing after 2015, and reaches approximately two percentage points in 2018. On the other hand, from the right panel, we see the scaled trading volume near the market close increases significantly since 2008 for both bins, but the magnitude of increase is much larger for the high passive one. In 2018, the scaled trading volume in the last half hour is approximately three percentage points higher for the high passive ownership bin than that for the low passive one. As stocks with low (resp. high) passive ownership are more likely to be traded by active (resp. index-based) strategies, the results in Figure 2.2 suggest an active-open, passive-close profile for intraday trading, i.e., there is more discretionary trading in the morning, and the index-based strategies tend to concentrate their trading near the market close.

The surge in trading volume near the market close, as a consequence of concentrated trading from passive investors, has been widely noticed by financial press (see, e.g., Driebusch et al. (2018) and Strumpf (2015)). The motivations for such behavior include, among others, to minimize the tracking errors of orders benchmarked to the closing price, to efficiently deploy capital to hundreds of underlying stocks using market-on-close order, and to reduce the inventory risk for redemption and creation settlement (see e.g, Cushing and Madhavan (2000), Foucault et al. (2005), and Wu (2019)). Besides, the active-open, passive-close trading profile is also obtained in Min et al. (2018) using the intraday trading volume data of S&P 500 constituents in 2017.

2.2.2 Evidence from Index Removal Effect

We have shown in the previous section that the trading from index-based strategies drives up the trading volume near the market close. In this section, we provide consistent

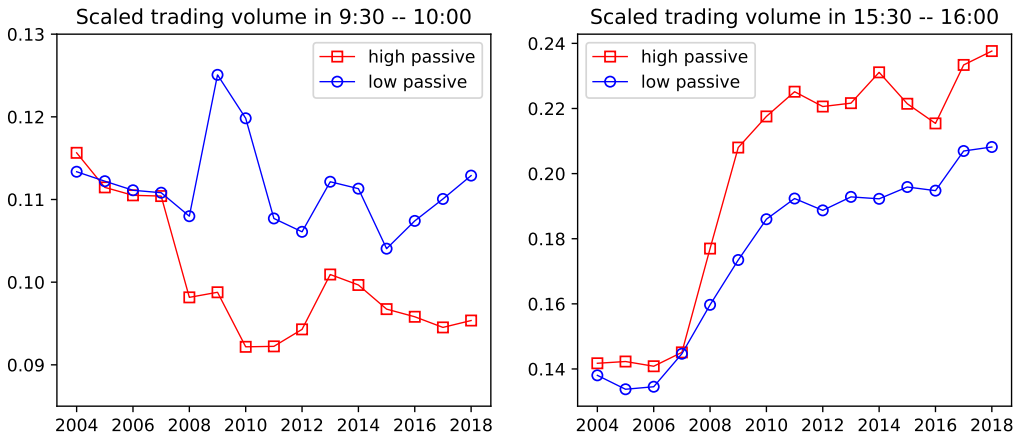


Figure 2.2: Scaled trading volume by passive ownership bins

Note: Scaled trading volume in 9:30 – 10:00 (left) and 15:30 – 16:00 (right) for the low and high passive ownership bins.

evidence for such effect using an impact analysis for the stocks that are removed from the S&P 500 Index. Specifically, we find that after a stock is removed from the index, thus less tracked by index-based strategies, its scaled trading volume tends to drop dramatically near the market close.

We select the stocks that are removed from the S&P 500 Index between 2011 and 2018, during which the index-based strategies have become more prevalent. Moreover, to mitigate the potential impact of delisting, we restrict to the stocks that have at least 60 days of observations in the Trade and Quote (TAQ) database after removal. This leaves us with 54 stocks between 2011 and 2018. We then compute the average scaled trading volume for these stocks in the 30 trading days before and after their removal.

The left panel of Figure 2.3 shows the intraday pattern of scaled trading volume before (red) and after (blue) the removal, which is computed for each half hour interval with a moving step of five minutes from 9:30 to 16:00. As shown in the left panel, the intraday curve of scaled volume is almost unchanged after the removal, except for the last point representing the time interval 15:30 – 16:00. This can be seen more clearly from the absolute change in the right panel, where the shaded area represents the 95% confidence interval of the estimates. We see that after the removal from the S&P 500 Index, the scaled trading volume significantly drops near market close. On average, the scaled trading volume in

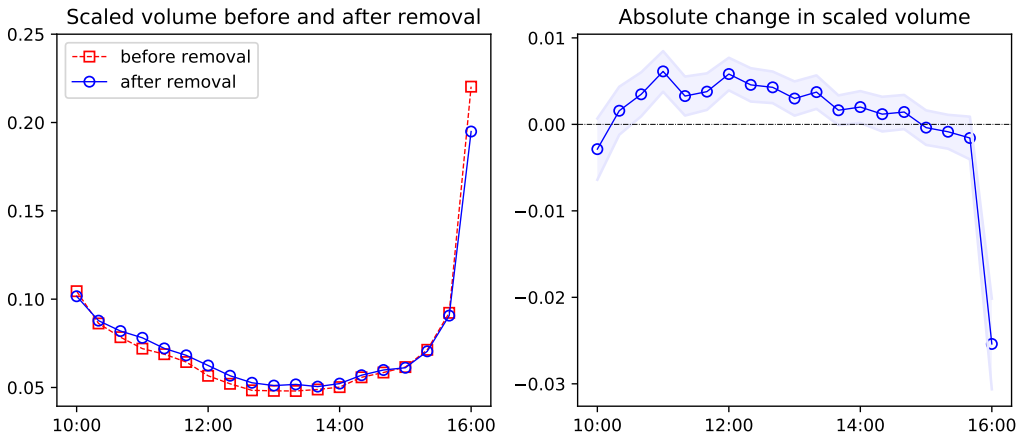


Figure 2.3: Impact on scaled volume by index removal

Note: The left panel shows the intraday scaled trading volume before and after removal. The right panel shows the absolute change.

15:30 – 16:00 drops by 2.5 percentage points, which translates to a relative drop of more than 11%. This decrease is offset by the increases in scaled trading volume before 14:00. However, such increases are much smaller in magnitude and more scattered across the day.

After a stock is removed from the S&P 500 Index, it will be less traded by index-based strategies. This analysis shows that the major impact of the removal is the drop in the proportion of trading volume near the market close. Similar effect is also observed in the literature from other perspectives. For example, Grynkiv and Russell (2015) use the data from 2012 to 2015 and find that the increase in trading volume at the end of trading session is more significant for S&P 500 constituents than for stocks in the less liquid exchange-traded products (ETPs). These results provide consistent evidence for the conclusion that index-based strategies tend to concentrate their trading at the end of trading session.

2.2.3 Hypothesis on Intraday Stock Patterns

From the two empirical studies, we see the growth of index-based strategies and their trading activities substantially impact the intraday stock dynamics. However, unlike the trading volume, the implication on other important aspects, e.g., correlation and beta, remain less studied. In the rest of this chapter, we aim to explore these implications with a large high-frequency dataset and efficient estimation methods.

To begin with, we propose four hypotheses on the intraday patterns of stock dynamics based on the active-open, passive-close trading profile. We expect the patterns described in the hypotheses to be more significant in the recent decade, during which the index-based strategies have become more prevalent.

First, we expect the correlation between different stocks to be lower in the morning, but higher near the market close. This is because the index-based strategies tend to trade multiple stocks in the same direction simultaneously, thus driving up the correlation at the end of trading session. Some examples include buying all S&P 500 constituents and investing in all stocks in target sectors. On the other hand, more discretionary trading from active strategies in the morning tends to result in lower correlation, as active strategies focus more on the specific shocks related to individual stocks. The above analysis leads to our following hypothesis on the intraday pattern of correlation.

Hypothesis 1 (**H1**)**.** *The intraday correlation is low at the market open and high near the market close.*

Next, we propose a hypothesis on the intraday pattern of beta, which measures the level of systematic risk of individual stocks. With more index-based trading near the market close, we expect the betas of different stocks to move towards one at the end of trading session, as individual stock returns are more driven by index-level orders. By contrast, we expect betas of different stocks to be more dispersed at the market open, as discretionary trading captures the heterogeneity in their levels of systematic risk. This leads to our second hypothesis as below.

Hypothesis 2 (**H2**)**.** *The intraday betas of different stocks are more dispersed in the morning, but move towards one near the market close.*

In addition, it has been empirically observed that the trading from institutional investors is highly persistent across days (Campbell et al., 2009). As the index-based strategies are mostly executed by institutional investors (via passive vehicles), we expect lower daily volume dispersion near the market close, i.e., the corresponding trading volume is more stable across days. On the other hand, the active strategies at the market open tend to

focus more on the short-term price fluctuations and incoming news flows. This leads to trading activities that vary much across days. Thus, we expect the daily volume dispersion to be higher at the market open. To summarize, we have following hypothesis on daily dispersion in trading volume.

Hypothesis 3 (H3). *The daily dispersion in trading volume is high at the market open and low near the market close.*

Finally, we propose a hypothesis on the intraday pattern of stock volatilities. As there are more information acquisitions in the morning (e.g., reacting to overnight news flows), we expect the volatility to be higher at the market open. The information acquisitions can be driven by both active and index-based investors. For example, a good news about a specific company may motivate the active investors to buy its stock, and a macro news that benefits the overall equity market (e.g., a rate cut announcement) may trigger the index-based strategies to buy the components of S&P 500 Index. On the other hand, we also expect the volatility to increase near the market close, as there are more trading activities from index-based strategies and the demand for inventory management.

Hypothesis 4 (H4). *The volatility is high at the market open and near the market close, but low during the day.*

In subsequent sections, we use a large high-frequency dataset to show the intraday patterns discussed in the four hypotheses indeed hold for S&P 500 constituents, especially in the recent decade. This reveals, from multiple aspects, the substantial implications of index-based investment on intraday stock dynamics.

2.3 Estimation Methodologies in High-Frequency Setting

In this section, we introduce the estimation methods in our high-frequency setting. Specifically, we define the estimators for realized variance, covariance, correlation, and beta. The estimators account for both market microstructure noise and observation asynchronicity, and can be efficiently implemented on our large dataset.

2.3.1 Estimators for Realized Variance and Covariance

We first introduce the estimators for realized variance and covariance with high-frequency data, which serve as an indispensable foundation for the estimation of realized correlation and beta. We employ the Two-Scale Realized Variance (TSRV) and Two-Scale Realized Covariance (TSRCV) estimators developed in Zhang et al. (2005), Aït-Sahalia et al. (2011), and Zhang (2011). The two estimators are unbiased under market microstructure noise and asynchronicity, and avoid information loss by using all price observations.

The TSRV estimator is established as follows. Suppose we estimate the realized variance for stock Y over a target time interval. We observe the log price Y_i at a series of time points $i = 0, 1 \ldots, n$. Here we consider a fixed grid of sampling intervals, e.g., every five seconds. We select the last observation in each interval, or use the most recent one if there is no observation in the current interval. This is analogous to the previous-tick interpolation commonly used in high-frequency literature (see, e.g., Gençay et al. (2001)). As we focus on the S&P 500 constituents in this study, the stocks considered are generally highly liquid with frequent price observations.

The observed price Y_i can be viewed as a sum of the true underlying price and the market microstructure noise. This introduces an essential challenge for estimating realized variance with high-frequency data. The most naive way to estimate realized variance is to sum all the squared returns in the time interval, i.e., $\mathrm{RV}^{(nv)} = \sum_{i=0}^{n-1}(Y_{i+1} - Y_i)^2$. However, as shown in Ait-Sahalia et al. (2005), this naive estimator is biased by market microstructure noise, and the bias increases in the number of observations n. Thus, this estimator can be severely contaminated when sampling frequency is high. The most straightforward remedy for this is to sample sparsely. For instance, the estimator with observations sampled every J steps can be constructed as $\mathrm{RV}^{(sp)} = \sum_{i=0}^{n/J-1}(Y_{(i+1)J} - Y_{iJ})^2$. This sparse estimator is widely employed in the literature, with the sampling interval chosen in an ad hoc way from 5 to 30 minutes (see, e.g., Gençay et al. (2002) and Barndorff-Nielsen and Shephard (2002)). While the sparse estimator reduces bias, it inevitably leads to information loss. Such loss can be significant when sampling frequency is high: if we sample every minute for 1-second level price observations, we implicitly discard 59/60 of the original data as only the last

observation of each minute is used. An explicit analysis of the naive and sparse estimators can be found in Zhang et al. (2005).

To overcome the above dilemma, we estimate realized variance by the TSRV estimator proposed in Zhang et al. (2005) and further developed in Aït-Sahalia et al. (2011). The TSRV estimator circumvents the two challenges discussed above: it uses all price observations, but yields an unbiased and consistent estimator of the underlying integrated volatility[3]. The spirit of the TSRV estimator is to correct the bias by combining the returns from two time scales, i.e., a fast and a slow one. For a time scale J, define the following sum of squared returns

$$[Y,Y]^{(J)} = \frac{1}{J}\sum_{i=0}^{n-J}(Y_{i+J} - Y_i)^2.$$

Similar to the sparse estimator, the term $[Y,Y]^{(J)}$ is also based on J-step returns. However, it moves by one step each time and thus uses all the observations. This avoids any loss in the price information. Then, the TSRV estimator with fast scale J and slow scale K is given by

$$\mathrm{RV}^{(J,K)} = \frac{n}{(K-J)\bar{n}_K}\left([Y,Y]^{(K)} - \frac{\bar{n}_K}{\bar{n}_J}[Y,Y]^{(J)}\right) \quad \text{for } J < K, \tag{2.2}$$

where $\bar{n}_J = (n - J + 1)/J$ and $\bar{n}_K$ is defined similarly. Thus, the TSRV estimator is a linear combination of the squared terms of two time scales. As a nonparametric estimator, it can be efficiently implemented on a large set of stocks over a long period. This advantage is essential for our study, which involves estimation for all S&P 500 constituents across 15 years. The realized volatility is simply computed as the square root of the TSRV estimator in (2.2).

Next, we briefly introduce the TSRCV estimator for realized covariance. As we have discussed, the estimation of covariance under high-frequency setting is biased due to asychronicity and microstructure noise. To cope with these two challenges, we employ the TSRCV estimator proposed in Zhang (2011), which can eliminate the two types of bias simultaneously. The TSRCV estimator follows the same spirit of the TSRV estimator in (2.2), i.e., correcting the bias by combining the returns from two time scales. Besides, sim-

[3] Aït-Sahalia et al. (2011) further show the bias-corrected and consistent properties of the TSRV estimator hold even when microstructure noise exhibits time series dependence.

ilar to the TSRV estimator, the TSRCV estimator is nonparametric and can be efficiently implemented on a large set of stocks.

The TSRCV estimator is constructed as follows. Suppose we estimate the realized covariance for two stocks X and Y over a target time interval. We observe the log prices X_i and Y_i at a series of time points $i = 0, 1, \ldots, n$. Same as for the TSRV estimator, here we consider a fixed time grid and apply previous-tick interpolation to handle missing observation. For a time scale J, define the term $[X, Y]^{(J)}$ as

$$[X, Y]^{(J)} = \frac{1}{J} \sum_{i=0}^{n-J} (X_{i+J} - X_i)(Y_{i+J} - Y_i).$$

Then the TSRCV estimator is given by

$$\mathrm{RCV}^{(J,K)} = \frac{n}{(K-J)\bar{n}_K} \left([X, Y]^{(K)} - \frac{\bar{n}_K}{\bar{n}_J} [X, Y]^{(J)} \right) \quad \text{for } J < K, \tag{2.3}$$

where $\bar{n}_J = (n-J+1)/J$ and $\bar{n}_K$ is defined similarly. The TSRCV estimator is unbiased under both observation asynchronicity and market microstructure noise. More detailed analysis of its properties can be found in Section 8 of Zhang (2011).

2.3.2 Estimators for Realized Correlation and Beta

In this section, we develop the estimators for realized correlation and beta, which are based on the TSRV and TSRCV estimators in the previous section. Specifically, we develop two methods to estimate realized correlation. The first one estimates the realized correlation between stock pairs, while the second one estimates the portfolio-implied realized correlation between two sets of stocks.

2.3.2.1 Pairwise Realized Correlation

The estimator for pairwise realized correlation is simply the high-frequency counterpart of the traditional correlation, which is defined by

$$\rho_{X,Y} = \frac{\mathrm{Cov}(R_X, R_Y)}{\sqrt{\mathrm{Var}(R_X)} \cdot \sqrt{\mathrm{Var}(R_Y)}}$$

for stocks X and Y. To estimate their realized correlation, we just plug in the high-frequency counterparts of the covariance and variance, i.e.,

$$\mathrm{RCorr}_{X,Y}^{(J,K)} = \frac{\mathrm{RCV}_{X,Y}^{(J,K)}}{\sqrt{\mathrm{RV}_X^{(J,K)}} \cdot \sqrt{\mathrm{RV}_Y^{(J,K)}}}, \tag{2.4}$$

where J and K denote the two time scales employed in the TSRV and TSRCV estimators.

2.3.2.2 Portfolio-implied Realized Correlation

Besides, we propose an estimator for the realized correlation between two mutually exclusive sets of stocks. Unlike the pairwise estimator, the new estimator is based on the realized variances of suitably constructed portfolios. Consider two mutually exclusive stock sets A and B. Denote by $w_i > 0$ the weight of stock i (e.g., market-capitalization). Note that we do not require $\sum_i w_i = 1$ as long as the weights are fixed. Define the average return correlation between the two sets of stocks as

$$\bar{\rho}_{A,B} = \sum_{i \in A, j \in B} w'_{i,j} \rho_{i,j}, \tag{2.5}$$

where $\rho_{i,j}$ is the return correlation between stocks i and j; $w'_{i,j}$ is defined by

$$w'_{i,j} = \frac{w_i w_j \sigma_i \sigma_j}{\sum_{k \in A, l \in B} w_k w_l \sigma_k \sigma_l},$$

where σ_i denotes the standard deviation of the return of stock i. Thus, $\bar{\rho}_{A,B}$ in (2.5) is a weighted average of the pairwise correlations $\rho_{i,j}$ for $i \in A$ and $j \in B$. It puts more weights on the stock pairs with larger portfolio weights (w_i and w_j) or more volatile returns (σ_i^2 and σ_j^2).

We now propose the estimator for (2.5) under high-frequency setting. We construct three portfolios using the stock weights w_i: the first two include stocks in A and B respectively, and the third one combines stocks from both A and B. Then, the average correlation (2.5) in high-frequency setting can be estimated by

$$\mathrm{RCorr}_{A,B}^{(J,K)} = \frac{\mathrm{RV}_S^{(J,K)} - \mathrm{RV}_A^{(J,K)} - \mathrm{RV}_B^{(J,K)}}{2 \sum_{i \in A, j \in B} w_i w_j \sqrt{\mathrm{RV}_i^{(J,K)}} \cdot \sqrt{\mathrm{RV}_j^{(J,K)}}}. \tag{2.6}$$

Here $\mathrm{RV}_A^{(J,K)}$, $\mathrm{RV}_B^{(J,K)}$, and $\mathrm{RV}_S^{(J,K)}$ denote the realized variances of the three portfolios respectively; $\mathrm{RV}_i^{(J,K)}$ denotes the realized variance of stock i. They are estimated by the

TSRV estimator with time scales J and K. Equation (2.6) defines the portfolio-implied estimator for realized correlation between two sets of stocks. We document its explicit derivation in Appendix B.1.

When A and B only contain one stock each, we can show by simple algebraic calculation that the portfolio-implied estimator $\mathrm{RCorr}_{A,B}^{(J,K)}$ coincides with the pairwise estimator in (2.4). However, when the two sets have multiple stocks, the portfolio-implied estimator (2.6) significantly reduces the computational burden. In particular, if both sets have N stocks, the portfolio-implied estimator only needs to estimate $2N + 3$ realized variances, while the average correlation based on the pairwise estimator needs to estimate N^2 realized covariances (for each pair) and $2N$ realized variances (for each stock).

2.3.2.3 Estimation of Realized Beta

Finally, we propose the estimator for realized beta under high-frequency setting. A stock's beta measures the level of systematic risk in its return. However, the study of intraday beta with high-frequency data, to our best knowledge, is relatively rare.

Denote the market return by $R_{M,t}$. The traditional beta of stock i is estimated by

$$\beta_i = \frac{\mathrm{Cov}(R_i, R_M)}{\mathrm{Var}(R_M)}, \tag{2.7}$$

where $\mathrm{Cov}(R_i, R_M)$ is the covariance between individual and market returns; $\mathrm{Var}(R_M)$ is the variance of market return. In high-frequency setting, we estimate the realized beta by plugging in the realized variance and covariance into above equation, i.e.,

$$\mathrm{RBeta}_i^{(J,K)} = \frac{\mathrm{RCV}_{M,i}^{(J,K)}}{\mathrm{RV}_M^{(J,K)}}, \tag{2.8}$$

where $\mathrm{RCV}_{M,i}^{(J,K)}$ and $\mathrm{RV}_M^{(J,K)}$ denote the TSRCV and TSRV estimators with time scales J and K. In this study, we use the S&P 500 ETF from SPDR (ticker SPY) to compute the market return, as its high-frequency data are conveniently available in the Trade and Quote (TAQ) database.

2.4 Data and Implementation Details

2.4.1 Data

In this study, we use all the stocks in the S&P 500 Index from 2004 to 2018. The universe is adjusted dynamically to reflect the quarterly rebalancing of the index. The high-frequency data is obtained from the Trade and Quote (TAQ) Daily Product database. The database contains intraday transaction data (both trades and quotes) for all securities listed on US equity exchanges since September 2003. The original data provided is in millisecond level. In our study, we use the trade data and sample every five seconds. By the previous-tick interpolation, we use the last price observation for each interval, or the most recent observation if no trade happens in the interval. Accordingly, trade sizes are summed within each five-second interval to measure the trade volume in the interval. Days with only morning trading hours are discarded, including the days before Independence Day, Thanksgiving, and Christmas.

We get other information on stocks from the Center for Research in Security Prices (CRSP) and Compustat databases. This includes daily prices, market capitalization, and sector information. The TAQ and CRSP databases are linked via the ticker and permno code of stocks. We discard stocks with multiple share classes as well as preferred stocks. Besides, we obtain fund holding data used in Section 2.2 from Thomson Reuters Mutual Fund Holdings and mutual fund classification from CRSP.

Before estimation, we first identify and handle errors and outliers in the high-frequency data by the following two filters. In the first one, we handle "bounce-backs" where price moves by a large amount but then returns to almost the same level immediately. This filter is also employed in previous literature on realized variance (see e.g., Aït-Sahalia et al. (2011)). Denote three consecutive prices as p_1, p_2, and p_3 (each for a five second interval). We regard p_2 as a "bounce-back" if both conditions below are satisfied

$$|r_2| = \left|\ln\left(\frac{p_2}{p_1}\right)\right| > 0.001 \quad \text{and} \quad |p_3 - p_1| < 0.001.$$

That is, the first five second return is larger than 0.1%, and the difference between the first and last prices is smaller than 0.001.

In the second filter, we handle those consecutive outliers which can not be captured by the first one. At each time point t, we first compute the one-minute moving average of the prices, i.e.,

$$MA_t = \frac{1}{12}\sum_{i=0}^{11} p_{t-i}.$$

We regard p_t as an outlier if

$$\left|\ln\left(\frac{p_t}{MA_t}\right)\right| > 0.01,$$

i.e., if it is 1% away from the one-minute moving average. For the identified outliers, we set their price levels using the most recent observation and set the corresponding trading volume to be zero. Through experiments, we find the two filters identify fewer than 0.3% of the observations as outliers in normal years, and fewer than 1% in 2008 and 2009. For robustness checks, we find our results are not impacted when using other thresholds for the two filters.

2.4.2 Implementation Details

In all our empirical studies, we set the length of estimation interval to be 30 minutes. This choice of time interval balances two considerations. First, to estimate the intraday patterns, we prefer short estimation interval to enhance granularity. Second, there need to be enough observations in each time interval to obtain reliable estimates. We regard 30 minutes as a good balance between the two. With a length of 30 minutes, each interval contains 360 price and volume observations sampled every five seconds. Besides, we apply a moving step of five minutes to obtain smooth intraday patterns. Consequently, there are in total 73 time intervals for each day, corresponding to 9:30 to 10:00, 9:35 to 10:05, ..., and 15:30 to 16:00.

For all two-scale based estimators in Section 2.3, we set the fast and slow scales to be

$$J = 2 \text{ and } K = 12.$$

As the prices are sampled every five seconds, the fast and slow scales correspond to the returns over ten seconds and one minute respectively. It has been shown empirically that

the two scale estimators are robust to the specific choice of time scales (Aït-Sahalia et al., 2011).[4]

To obtain the intraday patterns, we average the estimates from individual stocks and trading days in each year. Before computing the average, we winsorize the individual estimates between their first and 99th percentiles to mitigate the impact from outliers. Besides, in very rare cases, the estimated realized variance can be negative and the realized correlation can fall outside of $[-1, 1]$. Such anomalies are probably due to large price jumps. When the estimated realized variance is negative, we replace it with the average of the estimates in that day, and we truncate the realized correlation to $[-1, 1]$. The main results are not impacted when we use different thresholds for winsorizing or discarding the anomalies from the estimates entirely.

2.5 Empirical Results of Intraday Stock Patterns

In this section, we provide the empirical results of intraday stock patterns estimated from our high-frequency data, including realized correlation, beta, trading volume, and volatility.

2.5.1 Intraday Realized Correlation Between Stock Pairs

In this section, we report the estimation results for intraday realized correlation between stock pairs. While there has been some literature on the comovement in the trading volume of different stocks (see, e.g., Karolyi et al. (2012), Koch et al. (2016), and Min et al. (2018)), the intraday correlation of stock returns is much less studied. We shed light on this topic by estimating realized correlation from a large high-frequency dataset with robust estimators. Our results reveal that the intraday realized correlation indeed shows specific patterns that change over years. The findings support our statement in hypothesis H1 on the implication of index-based investment.

[4]As an additional robustness check, we select several examples and compare the estimated realized variances by the TSRV method with that by the parametric MLE method in Ait-Sahalia et al. (2005). The results match well in most cases. Note that the MLE method requires separate optimization in each estimation. Thus it can not be practically implemented on our large dataset.

The estimation proceeds as follows. Using the pairwise estimator (2.4), we estimate the realized correlation for each stock pair in the S&P 500 Index. Then, the most convenient way to obtain the general intraday pattern is to average across all stock pairs. However, the correlation between two stocks can be very different across pairs. For example, correlation may be positive or negative depending on the fundamental similarity of the two stocks. Such variation can have substantial impact on the intraday realized correlation as well. To capture the potential heterogeneity in realized correlation, we divide stock pairs into bins based on the correlation of their daily returns, and compute the average realized correlation for each bin separately.

We construct twelve bins of stock pairs as follows. For each year, we pick those stocks that are in the S&P 500 Index for the entire year. This leaves us with a set of 475 stocks on average in each year[5]. For each stock pair constructed from this set, we compute its daily correlation using the daily returns of the two stocks in the year. We then divide the stock pairs into different bins by the levels of their daily correlations, which can be regarded as a measurement of the fundamental similarity between the two stocks. Denote by p_α the α-th percentile of the daily correlations across all stock pairs in the given year. The first three bins include stock pairs with daily correlations within $[p_0, p_1]$, $(p_1, p_5]$, and $(p_5, p_{10}]$, respectively, i.e., the stock pairs with the bottom 10% daily correlations. The other nine bins contain the stock pairs with daily correlations within $(p_{10n}, p_{10(n+1)}]$ for $n = 1, 2, ..., 9$. As most stocks in the S&P 500 Index are positively correlated, we use a more granular partition via the first three bins for the stock pairs with low, potentially negative, daily correlations. The average daily correlation for each bin is reported in Table B.3 in Appendix C.2.3.

For each pair bin denoted by B_j $(j = 1, 2, ..., 12)$, we compute its realized correlation in time interval t as the equal weight average of all the stock pairs and trading days, i.e.,

$$\overline{\text{RCorr}}_{jt} = \frac{1}{N \times |B_j|} \sum_{d=1}^{N} \sum_{(i_1, i_2) \in B_j} \text{RCorr}_{dt}^{i_1, i_2},$$

where $\text{RCorr}_{dt}^{i_1, i_2}$ denotes the realized correlation between stocks i_1 and i_2 in time interval t of day d; $|B_j|$ is the number of pairs in bin j and N is the number of trading days. Note that even the smallest bin contains a large number of stock pairs. For example, with 475

[5]The numbers of such stocks for each year are reported in Table B.2 in Appendix C.2.3.

stocks in a year, we would have in total $475 \times 474/2 = 112,575$ stock pairs. Thus even the smallest bin (below the first percentile) includes over 1,000 stock pairs, which translates to more than $1,000 \times 250 = 250,000$ samples every year for a given time interval t. Such large sample size improves the robustness of the estimated intraday pattern.

The results for intraday realized correlation of different bins are shown in Figure 2.4. Each panel represents a given year between 2004 and 2018. The horizontal axis corresponds to the trading hours in a day, where the first (resp. last) point represents the half hour time interval 9:30 – 10:00 (resp. 15:30 – 16:00). For each year, we plot the estimated intraday realized correlation for six selected bins: the first three bins for low-correlated pairs ($[p_0, p_1]$, $(p_1, p_5]$, and $(p_5, p_{10}]$), the bin with median correlation level ($(p_{40}, p_{50}]$), and the two bins for high-correlated pairs ($(p_{80}, p_{90}]$ and $(p_{90}, p_{100}]$). The six bins with daily correlation from high to low are represented by the red, purple, orange, green, cyan, and dark blue lines, respectively. The results for other bins are qualitatively similar. The standard deviations for the intraday curves are also estimated. They are generally very small thanks to the large sample size. Indeed, the standard deviation for the estimates in Figure 2.4 is smaller than 0.006 in all cases[6].

By Figure 2.4, we have two direct observations for the intraday realized correlation. First, in most cases, the realized correlation is positive. This is not surprising as most stocks in the S&P 500 Index are positively correlated. Besides, the relative ranking in daily correlation is mostly preserved in realized correlation. That is, the bin with higher daily correlation also has higher realized correlation in a given time interval.

We then take a closer look at the intraday pattern of realized correlation. Comparing the intraday curves for different bins and over years, we can see the realized correlation indeed demonstrates specific patterns that change over time. We summarize our findings in the following three points.

First, in the period 2004 – 2007 (the first four panels), the intraday realized correlation generally shows an M-shape pattern for all the six bins: it starts lowest at the market open and increases to a peak around 11:00, stays flat around noon, further increases to the highest

[6]The estimates of standard deviations for these as well as other intraday curves in the chapter are available from authors upon request.

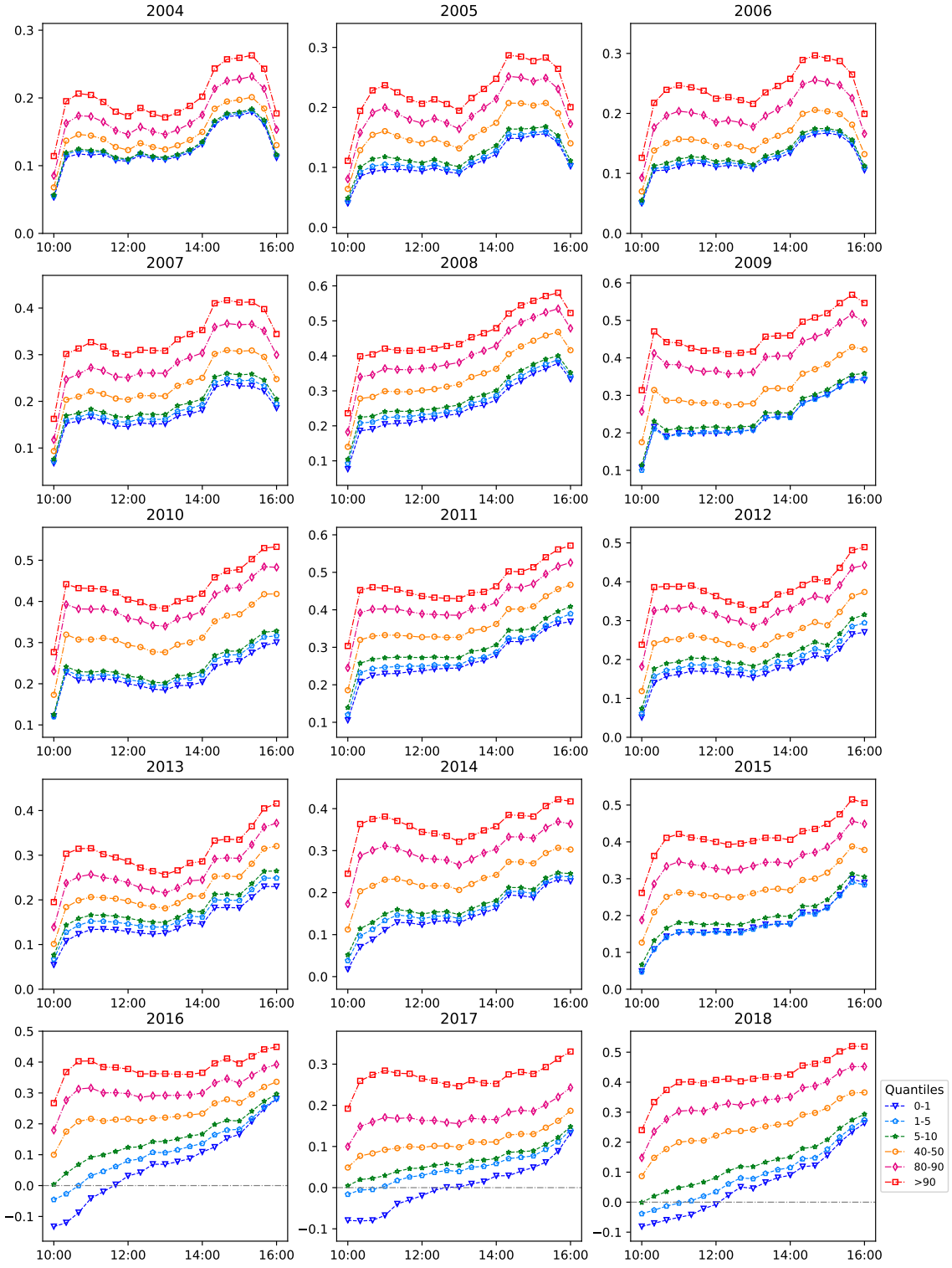


Figure 2.4: Intraday realized correlation for different stock pair bins

level around 15:00, and finally drops near the market close.

Second, after 2009, the drop in realized correlation near the market close vanishes. Instead, the realized correlation increases in the entire afternoon, and reaches the highest level near the market close for all bins. The market open (9:30 to 10:00) still witnesses the lowest level of realized correlation. The curve in the middle of the day (e.g., 11:00 to 14:00) becomes flatter, indicating stable periods for realized correlation. This holds except for the three low correlation bins in 2016 – 2018, which is further discussed below.

Finally, in recent years from 2016 to 2018, the intraday pattern for the low correlation bins (green, cyan, and blue lines) becomes different from that in previous years as well as that for the high correlation bins. For these three bins, their realized correlation at the market open further decreases, even to the negative regime, which is not seen in previous years. Besides, the middle part of their intraday curves become steeper, showing their realized correlation increases rapidly during the day. This change is not seen in the intraday pattern for the high correlation bins, which still stays flat during the middle of the day. For all bins, the realized correlation reaches the highest level near the market close. This suggests the stock prices are more likely to move in the same direction at the end of trading session.

The above observations are illustrated more directly in Figure 2.5, where we plot the intraday realized correlation for the stock pairs with the top 10% (left) and bottom 10% (right) daily correlations. Each line in the panel represents the average over different years in a given period (2004 – 2006, 2007 – 2010, 2011 – 2015, and 2016 – 2018). In 2004 – 2006, the realized correlation for both high and low correlated pairs show an M-shape (purple lines). After that, the realized correlation significantly increases near the market close. Moreover, in 2016 – 2018, the realized correlation for the bottom 10% pairs starts below zero at the market open, and monotonically increases during the day to above 0.2 at the end of trading session (red line in the right panel).

The intraday patterns of realized correlation in the recent decade confirms our hypothesis H1 in Section 2.2.3, i.e., correlation is low in the morning and high near the market close[7].

[7]The difference between the first and last half hour intervals is statistically significant at 0.1% level.

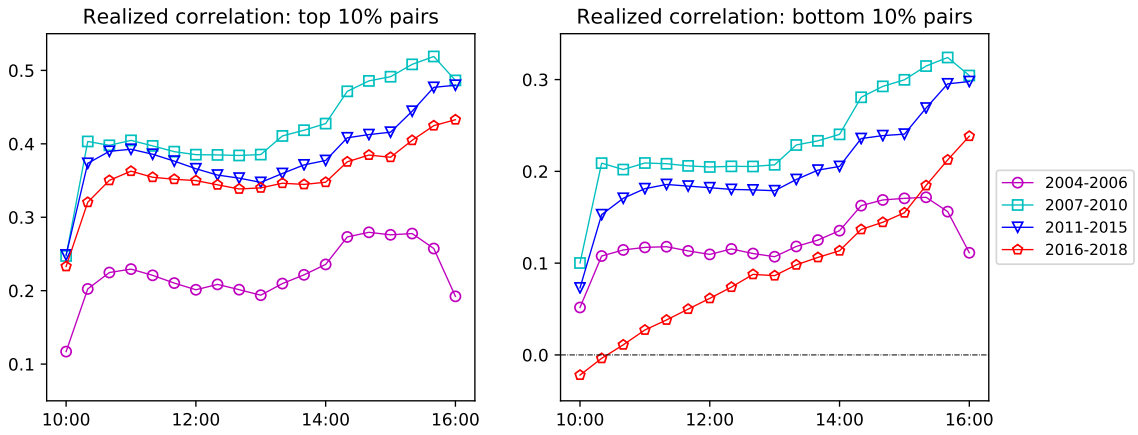


Figure 2.5: Realized correlation for top and bottom stock pairs

Note: Realized correlation for the stock pairs with daily correlations above the 90th percentile (left) and below the 10th percentile (right), averaged over the yearly results in different periods.

This demonstrates the implication of the growth of index-based strategies and the active-open, passive-close trading profile. In particular, more discretionary (resp. index-based) trading tends to decrease (resp. increase) the correlation between different stocks at the market open (resp. close). Besides this hypothesis, our estimation results reveal other significant pattern shifts for intraday realized correlation in recent years, especially the even lower starting level and monotonically increasing shape for the low correlated pairs.

In the recent work of Buccheri et al. (2020), they propose a score-driven model to estimate the covariance dynamics under high-frequency setting. With a much smaller dataset (transaction data of ten stocks in 2014), they show the opening hours are dominated by idiosyncratic risk and a common market factor emerges in the afternoon. This result is consistent with the intraday patterns we obtained for realized correlation. Moreover, in Section 2.6, we develop a market impact model with time-varying liquidity provision from different types of investors, and use it to show the active-open, passive-close trading profile indeed generates the intraday correlation pattern that is qualitatively similar to the one observed in recent years. This provides additional theoretical support for our interpretation.

The intraday pattern of realized correlation has various applications, especially in intraday trading and portfolio execution. For example, if the traders want to exploit the low

(resp. high) correlation in the intraday price movement of different stocks, they may set up their positions around the market open (resp. market close). On the other hand, if the portfolio managers prefer a period with stable correlation for order execution, the middle of the day appears as a better choice. These applications, among others, can be topics for future research.

2.5.2 Intraday Realized Correlation Between Sectors

In this section, we estimate the intraday realized correlation between different sectors. While we can compute the average between-sector correlation using the pairwise correlations of corresponding stocks, here we employ the portfolio-implied approach developed in Section 2.3.2.2. This allows us to estimate the realized correlation from the portfolio level that incorporates different weights of stocks. We reveal that the intraday realized correlation between sectors demonstrates similar patterns to that for stock pairs observed in Section 2.5.1.

We implement the portfolio-implied estimator in (2.6) as follows. As for pairwise correlations, we choose the stocks that are in the S&P 500 Index for the entire year, and divide them to eleven mutually exclusive sectors based on the first two digits of their GICS codes. The mapping from GICS codes to sector names and the numbers of stocks in each sector are summarized in Tables B.4 and B.5 in Appendix C.2.3. We exclude the real estate sector as it is not formally included as a sector before 2016, and the number of stocks in this sector is small in early years. We set the weight of each stock proportional to its average market capitalization throughout the year. The sector and market capitalization data are obtained from CRSP.

With ten remaining sectors, there are in total 45 sector pairs. For each year, we focus on the sector pairs with high and low average daily correlations[8]. Specifically, we report the estimation results for six sector pairs, three with the highest average daily correlation and three with the lowest. The sector names for the six selected pairs in each year are reported in Table B.6 in Appendix C.2.3. The results for between-sector intraday realized

[8]The average daily correlation between two sectors are simply computed as the average of the pairwise daily return correlations between their constituents.

correlation are shown in Figure 2.6. The six sector pairs, with average daily correlations from high to low, are represented by the red, purple, orange, green, cyan, and dark blue lines respectively. At first glance, we see the ranking of sector pairs is generally preserved in intraday realized correlation: the three pairs with higher daily correlations also have higher realized correlation than the other three for most of the time during the day.

A more interesting finding is revealed by comparing the intraday patterns in Figures 2.4 and 2.6. Recall Figure 2.4 plots the intraday realized correlation for stock pairs estimated by the pairwise estimator (2.4). Thus the realized correlation in the two figures are very different with respect to both the estimated object and the estimation method. Surprisingly, however, we see the patterns in the two figures are similar, especially in the recent decade. Thus the discussion in previous section regarding the shapes of intraday patterns generally applies here. In the period 2004 – 2007, the intraday realized correlation demonstrates an upward M-shape for most sector pairs. After 2009, the realized correlation starts low in the morning, stays relatively flat in the middle of the day, and increases significantly near the market close. Moreover, in 2016 to 2018, the realized correlation of the three low correlated sector pairs starts negative at the market open and increases quickly during the day, which is similar to the patterns of the three low correlation bins in Figure 2.4.

The consistent results for sector pairs highlight the generality and robustness of the intraday correlation patterns observed in Figures 2.4 and 2.6. It demonstrates, from the portfolio-level, the implication of the growth of index-based strategies and the active-open, passive-close trading profile. Specifically, concentrated trading from index-based strategies drives up the correlation between different sectors near the market close, while more discretionary trading tends to lower the correlation in the morning. Such patterns have important applications in sector-based intraday trading and portfolio execution, which are deferred to future research.

2.5.3 Intraday Pattern of Realized Beta

In this section, we analyze the estimation results for intraday realized beta of S&P 500 constituents. We find that in recent years, the realized betas of different stocks start dispersed at the market open, but generally move towards one near the market close. This

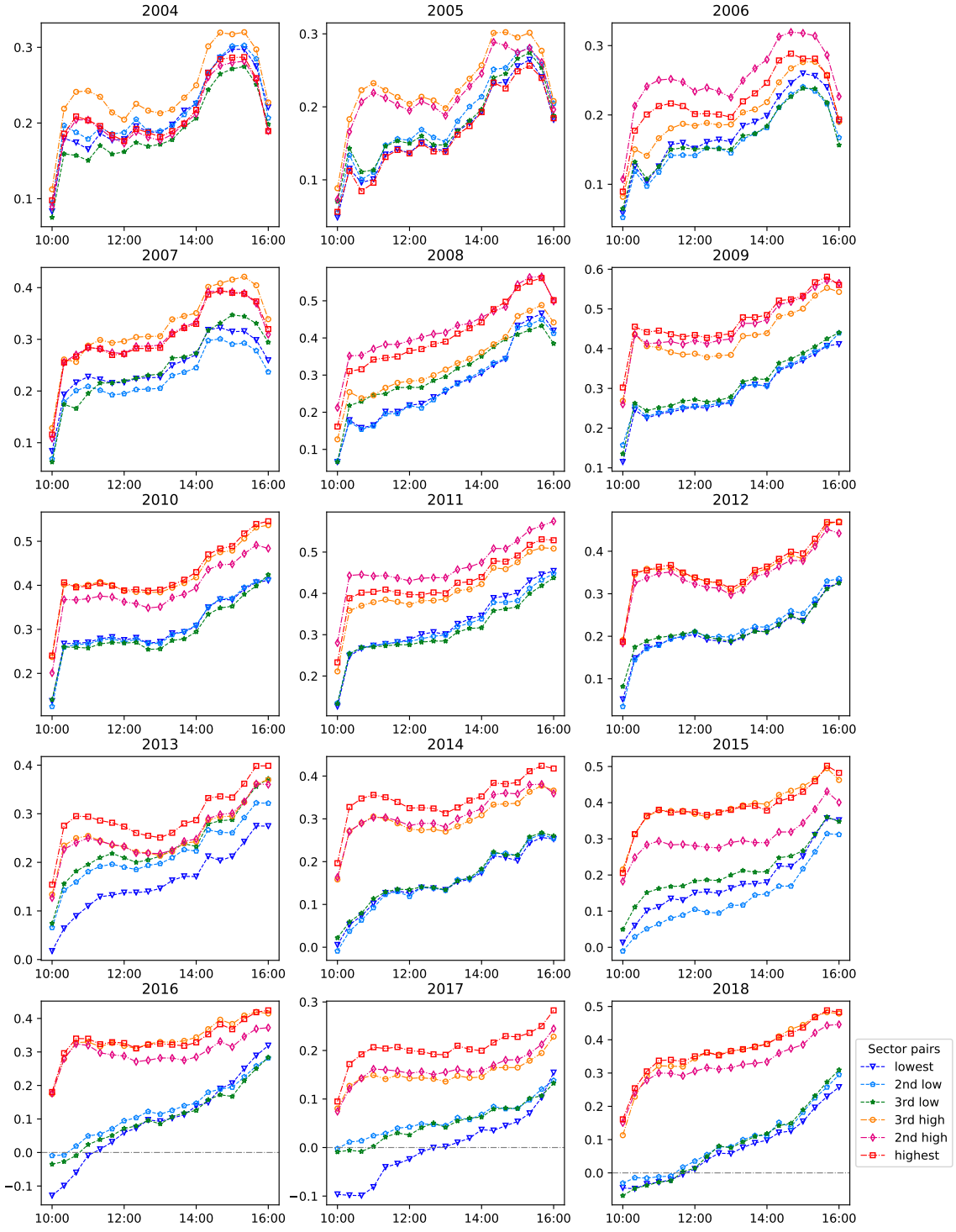


Figure 2.6: Intraday realized correlation for sector pairs

confirms our second hypothesis H2 in Section 2.2.3. The convergence pattern of realized beta echos our results for realized correlation discussed in previous sections, and shows the impact of the index-based strategies and the active-open, passive-close trading profile.

We estimate intraday realized beta using the estimator (2.8) as follows. Similar to the study of pairwise realized correlation, we select the stocks that are in the S&P 500 Index for the entire year, and divide stocks into bins based on their daily betas, which are computed by (2.7) using daily returns. We construct eleven bins for each year. Denote by p_α the α-th percentile of daily betas among all stocks. The first two bins include the stocks with daily betas in $[p_0, p_5]$ and $(p_5, p_{10}]$, i.e., stocks with the bottom 10% daily betas. The other nine bins consist of the stocks with daily betas in $(p_{10n}, p_{10(n+1)}]$ for $n = 1, 2, ..., 9$. The average daily beta for each bin is reported in Table B.7 in Appendix C.2.3. For each bin B_j $(j = 1, 2, ..., 11)$, we estimate its realized beta over time interval t as

$$\overline{\text{RBeta}}_{jt} = \frac{1}{N \times |B_j|} \sum_{d=1}^{N} \sum_{i \in B_j} \text{RBeta}^i_{dt},$$

where RBeta^i_{dt} denotes the realized beta of stock i in time interval t of day d; N and $|B_j|$ are the number of trading days and number of stocks in bin j, respectively. With 475 stocks in a year, the smallest bin would have 24 stocks, which translates to approximately $250 \times 24 = 6,000$ samples every year for a given time interval t.

The estimation results of intraday realized beta are shown in Figure 2.7. For each year, we plot the estimated results for six selected bins: the first three bins with low daily betas ($[p_0, p_5]$, $(p_5, p_{10}]$, and $(p_{10}, p_{20}]$), and the three with high daily betas ($(p_{70}, p_{80}]$, $(p_{80}, p_{90}]$, and $(p_{90}, p_{100}]$). The six bins with daily betas from high to low are represented by the red, purple, orange, green, cyan, and dark blue lines, respectively. The horizontal dashed black line denotes the level of beta equal to one. The standard deviations of the estimated intraday curves in Figure 2.7 are below 0.015 in all cases, which are relatively small compared with the estimated levels.

By Figure 2.7, we have the following observations of the intraday pattern of realized beta. First, the ranking of daily beta across different bins is mostly preserved in the intraday pattern: the bin with higher daily beta also has higher realized beta across the day. Next, we see the intraday patterns of realized beta indeed show specific shapes that evolve over

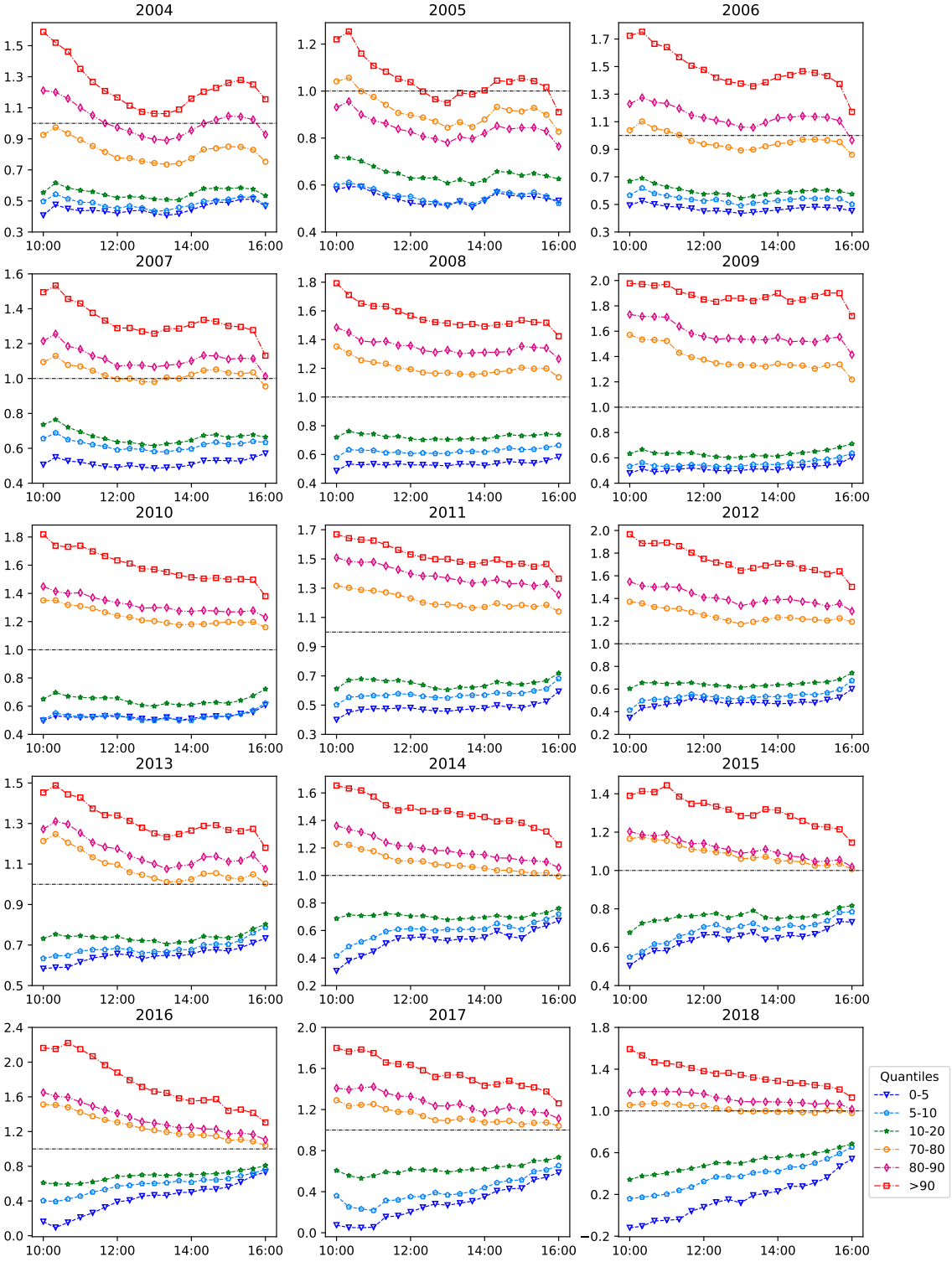


Figure 2.7: Intraday realized beta for different stock bins

time. For the high beta bins (red, pink, and yellow lines), their intraday realized beta exhibits a smirk pattern in 2004 – 2007, and a monotonically decreasing pattern in the years after. For the low beta bins (green, cyan, and dark blue lines), their intraday realized beta stays relatively flat before 2013, but significantly increases during the day in the years after. Consequently, in the recent decade, we see the realized betas of different bins are more dispersed in the morning, but generally move towards one near the market close. This confirms our second hypothesis H2 in Section 2.2.3, and shows the implication of the active-open, passive-close trading profile on intraday beta.

To demonstrate the changes in the intraday patterns more directly, we plot in Figure 2.8 the realized beta for the stocks with top (left) and bottom (right) 10% daily betas, averaged over the years in four different periods. By Figure 2.8, we see significant changes in the intraday patterns in recent years (2016 – 2018), as shown by the red lines in the two panels. First, at the market open, the realized betas of the high beta and low beta stocks become more dispersed than previous years, as shown by the even higher (resp. lower) red line in the left (resp. right) panel. On the other hand, the magnitude of the intraday movement is larger, i.e., the realized beta drops (resp. increases) more during the day for the high (resp. low) beta stocks. Such changes are especially noticeable for the low beta stocks: their average realized beta starts below 0.2 at the market open, but rises dramatically to above 0.6 at the end of trading session.

As a consequence of above changes, we see a more significant convergence pattern of intraday realized beta in recent years. Specifically, the divergence in realized betas of different stocks shrinks during the day. At the end of trading session, the realized betas of all bins move towards one, suggesting the individual stock returns are more similar to the market return. This can be attributed to the growth of index-based investment, and the active-open, passive-close trading profile. In particular, more discretionary trading in the morning make realized betas more dispersed across stocks, while more index-based strategies drive realized betas towards one near the market close. In Section 2.6, we provide theoretical support for such interpretation using a market impact model with time-varying liquidity provision from single-stock and index-fund investors. The intraday pattern of realized beta has potential applications for intraday trading strategies that exploit the levels of systematic

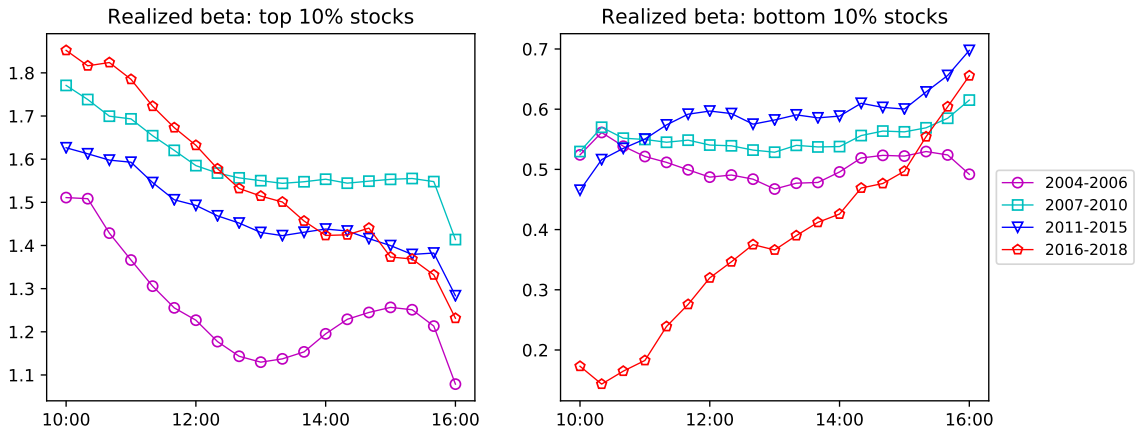


Figure 2.8: Realized beta for stocks with top and bottom daily betas

Realized beta for the stocks with daily beta above the 90th percentile (left) and below the 10th percentile (right), averaged over the yearly results in different periods.

risk in stocks' returns. For example, it may be better to execute strategies that hinge on the heterogeneity in stocks' systematic risk levels in the morning rather than in the afternoon.

2.5.4 Intraday Patterns of Trading Volume and Realized Volatility

In this section, we look into the intraday patterns of trading volume and realized volatility. First, we show in Figure 2.9 the intraday pattern of scaled trading volume defined in (2.1). The results are computed as the equal weight average of all S&P 500 constituents in a given year. The four panels plot the estimation results for 2004 – 2007, 2008 – 2010, 2011 – 2014, and 2015 – 2018 respectively. Comparing across the four panels, we see the intraday pattern of scaled trading volume indeed changes over time. In 2004 – 2007, the scaled trading volume demonstrates a symmetric U-shape pattern that is relatively stable across years. The trading volume near the market close is quite close to that at the market open. However, from 2008, the trading volume near the market close increases dramatically, and the symmetric U-shape pattern becomes skewed to the right. This trend becomes more significant in recent years, as the trading volume near the market close keeps increasing. In 2018, the final half hour 15:30 – 16:00 (the last point) consists of more than 20% of the total trading volume in the day.

The change in the intraday pattern of trading volume can be seen more clearly from the

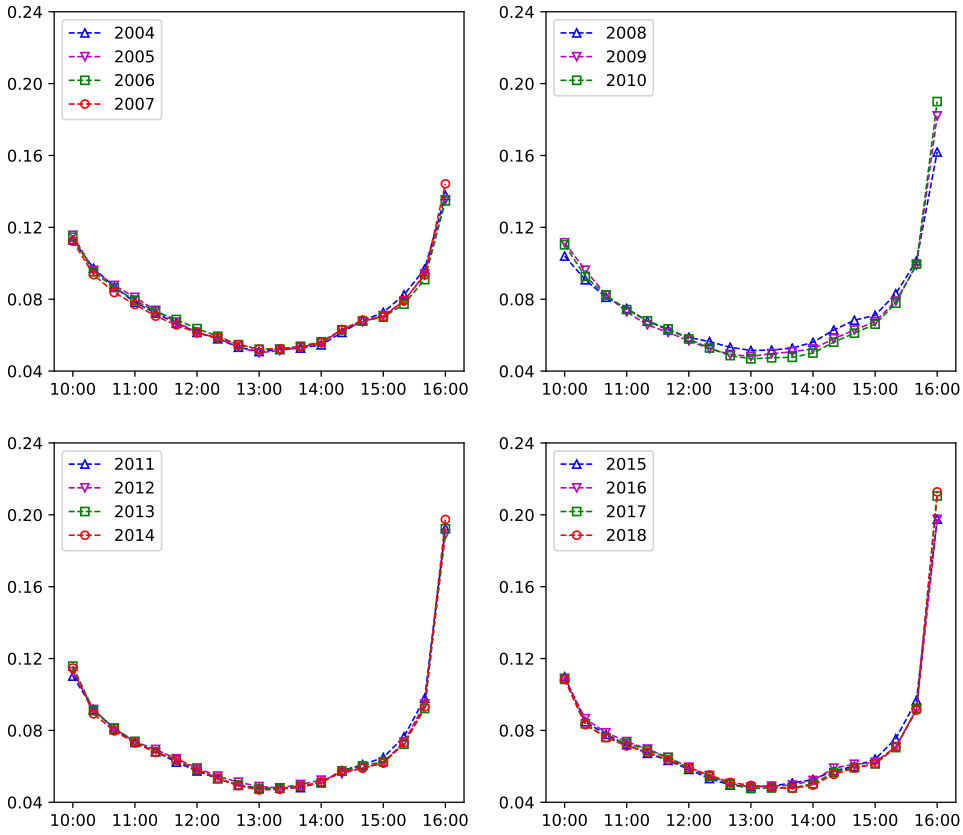


Figure 2.9: Intraday scaled trading volume

left panel of Figure 2.10, where we fix four time intervals (9:30 – 10:00, 11:30 – 12:00, 13:30 – 14:00, 15:30 – 16:00) and plot the scaled trading volume across different years. By the red line, we see the scaled trading volume in 15:30 – 16:00 increases significantly, especially during 2007 to 2010 and after 2016. Such increase is offset by the drop in scaled trading volume in the middle of the day, although the magnitude is much smaller. The results observed here indicate the increase in end-of-day trading volume holds for our large stock universe, thus generalize the finding in Figure 2.2 for the stocks with low and high passive ownership. Such increase can be attributed, in part, to the growth of index-based strategies and their concentrated trading near the market close. For instance, Cushing and Madhavan (2000) and Foucault et al. (2005) point out the importance of closing price for institutional investors, who execute most of the index-based strategies. The recent work in Wu (2019) finds ETF flows make increased usage of market-on-close orders.

Next, motivated by hypothesis H3 in Section 2.2.3, we study the intraday pattern of

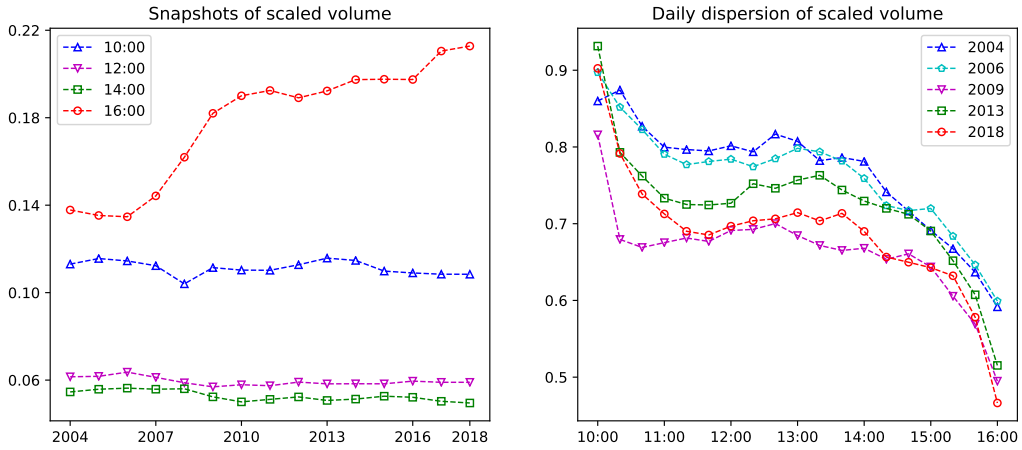


Figure 2.10: Snapshots and daily dispersion of scaled volume

The left panel shows the scaled trading volume in four time intervals. The right panel shows the intraday pattern of daily trading volume dispersion for five selected years.

daily dispersion in trading volume. For stock i, its daily volume dispersion in time interval t is defined as

$$\text{VolmDisp}_{it} = \frac{\text{StdVolm}_{it}}{\text{AvgVolm}_{it}}. \tag{2.9}$$

Here AvgVolm_{it} and StdVolm_{it} denote the mean and standard deviation of the corresponding trading volume across different days, i.e.,

$$\text{AvgVolm}_{it} = \frac{1}{N}\sum_{d=1}^{N} \text{Volm}_{idt},$$

and

$$\text{StdVolm}_{it} = \sqrt{\frac{1}{N-1}\sum_{d=1}^{N} \left(\text{Volm}_{idt} - \text{AvgVolm}_{it}\right)^2},$$

where N is the number of trading days for stock i (e.g., roughly 250 trading days in a year). The normalization by AvgVolm_t in the denominator of (2.9) allows us to average and compare across stocks with very different trading volume levels.

We show the intraday pattern of daily volume dispersion for five selected years (2004, 2006, 2009, 2013, and 2018) in the right panel of Figure 2.10, which are computed as the equal weight average of all the stocks in each year. The results for other years are qualitatively similar and available from authors upon request. We have the following observations for the intraday pattern of daily volume dispersion. First, we see the daily volume dispersion demonstrates a similar intraday pattern in all the five years: starts high at the market

open, drops and stays flat in the middle of the day, and further decreases near the market close. This pattern suggests the trading volume is more volatile across days in the morning, but much less near the market close. The standard deviations of the estimates in Figure 2.10 are smaller than 0.018 in all cases, and the difference between market open and close is statistically significant at 0.1% level for all five years. It confirms our hypothesis H3 on daily volume dispersion and shows the impact of the growth of index-based strategies and the active-open, passive-close trading profile. Specifically, the trading from discretionary investors in the morning varies more across days, while the trading from index-based strategies is more persistent.

Besides the general decreasing pattern of daily volume dispersion, more interesting results are revealed by comparing the intraday curves of the five different years. First, the overall dispersion level in 2009 (purple) is the lowest among the five years. This is likely a consequence of the financial crisis, during which trading volume was consistently high. More importantly, we see the magnitude of the intraday drop from market open to close increases significantly in 2013 and 2018 compared with that in previous years[9]. The increased drop is mainly driven by the lower dispersion levels at the end of trading session. This shift can be attributed to the prevalence of index-based strategies in the recent decade, and particularly, their concentrated trading near market close.

We further support this interpretation by comparing the intraday pattern of daily volume dispersion for the high and low passive ownership bins defined in Section 2.2.1, i.e., stocks with passive ownership below the fifth percentile and above the 95th percentile each year. The left and right panels of Figure 2.11 plot the average of yearly results for the two bins in the first (2004 – 2010) and second periods (2011 – 2018), respectively. By the left panel, we see the daily volume dispersion is quite close for the two bins in 2004 – 2010. However, in 2011 – 2018, the daily volume dispersion for high passive ownership bin becomes significantly lower than that for the low passive one. The increased gap in daily volume dispersion can be explained by the prevalence of passive mutual funds in the recent decade, which leads to larger difference in the degree of passive ownership between the two bins. Indeed, the

[9]The increase in drop magnitude is also observed for other years after 2010.

average difference in passive ownership between the two bins increases from 7% for the first period to 20% for the second period, which can be computed from Table B.1.

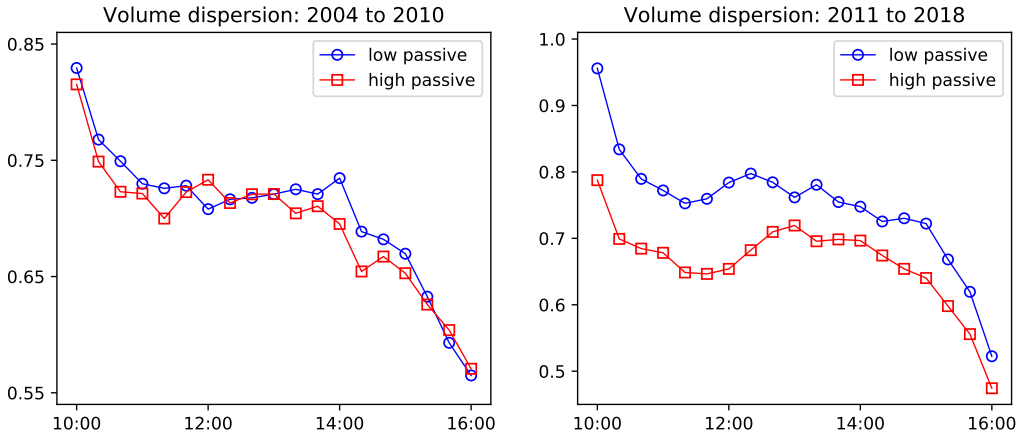


Figure 2.11: Daily volume dispersion by passive ownership bins

Daily volume dispersion of the low and high passive ownership bins, averaged over the first period 2004 – 2010 (left) and the second period 2011 – 2018 (right) respectively.

Finally, we look into the intraday pattern of realized volatility, as we discussed in Hypothesis 4. While the intraday volatility pattern has been widely studied in previous literature, as far as we know, here we use a large dataset with all S&P 500 constituents over 15 years, instead of limiting to the market index or a few selected stocks. The results are shown in Figure 2.12, with each curve computed as the equal weight average of all stocks in a given year. The results are similar when we compute the average using the market-cap weight.

We have following observations for the intraday pattern of realized volatility. First, the intraday volatility shows a U-shape pattern skewed to the left: it starts relatively high in the morning and decreases during the day. In the majority of years, the realized volatility reaches the lowest level around noon, and slightly increases near the market close. Not surprisingly, the financial crisis period (2008 and 2009) is associated with extremely high realized volatility in the entire day, as seen from the magnitude of vertical axis in the upper-right panel. The pattern observed here generally supports our hypothesis H4.

This U-shape pattern of intraday volatility is widely observed in the literature (see, e.g., Wood et al. (1985), Pagano et al. (2008), and Stroud and Johannes (2014)). In some of their results, the spike in volatility near the market close is more significant than that seen in

Figure 2.12. The mild difference in the intraday patterns can be potentially explained by the different samples used in our and other studies: Figure 2.12 shows the equal-weight average volatility of all S&P 500 constituents over each entire year, while other studies mainly focus on the volatility of the market index or several stocks over a shorter horizon (e.g, a month). Finally, we notice that the increase in realized volatility near the market close vanishes after 2013, i.e., the "tail" of the intraday curve tends to flatten in recent years. This may be explained by the concentrated trading from index-based strategies at the end of trading session.

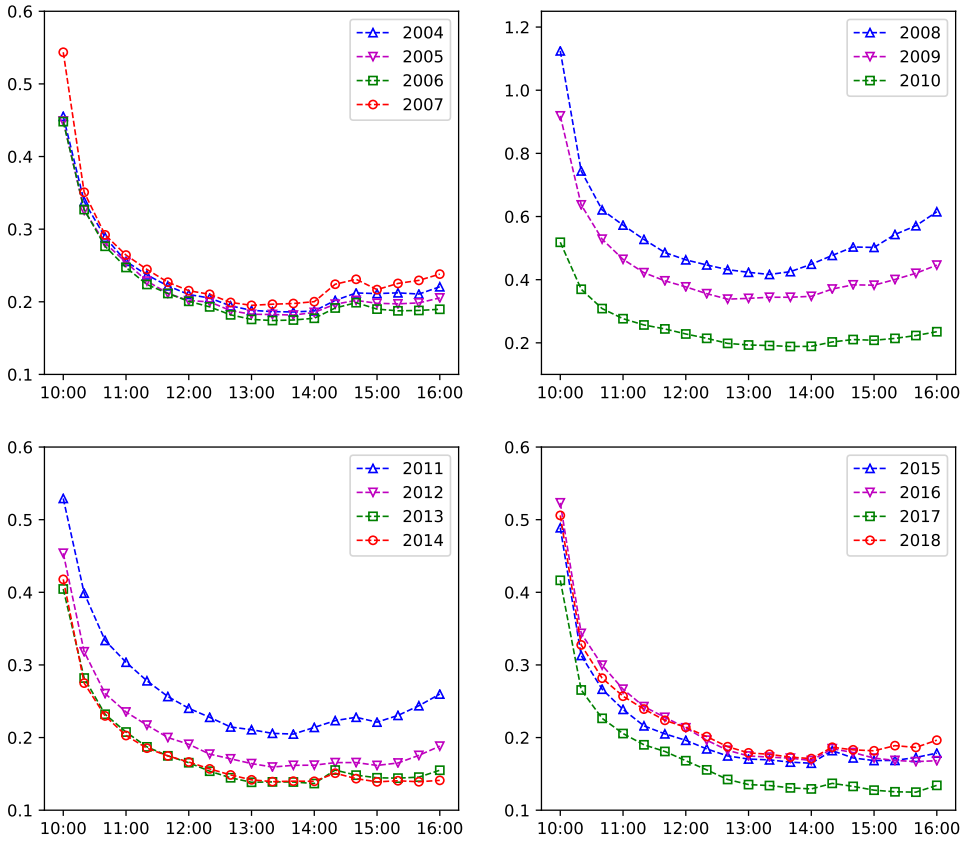


Figure 2.12: Intraday realized volatility (annualized)

2.6 A Market Impact Model with Time-varying Liquidity Provision

In this section, we provide additional theoretical support for the relation between the growth of index-based strategies and the intraday patterns of stock dynamics. In particular, we use a market impact model to show that the active-open, passive-close trading profile indeed generates the intraday pattern of realized correlation and beta observed in Sections 2.5.1 and 2.5.3.

Suppose the market has N individual stocks, indexed by $i = 1, 2, ..., N$, and a fund consisting of these stocks. The fund can be an ETF or a mutual index fund. The weight of each stock in the fund is given by w_i, which is assumed to be positive with $\sum_{i=1}^{N} w_i = 1$. For stock i, denote its actual price and investors' reservation value by p_i and r_i, respectively. We use the vector representations $\mathbf{w} = (w_1, w_2, ..., w_N)^\top$, $\mathbf{r} = (r_1, r_2, ..., r_N)^\top$, and $\mathbf{p} = (p_1, p_2, ..., p_N)^\top$.

There are both single-stock and index-fund investors in the market. The single-stock investors buy or sell individual stocks in response to the change in the gap between the actual price and the reservation value. Specifically, active investors will buy (resp. sell) $\psi^a_{i,t}$ shares if the gap $r_i - p_i$ increases (resp. decreases) by one dollar in time interval t. This linear assumption is often assumed in microstructure literature (Kyle, 1985), and can be justified under the case of CARA utility investors and normally distributed beliefs. In contrast, the index-fund investors only trade the fund based on its price and reservation value implied by the stocks. They will buy (resp. sell) ψ^f_t share of the fund if the gap $\mathbf{w}^\top\mathbf{r} - \mathbf{w}^\top\mathbf{p}$ increases (resp. decreases) by one dollar. Such index-based strategy trades stocks on a portfolio-level, translating to $\mathbf{w}\psi^f$ position change for each stock.

The parameters $\psi^a_{i,t}$ and ψ^f_t measure the liquidity provided by investors, which are allowed to be time-varying during the day. This set-up is similar to the liquidity provision model in Min et al. (2018). The trading from both types of investors impacts the stock prices. For illustration purpose, we use a simple linear function to model the price impact. The linear price impact model is widely used in microstructure literature (see, e.g., Huberman and Stanzl (2004) and Alfonsi et al. (2012)). In particular, we assume every share bought

(resp. sold) of stock i would increase (resp. decrease) its price by ϕ_i, which measures the sensitivity of price to trading demand.

The market evolves as follows. In time interval t, a random shock $\varepsilon_{i,t}$ happens to the reservation value of stock i. Natural sources for the random shocks can be news flow or unpredictable announcements that impact the stock valuation. We assume the shocks are i.i.d. over different time intervals, but allow them to be correlated across stocks. This captures the fact that different stocks may be driven by some common factors (e.g., news for sectors). Denote the vector representation of the shocks by $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, ..., \varepsilon_{N,t})^\top$. After the random shocks, both types of investors trade the stocks (or index fund) based on the new valuation. This leads to a new equilibrium price level that accounts for both the shock in reservation value and the market impact from trading.

Denote the equilibrium price change by an N-dimensional vector $\Delta\mathbf{p}_t$. Then, the change in the gap between price and reservation value is equal to $\varepsilon_t - \Delta\mathbf{p}_t$. The trading demand comes from both single-stock and index-fund investors. First, single-stock investors will buy $\Psi_t^a(\varepsilon_t - \Delta\mathbf{p}_t)$ shares of each stock, where the matrix $\Psi_t^a = \text{diag}_{i=1}^N(\psi_{i,t}^a)$. Next, index-fund investors will buy $\psi_t^f \mathbf{w}^\top(\varepsilon_t - \Delta\mathbf{p}_t)$ shares of the fund, which translates to $\mathbf{w}\psi_t^f \mathbf{w}^\top(\varepsilon_t - \Delta\mathbf{p}_t)$ shares for individual stocks. Summing up the two sources of demand, the market impact model implies

$$\Phi \cdot \left[\Psi_t^a(\varepsilon_t - \Delta\mathbf{p}_t) + \mathbf{w}\psi_t^f \mathbf{w}^\top(\varepsilon_t - \Delta\mathbf{p}_t)\right] = \Delta\mathbf{p}_t,$$

where $\Phi = \text{diag}_{i=1}^N(\phi_i)$ measures the market impact. From this linear equation, we can solve $\Delta\mathbf{p}_t$ as

$$\Delta\mathbf{p}_t = M_t\varepsilon_t, \tag{2.10}$$

which is proportional to the shock ε_t. The matrix M_t is given by

$$M_t = (I_N + \Phi\Gamma_t)^{-1}\Phi\Gamma_t,$$

where I_N is the N-dimensional identity matrix; Γ_t is defined by

$$\Gamma_t = \Psi_t^a + \mathbf{w}\psi_t^f \mathbf{w}^\top.$$

The matrix Γ_t measures the total market liquidity from both single-stock and index-fund investors, which is allowed to be time-varying during the day. As the sum of two symmetric and strictly positive-definite matrices, it is also symmetric and strictly positive-definite.

By (2.10), the covariance matrix of $\Delta \mathbf{p}_t$ can be computed as

$$\Sigma_t^p = M_t \Sigma^r M_t^\top, \tag{2.11}$$

where Σ^r is the covariance matrix of the random shocks ε_t in stocks' reservation values. When there are only single-stock investors, i.e., $\psi_t^f \equiv 0$, it is easy to verify the matrix Γ_t (and M_t) becomes diagonal, and the price change of stock i follows by

$$\Delta p_{i,t} = \frac{\phi_i \psi_{i,t}^a}{1 + \phi_i \psi_{i,t}^a} \varepsilon_{i,t}.$$

In this case, the price change of a stock only depends on the shock to its own reservation value. With deterministic liquidity parameters $\psi_{i,t}^a$, the correlation between price changes equals to that between the corresponding random shocks. As we assume the distribution of random shocks does not change over time, the pairwise correlation between $\Delta p_{i,t}$ is constant even if the liquidity parameters $\psi_{i,t}^a$ are time-varying. However, this does not hold when there are index-fund investors, as positive ψ_t^f leads to non-zero off-diagonal entries in Γ_t (and M_t). By (2.10), the change in the fund price is given by

$$\Delta p_t^f = \mathbf{w}^\top \Delta \mathbf{p}_t = \mathbf{w}^\top M_t \varepsilon_t.$$

We compute the beta of stock i using the price changes as

$$\beta_{i,t} = \frac{\mathrm{Cov}(\Delta p_{i,t}, \Delta p_t^f)}{\mathrm{Var}(\Delta p_t^f)}. \tag{2.12}$$

Accordingly, its average beta over all intraday intervals is defined by

$$\bar{\beta}_i = \frac{1}{T} \sum_{t=1}^{T} \beta_{i,t}. \tag{2.13}$$

We conduct following numerical experiment to show the impact of time-varying liquidity provision on intraday stock dynamics. We model the intraday liquidity provision in line with the active-open, passive-close pattern. We divide the trading hours of each day (9:30

to 16:00) to 78 five-minute intervals, indexed by $t = 1, 2, ..., 78$. We assume the liquidity parameters $\psi^a_{i,t}$ and ψ^f_t vary parametrically as

$$\psi^a_{i,t} = \alpha_t \bar{\psi}^a_i \text{ and } \psi^f_t = \beta_t \bar{\psi}^f,$$

with $\sum_{t=1}^{78} \alpha_t = \sum_{t=1}^{78} \beta_t = 1$ and constants $\bar{\psi}^a_i$ and $\bar{\psi}^f$. The time-varying parameters α_t and β_t determine the liquidity profile of single-stock and index-fund investors throughout the day. For illustration purpose, we consider following simplified liquidity profile:

$$\alpha_t = \begin{cases} 0.0256, \text{ for } t = 1, 2, ..., 6 \\ 0.0105, \text{ for } t = 7, 8, ..., 72 \\ 0.0256, \text{ for } t = 73, 74, ..., 78 \end{cases} \quad \text{and} \quad \beta_t = \begin{cases} 0.0075, \text{ for } t = 1, 2, ..., 72 \\ 0.0769, \text{ for } t = 73, 74, ..., 78 \end{cases}.$$

That is, the liquidity from single-stock investors is higher at the market open (9:30 – 10:00) and near the market close (15:30 – 16:00), while the liquidity from index-fund investors concentrates near the market close (15:30 – 16:00). Thus, there is more discretionary trading in the morning, and the end of trading session is dominated by index-based strategies. This liquidity profile is qualitatively analogous to the result in Min et al. (2018) (Figure 2 therein), which is calibrated from the intraday trading volume data of S&P 500 constituents in 2017.

We assume the market has $N = 478$ stocks, which is the number of stocks that are in the S&P 500 Index throughout the entire 2018. The index fund is benchmarked to the average price of the stocks, i.e., $\mathbf{w} = (1/478, 1/478, \ldots, 1/478)^\top$. For illustration purpose, we assume the random shocks to stocks' reservation values have standard deviation of 1, and are correlated between the stocks following the daily return correlation matrix in 2018. We set the liquidity and market impact parameters as $\bar{\psi}^a_i = 10$, $\bar{\psi}^f = 843.5$, and $\phi_i = 0.9$ for $i = 1, 2, \ldots, 478$. Thus, the index-fund investors provide on average $\bar{\psi}^f / (\bar{\psi}^f + \sum_i \bar{\psi}^a_i) = 15\%$ of total market liquidity, which is consistent with the passive ownership degree in Figure 2.1. Besides, these parameters imply that, if there are only single-stock investors, the equilibrium price will increase by 0.9 dollar for every dollar increase in the reservation value.

Similar to previous studies in this chapter, we employ half-hour intervals with a moving step of five minutes to estimate the intraday curves for correlation and beta. For each half an hour, we compute the average correlation in $\Delta \mathbf{p}_t$ and stock beta $\beta_{i,t}$ over the six 5-minute intervals, which can be explicitly obtained from (2.11) and (2.12). Denote by p^c_α

the α-th percentile for the pairwise correlations between the random shocks $\boldsymbol{\varepsilon}_t$, i.e., daily return correlation in our study. We focus on the intraday correlation pattern for three bins of stock pairs, which have underlying correlations (between corresponding random shocks) below p^c_5, between $[p^c_{45}, p^c_{55}]$, and above p^c_{95} respectively. They correspond to the stock pairs with high, median, and low correlations. Besides, denote by p^b_α the α-th percentile of the average stock beta $\bar{\beta}_i$ given in (2.13), we study the intraday beta pattern for the stocks with $\bar{\beta}_i$ below p^b_5 and above p^b_{95}, i.e., the low and high-beta stocks of the bottom and top 5%.

Figure 2.13 plots the intraday patterns of pairwise correlation in $\Delta\mathbf{p}_t$ (left) and the stock beta $\beta_{t,i}$ (right). By the left panel, we see the correlations indeed change substantially during the day, as a consequence of the index-fund investors and time-varying liquidation provision. At the market open, the correlations are lower for all three bins, and are closer to the underlying levels when there is no index-fund investor. This shows the implication of more discretionary trading from single-stock investors in the morning. On the other hand, the correlation for all bins increase near the market close. The effect is large, and even more significant for the low correlation bin, which increases from around 0 in the morning to above 0.2 at the end of trading session. This increase can be explained by the concentrated trading from the index-fund investors, which drives stocks to move in the same direction via index-based orders. Besides, by two panels on the right, we see the intraday beta starts relatively high (resp. low) for the high- (resp. low-) beta stocks, but drops (resp. increases) during the trading session. This generates the convergence pattern of intraday realized beta observed in Section 2.5.3.

The intraday patterns in Figure 2.13, which are solved explicitly from the model, are in line with the observations for realized correlation and beta in Sections 2.5.1 and 2.5.3, especially in the recent decade: The realized correlation starts low at the market open, stays relatively flat in the middle of the day, and increases significantly near the market close; the realized beta starts dispersed in the morning, but generally moves towards one near the market close. This theoretical study confirms the time-varying liquidity provision, in particular, the active-open, passive-close trading profile, indeed contributes to the observed intraday patterns of realized correlation and beta.

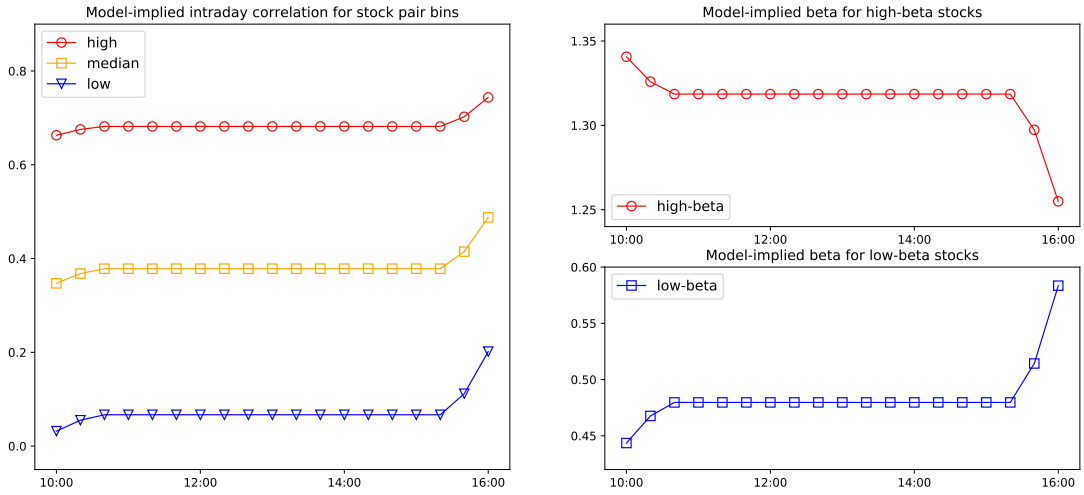


Figure 2.13: Model-implied realized correlation and beta

The left panel plots the model-implied intraday correlation for three pair bins. The right panels plot the intraday beta for high- (upper) and low- (lower) beta stocks.

2.7 Conclusion

The rapid growth of passive investment and index-based strategies has drawn much attention in the recent years. In this study, we demonstrate its implication on various aspects of intraday stock dynamics. In particular, we estimate intraday correlation, beta, volatility, and trading volume with a large high-frequency dataset, i.e., 1-second level trade data for all S&P 500 constituents from 2004 to 2018. We find the intraday patterns indeed change over time. In the recent decade, the realized correlation starts low in the morning and increases near the market close; the realized betas of different stocks start dispersed and generally move towards one at the end of trading session. Besides, we find the trading volume is more volatile across days in the morning than that in the afternoon. These patterns confirm our hypotheses on the implication of index-based strategies, which have become more prevalent in the recent decade.

With the development of financial technologies, high-frequency data has been widely used in various fields of finance. However, estimators under high-frequency setting are often biased by market microstructure noise and observation asynchronicity. Besides, large-scale application of high-frequency data is inevitably hindered by heavy computational burden.

Due to these reasons, most previous studies use a limited dataset (e.g., several stocks or indices over a short period) or employ ad hoc estimators that do not take market noise and asynchronicity into account. In this study, we overcome these challenges with estimators that can be efficiently implemented on large set of stocks and account for both market microstructure noise and observation asynchronicity. Furthermore, the size of the dataset allow us to obtain the general intraday patterns of US stocks and examine how they evolve over time.

The estimation of intraday patterns have various practical applications, including trading strategies, portfolio execution, and risk management. Besides, the intraday patterns facilitate the development of market impact and liquidity provision models that incorporate stylish and realistic features of intraday stock dynamics, e.g., time-varying correlation between stocks. Moreover, with the richness of the dataset, the cross-sectional dimension of the intraday estimates can be leveraged to study the determinants of intraday returns across stocks. Another interesting direction is to examine the intraday patterns for other markets, especially those emerging markets where institutional investors play a less important role. These potential directions can be explored in future research.

Chapter 3

Structural Estimation of Intertemporal Externalities on ICU Admission Decisions

3.1 Introduction

In managing service systems, one often has to consider the short term and long term implications of each decision. This is particularly relevant when the system is operated in a resource constrained environment. Indeed, managers often need to carefully balance between providing immediate access to service in order to address considerations such as minimizing waiting time versus the impact such actions may have on the ability to provide service to future customers. In this work, we study how hospitals balance this trade-off between short-term versus long-term considerations when making Intensive Care Unit (ICU) admission decisions and how this behavior impacts various system performance metrics.

The tension between the short-term and long-term considerations is particularly relevant in health care settings which are often resource constrained and where timely access can have substantial clinical implications. We study this balancing act in the case of ICUs where one could think of the short-term horizon as the next two to six hours versus the longer-term horizon of the next 24 hours. ICUs are specialized inpatient units which provide the highest level of care for the most critically ill patients. They are expensive medical resources, comprising a large portion of operating costs, with the cost of patient care being several times higher than regular wards (see, e.g., Coopersmith et al. (2012); Halpern and Pastores (2015)). Additionally, ICUs often operate at high occupancy levels (see e.g., Halpern and Pastores (2010, 2015)). Consequently, the ICU is often identified as a critical process bottleneck; ICU congestion can have serious repercussions on patient flow and patient outcomes (see, e.g., Kc and Terwiesch (2012); Allon et al. (2013); Kim et al. (2015); Chan et al. (2016)). All

these factors make ICU beds a key resource in a hospital that must be managed effectively.

In this work, we study ICU admissions for patients admitted to the hospital from the emergency department (ED) in a large hospital network. When a patient arrives in the ED, an ED physician will stabilize the patient and assess the patient's needs. If the patient may require ICU admission, an intensivist is called for a consultation. The ultimate decision regarding the patient's disposition requires the coordination of many people including the ED physician, the intensivist, and hospital administrators, as well as various aspects of the hospital system response which could include temporizing actions or adding additional resources (e.g. floating nurses). For conciseness, throughout this manuscript, we will refer to this composite decision maker – comprised of complex interactions of clinicians and the system response – as the 'system' or the 'hospital'. When considering the admission of patients to the ICU, the system must balance the benefit of providing ICU care to the current ED patient versus the potentially negative impact of increased ICU occupancy level as a result of the admission.

As the ICU provides the highest level of care among all hospital units, swift admission generally benefits the patient. However, the admitted patient will occupy an ICU bed, which may restrict access to ICU care for future, perhaps more severe, patients. The system might also take other actions in response to the increased ICU occupancy which might affect the admission of the current patient. For example, the system might increase efforts to stabilize less severe patients to prepare for the increasing needs for ICU beds. Additionally, the system might obtain additional resources such as nursing staff in order to prepare for the increasing occupancy in the ICU, which might also slow down the admission of current patients. In other words, the system's admission response not only impacts current patients, but also changes the system state which affects future patients. This trade-off introduces intertemporal externalities on the ICU admission decisions, i.e., both current and future 'payoffs'[1] matter. While there are undoubtedly many factors which influence the ICU admission decision, our goal in this manuscript is to develop an understanding of how

[1]Throughout the manuscript, we will use the economic term 'payoff' to capture clinical, operational, and (possibly) financial impacts of each decision. This term is not intended to imply there are explicit financial payoffs associated with each decision.

much variation in how hospitals react to ICU congestion could, in part, be explained by intertemporal externalities.

A priori, it is not clear how forward thinking the system is when making ICU admission decisions as there are supporting arguments for responding early as well as late. On one hand, the system's primary goal is to provide the best care for the patient. Thus, if ICU care will benefit the patient, the system should admit the patient as long as there is an available bed in the ICU. This would suggest that the system behaves in a manner which does not respond much in advance to the increasing ICU occupancy. On the other hand, there is empirical evidence indicating that the system may reduce the likelihood of ICU admissions in order to 'save' ICU capacity for future, potentially sicker, patients, or take other actions in advance to prepare for the higher needs of ICU beds in the future. Such behavior is particularly noticeable when the ICU occupancy is high – i.e., when the remaining capacity and resources for future patients is limited. When the ICU is congested, the system is less likely to admit ED patients across all severity levels (Kim et al., 2015), and patients tend to board longer in the ED before being admitted to the ICU (Chan et al., 2016). These empirical findings suggest that the hospital might indeed take into consideration the ability to service future patients and the efforts needed to obtain additional resources in preparation for future admissions. Thus, at least to some extent, the system behaves in a manner that is consistent with forward-thinking and accounts for the long-term when making ICU admission decisions. Given the complexity of the admission process and the various constraints in the system, the system might not be fully aware of how much forward-thinking there is in the admission process. This study will shed light on this debate by utilizing data on which hospital unit patients are admitted to in order to empirically measure the degree of forward-thinking behavior embedded in the system's ICU admission decisions. While there are undoubtedly many factors which influence the flow of patients, we focus primarily on isolating the impact of one of these factors – the degree of forward-thinking behavior.

The degree of the system's forward-thinking behavior is very challenging to empirically quantify directly. First, to the best of our knowledge, there is limited data on the decision-making process itself. Rather, we utilize observational data on the final decision and some system information which could (or could not) influence the decision without explicit in-

formation into the decision-making process. Some crude measures can be constructed from system statistics regarding how the system reacts to ICU occupancy in their admission decisions. For example, if we see a larger drop in admission probability or a greater increase in ED boarding time (i.e. the time spent waiting in the ED once an admission decision has been made) as the ICU becomes congested, this would suggest the system is more active in saving ICU beds or obtaining additional resources for future patients and thus consider the long-term when making decisions. However, such crude measures have several drawbacks. First, the hospital's reaction to ICU occupancy is also affected by various other factors, such as the ED and ICU capacities, the composition of the entire patient cohort treated at the hospital, as well as the arrival rate and average length-of-stay (LOS) of patients. Thus, while such measures can provide some evidence of the hospital's internalization of the short versus longer term, they cannot do so in an accurate and straightforward way. It is also difficult to use these measures to compare behavior across hospitals, as hospitals may differ dramatically in their sizes, workloads, resources, and patient cohorts. More importantly, these indirect measures do not allow us to conduct counterfactual studies to analyze the impact of the forward looking behavior on various hospital performance metrics and patient outcomes.

To handle these difficulties, we take a structural estimation approach to measure the degree to which the system is forward-thinking from observed data and quantify the impact of such behavior on key system performance metrics. In particular, we leverage a new econometric approach on a large retrospective data set to estimate the inter-temporal discount factor in a dynamic discrete choice model for ICU admission decisions. In this model, the hospital may consider the longer-term system dynamics and performance and take relevant actions in advance in order to maximize their expected accumulative utility. Their actions impact the current period utility, but also affect future utility through the transitions of the system states. The discount factor denotes the relative weight of the next period's utility in the hospital's objective function, and, thus, determines how they balance current and future payoffs when making admission decisions. We want to emphasize that our interpretation of the discount factor in the model includes the following forward-thinking behaviors that may slow down admissions when ICU occupancy is high: 1) actively saving beds for fu-

ture patients; 2) stabilizing less severe patients in the ED; 3) obtaining additional resources for ICU admission. We consider all of these possible behaviors as forward or longer-term thinking which are captured by the discount factor in the model. We acknowledge that the system might not be fully aware of their forward-thinking behavior or be able to choose the level of forward-thinking given the complexity of the admission process and various constraints it faces. In other words, the discount factor in our setting should be interpreted as the 'perceived discount factor' that describes the observed behaviors of hospitals. For conciseness and consistency with the rest of the literature, we use the term 'discount factor' and 'perceived discount factor' interchangeably throughout the chapter. A larger perceived discount factor implies that the system is more forward-thinking, while a smaller perceived discount factor implies they are less forward-thinking, and care more about the short-term.

We estimate the perceived discount factor and the costs associated with the system's actions jointly from the data. This is in stark contrast to the majority of empirical studies with dynamic models in the literature which assume the discount factor to be known and then estimate the cost parameters (see, e.g., Rust (1987), Bajari et al. (2007), and Mehta et al. (2017)). Moreover, the discount factor is generally set at a large level that is close to 1 – e.g., 0.90 or even 0.99. In other words, these works make an implicit assumption that the agent is relatively forward-looking. However, this assumption lacks empirical support and formal justifications. More importantly, the discount factors in dynamic models can vary dramatically depending on the context of the problem, or as a consequence of the behavioral variation of the decision makers. In the context of ICU admissions, using a prespecified perceived discount factor can be inappropriate, as there are conflicting arguments for both the immediate and longer-term aspects of the hospital's behaviors. The near versus long term behavior captured by the discount factor may vary across hospitals as well. To address these issues, we identify the discount factor from observed data instead of anchoring our analysis with an assumed, fixed perceived discount factor. To the best of our knowledge, we are the first to identify the discount factor in the healthcare setting.

The dynamic model cannot be identified from choice data without further restriction (see e.g., Manski (1993), Rust (1994), Magnac and Thesmar (2002)), which is a primary reason that most empirical studies assume a known discount factor. The non-identification of

dynamic models stems from the existence of observationally equivalent structures: multiple combinations of discount factors and costs can lead to the same choice probabilities for all states, making the discount factor and costs not jointly identifiable (Rust, 1994). We circumvent this difficulty by adapting the new econometric approach developed by Komarova et al. (2018) and relaxing an assumption on the action set in each period. We note that, much like other econometric methodologies, including the instrumental variable approach, applying the theoretical result in Komarova et al. (2018) to any empirical setting is not straightforward in practice. We need to appropriately adapt their result to our setting and show that our dynamic discrete choice model can be identified given the variation in our data, i.e., the discount factor and costs can be jointly estimated. Estimating the discount factor remains a challenging problem in the literature – it has only been done in two other recently published manuscripts in very different settings than ours (e.g. De Groote and Verboven (2019); Ching and Osborne (2020))). In both cases, they had to leverage special features in the data based on their respective empirical settings to achieve identification.

With the estimated structural model, we quantify the impact of the degree to which each hospital balances the short-term versus long-term considerations on key performance metrics of the ICU. This behavior can have a multifaceted impact on patient and system outcomes. A larger discount factor (longer-term focused behavior) can reduce ICU congestion by saving beds in advance, but this may also lead to longer boarding times for ED patients. Through counterfactual analyses, we explore the potential impact on hospital flow and patient outcomes if it were possible to increase the hospital's discount factor only.

From a broader perspective, our study builds an understanding of how human servers make decisions in a resource limited environment. We focus on the particular aspect of how servers internalize the trade-off between near versus longer term considerations. Our model can be viewed as a finite buffer queueing system. ED patients are admitted to the ICU or medical-surgical ward according to the hospital's decisions. Consequently, the transition of the system state is determined by random arrivals and departures as well as the behaviors of the hospital. In standard queueing systems, customers enter service immediately as long as there are available servers. However, in our setting, there are limited clear and/or objective criteria for which patients to admit to the ICU, as well as when to admit them. These

decisions inevitably depend on a patient's severity level and the system state (e.g. the availability of a bed), but also hinge on the behaviors of the hospital. In this work, we focus on how the hospital internalizes the intertemporal externalities on ICU admissions. This introduces a behavioral perspective to the queueing system which we will see has a substantial impact on system dynamics. Our main contributions can be summarized as follows:

- We build a new dynamic structural model of the system's ICU admission decisions that incorporates both consideration of the current patients as well as the decision's impact on the system's capacity to serve future patients. The structural model accounts for the observed patient severity level, unobserved patient characteristics that are only available to the hospital, random arrivals and departures, as well as the system capacity constraints. We measure the hospital's degree of near versus longer-term looking behavior by the discount factor in the model.

- We adapt a new methodological approach from the econometrics literature and we demonstrate that we can jointly estimate the discount factor and cost parameters in our data setting. We estimate the structural model with an extensive data set consisting of more than 300,000 hospitalizations from 21 Kaiser Permanente hospitals. We find there is large heterogeneity in the estimated discount factors across hospitals – i.e., some account more for the near-term, while others consider more longer-term impacts. In contrast to the standard approach in the literature, in the context of ICU admission decisions, it is inappropriate to assume a prespecified level for the discount factor without empirical support.

- With our estimated structural model, we perform counterfactual studies to evaluate the impact of the hospital's behavior on ICU performance metrics. We first consider operational interventions – adding one ICU bed or decreasing the external arrivals to ICU. These interventions naturally reduce ICU congestion, but can have substantial financial ramifications for a hospital. Next, we consider a behavioral intervention of increasing the discount factor from its current estimated level to 0.9 – i.e., changing the hospital's behavior so they use longer-term discounting. We show that for some

hospitals, the behavioral change of increasing the discount factor can lead to reductions in ICU congestion that are comparable to the costly act of adding one ICU bed or decreasing ICU external arrivals by 5%. We also discuss practical ways for increasing discount factor and propose a simple heuristic policy that can achieve most of the benefits. These analyses highlight the importance of understanding the behavior of the hospital in ICU capacity management.

The rest of the chapter is organized as follows. We conclude this section with a brief literature review. Section 3.2 describes the setting and data. Section 3.3 provides descriptive evidence for the discounting behaviors of hospitals in their ICU admission decisions. Section 3.4 develops the main structural model to measure the degree of near versus longer term looking behavior of the system, and establishes the identification results and algorithmic approach for estimation. Section 3.5 provides the estimation results. Section 3.6 conducts counterfactual studies and proposes a heuristic policy for admission decisions. Section 4.6 concludes this chapter and discusses future research directions.

3.1.1 Literature Review

Our work is related to four main streams of literature: (1) empirical healthcare operations management, particularly those related to ICU decisions; (2) structural estimation in operations management; (3) behavioral operations; and (4) econometrics tools for identifying dynamic models.

There has been a growing literature in the field of empirical healthcare management that examines patient flow in hospitals. A number of works study ICU and non-ICU admission decisions as we do in this work. For instance, Shmueli et al. (2003); Edbrooke et al. (2011) and Kim et al. (2015) study the impact of ICU admissions on patient outcomes including mortality, hospital length of stay (LOS), readmission rate, and patient transfers to higher levels of care. Patients who are not admitted to the unit of choice are typically rerouted to alternative units or even different levels of care. Song et al. (2019) and Dong et al. (2018) study off-placement of patients when bed availability in the primary unit is limited. Such off-placement has important clinical and operational implications as it can result in longer

LOS. Dong et al. (2018) find that by carefully coordinating admissions within the internal hospital network, ED boarding can be reduced. Indeed, patients waiting for access to care is highly undesirable. Chan et al. (2016) find that delays in ICU admission can increase ICU LOS, which, in turn, can create more congestion in an already busy unit. In the ED, another consequence of long waits is an increase in the likelihood of patients leaving without being seen (Batt et al., 2019). In contrast to this body of work which primarily focuses on the impact of admission decisions and waiting on outcomes, we focus on measuring the degree of forward-looking behavior in admission decisions and quantify its impact on hospitals and patients.

Clearly, bed availability, or conversely congestion, has substantial impacts on access to care and whether a patient is admitted, is rerouted, or waits. Indeed, there is substantial evidence the congestion influences patient flow. For instance, it can impact who is admitted to the ICU (Kim et al., 2015) and when patients are discharged (Kc and Terwiesch, 2012). These works demonstrate that clinicians may alter their actions based on congestion, potentially to the detriment of patient outcomes. There is evidence this impact of congestion on clinical behaviors also arises in the ED. Batt and Terwiesch (2016) find that when the ED becomes congested, certain test are initiated at triage in hopes of improving flow in the congested department.

Our work is also closely related to the literature on structural estimation in operations management. Structural models have been widely used in different fields of operations management including supply chains (Bray et al., 2019) and, more closely related to our work, service operations (Li et al., 2014). Akşin et al. (2013) take the structural estimation approach to study caller abandonment behaviors in a call center. Subsequent work including Akşin et al. (2016) and Yu et al. (2016) study the impact of delay announcement in call centers. The element of human customers and human servers often introduces interesting dynamics. For instance, Lu et al. (2013) find that observed queue lengths impact the purchasing behavior of customers in a super market setting. Emadi and Staats (2019) find that the attrition of agents at a management firm appears to be insensitive to salary.

Structural models have also been used specifically in the healthcare operations management literature. Olivares et al. (2008) use a newsvendor model to study how a hospital

balances the costs of reserving too much versus too little operating room capacity for cardiac surgery cases. In a different operating room setting, Rath and Rajaram (2018) use a choice model to estimate costs associated with operating room scheduling of anesthesiologists. Our work contributes to the application of structural estimation in healthcare operations management, but in the ICU setting. To the best of our knowledge, we are the first to estimate a dynamic structural model in healthcare operations.

The third stream of relevant literature is behavioral operations management and, particularly, its application in healthcare. There is evidence that the behaviors of physicians, staffs, or patients can have substantial impact. Green et al. (2013) studies nurse absenteeism and finds nurses exhibit aversion to higher levels of anticipated workload, leading to endogenous absenteeism rates that must be considered in nurse staffing. Song et al. (2015) compare the ED wait time between a system with dedicated queues versus a pooled queue. They find the wait time decreases when de-pooling and suggest that such a phenomenon has a behavioral explanation where physicians feel an increased ownership of patient wait time when faced with a dedicated queue. Ibanez et al. (2017) examines how radiologists view scans given their complete discretion to determine the order to complete tasks. This discretion can lead to inefficiencies in completing tasks. In a setting very similar to ours, Kim et al. (2019) studies the ICU admission decision from a behavioral perspective. They propose a behavioral model and use controlled experiments to understand whether and how physicians are impacted by occupancy when making admission decisions. They identify a number of factors, such as the availability of information, which can bias physician decisions. While we also look at the ICU admission decision, we focus on the behavior of the hospital and take a structural estimation approach to estimate the discount factor from data.

Finally, from the methodological aspect, our work is related to the literature on identification and estimation of dynamic discrete choice models. The dynamic discrete choice model we use resembles the work in the econometrics community pioneered by Rust (1987). Our study extends this line of literature by applying the dynamic discrete choice model in the ICU context to study the hospital's admission decisions. Furthermore, we estimate the discount factor and cost parameters jointly from empirical data. The identification of the discount factor is generally a very hard problem for dynamic models (Magnac and Thes-

mar, 2002). We rely on recent developments in Komarova et al. (2018) that establishes joint identification of discount factor and payoff parameters for dynamic choice models with linear structure. The adaption of this abstract methodology to our healthcare setting is not straightforward; it requires extending Komarova et al. (2018) to our setting with a state-dependent action space and verifying the exclusion criteria is satisfied by our model.

3.2 Setting and Data

We utilize a large data set from 21 Kaiser Permanente Northern California (KPNC) hospitals. The data contains over 312,306 hospitalizations over the period of two years. All patients are covered by KPNC insurance and received care at one of the KPNC hospitals. The hospitals cover a large geographic area and intra-hospital transfers are quite rare. As such, we will generally study each hospital separately.

Each observation in our data corresponds to a single hospitalization. For each hospitalization, we have patient level information such as age, gender, admitting hospital, admitting diagnosis, and three severity scores. The severity scores include a measure of the patient's chronic disease burden (COPS2), an acuity score (LAPS2), and a predicted in-hospital mortality risk score (CHMR). The LAPS2 score is the main severity measure we use in the analysis. It is assigned at hospital admission and measures the clinical severity of a patient based on labs and vital signs taken in the last 72 hours prior to admission – including any that may have been taken in the ED; a score of 110 is generally considered to capture a critically ill patient. More details about these scores can be found in Escobar et al. (2012) and Escobar et al. (2013). In addition to the patient level information, we also observe the admission and discharge time for each unit each patient stayed in during the hospitalization, as well as the type of care the unit provides – i.e., ICU, transitional care unit (TCU), general medical-surgical ward, operating room (OR), or the postanesthesia care unit (PAR). It is important to note that while we are able to see the full trajectory of each patient, including the ultimate decision outcome, the data does not include direct information on the decision making process. Additionally, while we have a rich dataset that includes detailed patient data, we do not have data on other factors that may influence the decision such as (i) nurse

staffing availability, (ii) ED patient census, (iii) diversion policies, (iv) specific physician coverage policies, and (v) the possibility to flex capacity.

We utilize the data from all hospitalizations to compute the maximum occupancy and real-time occupancy level of the ICU in each hospital. The maximum ICU occupancy varies from 7 to 36 beds across hospitals, and the average occupancy level varies from 34% to 76%. Among all ICU admissions, 63% are admitted via the ED to a medical service; 10% are admitted through a non-ED unit to a medical service; 10% are admitted via the ED to a surgical service, and 17% are admitted through a non-ED unit to a surgical service (i.e., these patients are scheduled surgeries).

Our study focuses on how the hospital internalizes the intertemporal externalities in ICU admissions. Since non-ED patients are more likely to be scheduled arrivals and because there are often fixed care protocols for surgical patients, our study is most relevant for patients admitted to a medical service via the ED. The ICU admission decision for these patients is made as follows. After a patient is stabilized in the ED, the ED physician provides an initial assessment about whether the patient needs to be admitted to the hospital. If the ED physician believes the patient needs to go to the ICU, an intensivist will be called to the ED for a consultation. While the intensivist makes the ultimate decision about whether and when the patient is admitted to the ICU, it is important to emphasize that the decision is determined by a system including various physicians, administrators, and possibly, patient family members, as well as the types of alternative interventions available (e.g. flexing capacity).

Next, we describe the data selection process for our study cohort. We start from a total of 312,306 hospitalizations. We restrict our study to the hospitalizations admitted to a medical service via the ED, which comprises the largest proportion of admitted patients ($> 60\%$). Note that for patients who are admitted via the ED, they appear in our data set as soon as the admission decision has been made; as such, we do not have information about patients discharged home from the ED nor patients for whom a disposition decision has not yet been made. We drop 12 hospitalizations with unknown gender and 9,128 (4.8%) hospitalizations for patients who experience hospital transfers or transports outside of KPNC. As we explain in more detail in Section Section 3.4, our study focuses on three possible decisions for each

patient in each decision epoch: keep the patient waiting in the ED, admit the patient to the ICU, or admit the patient to a non-ICU unit (e.g. the ward TCU). We drop 3,066 (1.7%) hospitalizations where the patient was admitted to other units – e.g., OR or PAR, from the ED. Finally, we drop 1,675 (1%) hospitalizations with ED waiting time longer than 12 hours as these episodes can be considered outliers (the average waiting time is shorter than two hours).

Because our data spans over two years, some hospitals might adjusted their ICU capacities during the sample period. As a result, we restrict our study cohort to the periods of each hospital with stable ICU capacity and occupancy. We follow three steps to select the sample. First, we discard the first and last month of data for all hospitals. Second, for several hospitals, we drop the period at either end of the sample where the ICU occupancy dramatically fluctuates or significantly differs from the more stable period in the middle. Finally, for hospital 21, we find that its ICU capacity experienced a substantial increase during the sample period (from 13 to 16). As a result, we split it into two parts, i.e., before and after the capacity change, and treat them as two hospitals in the estimation. We refer to 22 hospitals in our study cohort from here on. The number of days and hospitalizations for each hospital in the final study cohort are summarized in Table C.3 in Appendix C.2.3. In total, we drop 11,268 (6.4%) hospitalizations that are outside the stable periods.

The final study cohort consists of 164,167 hospitalizations. Out of them, 19,683 (12.0%) are admitted to the ICU, and the remaining admitted to a non-ICU unit. In Table 3.1, we summarize the patient characteristics of the final study cohort (left) and the subset which are admitted to the ICU (right). As expected, the admitted cohort has higher average severity scores than the complete cohort.

3.3 Descriptive Evidence for Discounting Behaviors

In this section, we provide descriptive evidence directly from the data on the discounting behaviors of hospitals with respect to the short versus longer term considerations. The results provide motivations for the decision model which describes this behavior in the next section.

Table 3.1: Summary Statistics of Patient characteristics of final study cohort and the subset of patients who are admitted to the ICU.

Final study cohort: N=164,167

	Min	Max	Mean	Median	SD
LAPS2	0.00	294.00	74.11	70.00	37.47
COPS2	0.00	306.00	44.97	28.00	43.09
CHMR	0.00	0.99	0.04	0.02	0.08
Male			0.53		
Age (years)	18.00	113.00	67.27	70.00	17.59
EDWait (hours)	0.02	12.00	1.30	0.88	1.41

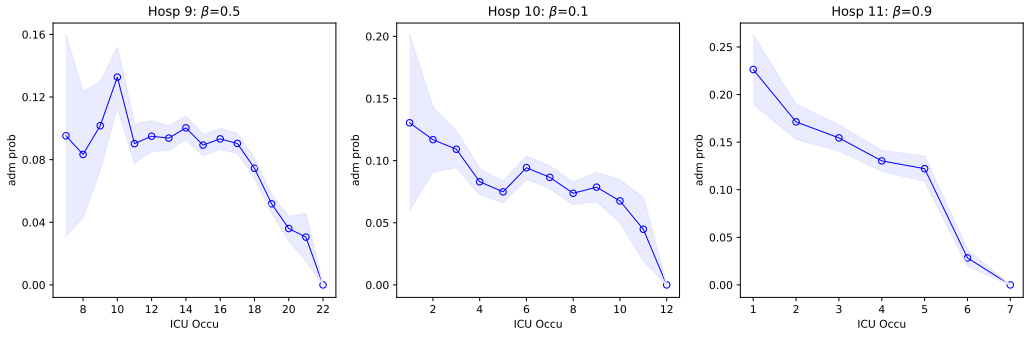
ICU admission cohort: N=19,683

	Min	Max	Mean	Median	SD
LAPS2	0.00	294.00	105.03	102.00	45.98
COPS2	0.00	276.00	48.67	30.00	47.27
CHMR	0.00	0.99	0.10	0.04	0.15
Male			0.48		
Age	18.00	111.00	64.52	67.00	17.48
EDWait	0.02	11.98	1.36	0.90	1.45

Note. LAPS2, COPS2, and CHMR are severity of illness scores. EDWait corresponds to the ED boarding time.

At a high level, one can think of the discounting behavior as how far into the future the hospital considers when making decisions. A hospital with a low degree of discounting behavior likely focuses primarily on the implications of the admission decision on the current patient and may consider system dynamics over a shorter horizon of two to six hours. On the other hand, a hospital with more a longer-term focus is concerned with the current patient as well as the impact any decision will have on the ability to treat patients who may arrive later, for example, within the next 24 hours. This discounting behavior can manifest itself by how much the hospital alters its admission decisions based on congestion. While the hospital may not be fully aware of its potential change in behavior due to congestion, we aim to understand the perceived discounting behavior that can be elicited from the admission decisions of patients.

Figure 3.1: Probability of ICU admission by ICU occupancy levels (selected hospitals)



In Figure 3.1, we show the ICU admission probability with respect to ICU occupancy level in three representative hospitals. The shaded area denotes the 95% level confidence interval. From the figure, we can see the ICU admission probability generally drops as ICU

occupancy increases. Moreover, the decrease happens well before the ICU is full. This provides descriptive evidence for the discounting behaviors, i.e., hospitals indeed take into account the ability to service future patients when making ICU admission decisions for current patient. These patterns cannot be explained by standard queueing systems.

We observe that hospitals can behave very differently in response to ICU congestion. For example, the ICU admission probability in Hospital 11 (right panel) starts to drop at low ICU occupancy levels, while this change happens at relatively high ICU occupancy levels in Hospitals 9 and 10 (left and middle panels). We find similar results through a comprehensive multinomial logit model in Appendix C.1. The observed variation suggests hospitals may have very different *degrees* of how much they internalize the intertemporal externalities. In order to quantify this behavior, we propose a structural model for ICU admission decisions in the next section. More importantly, the structural model allows us to conduct counterfactual analyses to quantify the impact of the forward looking behavior on key system performance metrics, which is not possible via descriptive statistics and reduced form analysis.

3.4 Structural Estimation

In this section, we first introduce the structural model which describes the ICU admission process within a hospital. Next, we explain the identification of the discount factor and describe the estimation procedure.

3.4.1 Dynamic Discrete Choice Model

In our structural model, we consider the admission decision at the level of the hospital. We model the ICU admission decision using a dynamic discrete choice model. There are three key features of the model: 1) in each period, the hospital considers three options for each patient: admitting the patient to the ICU, admitting the patient to a non-ICU unit (e.g., the ward), or keeping the patient waiting in the ED; 2) the decision depends on both patient severity and the current system status; 3) finally, the model allows for the hospital

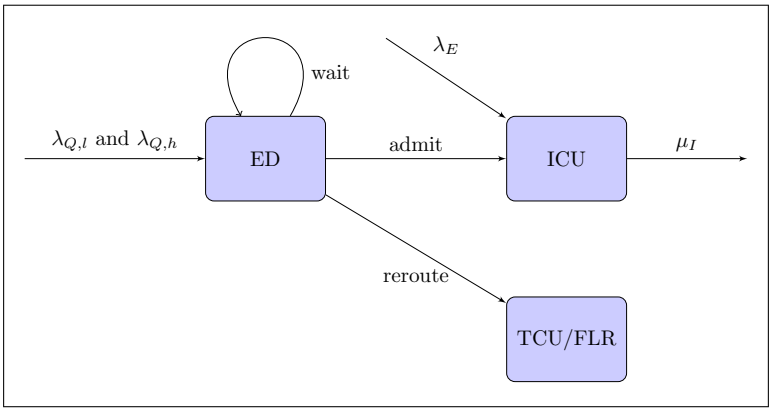
to consider the future and to make dynamic decisions. We provide detailed descriptions of the model below.

We divide the focal patient cohort into two classes: the low and high severity class, represented by subscripts l and h, respectively. We define the two classes based on the patient's LAPS2 score, which has been shown to be an effective measure of illness severity (Escobar et al., 2013). We assume that the ED has capacities Q_l and Q_h for the low and high classes, respectively. Note that in our model a patient arrival in the ED corresponds to when the decision to admit the patient to the hospital has been made (i.e. we only capture ED boarding patients), so the ED capacities can be regarded as capturing the limited boarding space in the ED. We defer the discussion about how various parameters in the model, including Q_l and Q_h, are estimated from the data to Section 3.4.3. The two classes of patients arrive in the ED every period according to the following distribution. For $i \in \{l, h\}$, let $A_{i,t}$ be the number of class i patients arriving to the ED in time period t. $A_{i,t}$ follows a truncated Poisson distribution with rate $\lambda_{Q,i}$. While in theory $A_{i,t}$ can be unbounded, we truncate it at the maximum number of patients arriving at ED in each period (denoted by M_{A_i}) observed in our data. This limits the state space of the model and helps to keep the estimation computationally feasible.

In addition to our focal patient cohort, there are other patients who can occupy ICU beds. E_t denotes the number of surgical and non-ED medical patients arriving in the ICU in period t, referred to as the external arrivals. E_t is distributed according to a Poisson distribution with arrival rate λ_E.

In each period, each ICU patient – from the ED or externally – departs from the ICU with probability μ_I. Thus, D_t, the number of patients departing from ICU in each period, follows a binomial distribution. The total number of beds in the ICU is B. We assume the ward units in the hospital have ample capacity. This helps with the computational complexity of our model; moreover, the ward is generally much less congested than the ICU. Indeed, the proportion of periods where the ward occupancy exceeds 95% of its capacity (median capacity of 95 beds) is less than 0.8%, and the proportion of periods with full ward occupancy is less than 0.05% (an order of magnitude less often than the ICU which has a median capacity of 16 beds). The system flow is depicted in Figure 3.2.

Figure 3.2: Overview of patient flow and potential paths in the ED-ICU/ward system



At the beginning of period t, the system state is given by a three dimensional vector,

$$s_t = (n^E_{l,t}, n^E_{h,t}, n^I_t),$$

where $n^E_{i,t}$ is the number of class i patients in the ED, $i \in \{l, h\}$, and n^I_t is the number of patients in the ICU. By the capacity constraints, $n^E_{i,t} \leq Q_i$ for $i \in \{l, h\}$ and $n^I_t \leq B$. For each patient in the ED, the hospital determines one of the following three decisions: admit to ICU, admit to non-ICU units, or keep them waiting in the ED. Since the patients are treated as identical in terms of their observables within each class, the hospital's action can be described by the following four dimensional vector

$$d_t = (a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}),$$

where $a_{i,t}$ and $r_{i,t}$ denote the numbers of patients admitted to ICU and non-ICU units of class i, $i \in \{l, h\}$, respectively. Due to the capacity constraint in the ICU, the admissible action set for system state s_t is

$$\Pi(s_t) = \left\{(a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}) : a_{l,t} + r_{l,t} \leq n^E_{l,t},\ a_{h,t} + r_{h,t} \leq n^E_{h,t},\ a_{l,t} + a_{h,t} \leq B - n^I_t\right\}. \quad (3.1)$$

$\Pi(s_t)$ specifies the following set of constraints: The first two constraints state that the sum of admitted patients (ICU and non-ICU units) must be smaller than or equal to the total number of patients currently in the ED. The last constraint requires that the total number of patients admitted to ICU must be smaller than or equal to the current number of available beds in the ICU.

In each period, a cost $c(s_t, d_t)$ associated with state s_t and action d_t is incurred. We note that these costs represent the hospital's assessment of the 'costs', or disutility, incurred for each state-action pair. They may capture clinical costs, operational costs, financial costs, etc. These costs are not necessarily consistent with each individual stakeholder's assessment (e.g. patient, ED physician, ICU physician, hospital administrator, etc.) of the costs associated with each state-action pair; they can also capture various system-level constraints. We assume admitting a patient to the ICU has zero costs for both classes of patients. This assumption is a reasonable approximation for the following reasons. First, ICUs have very high fixed operating costs, therefore, the marginal cost of admitting one patient is very small for either class of patients (Roberts et al., 1999; Kahn et al., 2008). For example, the extremely high set-up cost for specialized equipment such as ventilators and monitors in the ICU can be regarded as sunk costs, which do not affect the hospital's ICU admission decision. Additionally, hospitals tend to staff their ICU beds based on fixed nurse-to-bed ratios and rarely adjust staffing levels based on occupancy and severity of patients. Finally, ICU patients are the most severe type in a hospital. Therefore, it is reasonable to assume that the hospital's primary concern is to prioritize patient outcomes, and the immediate cost of ICU admission is negligible in comparison. In other words, the hospital's primary goal is to minimize undesirable patient outcomes due to lack of ICU care. As such, we assume zero admission costs for both classes of patients. This implies that all other costs are assessed relative to ICU admission. As we discuss in Section 3.4.2, the assumption of zero ICU admission cost is crucial for the identification of the discount factor in the dynamic discrete choice model.

We assume that admitting each low (high) severity patient to non-ICU units incurs a non-ICU routing cost $c_{r,l}$ ($c_{r,h}$), while keeping each low (high) severity patient waiting in the ED incurs a waiting cost $c_{w,l}$ ($c_{w,h}$). These cost parameters represent the average non-ICU admission or waiting cost across patients within each class. We restrict the waiting costs for both classes to be positive, i.e., $c_{w,l}, c_{w,h} > 0$. This is a reasonable assumption since longer ED boarding has been shown to be associated with increased mortality risk and hospital LOS (Singer et al., 2011). Thus, on average, keeping patients in the ED is more likely to lead to negative outcomes compared with admitting them to ICU. As we

assume ICU admission incurs zero costs for both classes, the average waiting costs $c_{w,l}$ and $c_{w,h}$ should both be positive as longer waiting time is less desirable than immediate ICU admission. On the other hand, we do not restrict the sign of the non-ICU admission costs $c_{r,l}$ and $c_{r,h}$. Patients who are not critically ill can often receive sufficient care in the ward, so admitting them to the ward is not necessarily worse than admitting them to the ICU. Therefore, the average cost of non-ICU admission, compared with ICU admission, can be positive or negative. We note the relative differences in costs across the low versus high severity patients captures the tradeoff between admitting (or delaying) the high versus low severity patients. Finally, there are no costs associated with external arrival patients in our model. The majority (> 70%) of the external arrivals to the ICU are surgical patients (63% of which are scheduled surgeries) for whom ICU beds are often reserved in advance; thus, our model focuses on the costs associated with the patients for whom there are less clear protocols – those admitted to a medical service via the ED.

Given the system state, s_t, and the action, d_t, the total per period cost is given by,

$$c(s_t, d_t) = c_{r,l} r_{l,t} + c_{w,l}\left(n^E_{l,t} - a_{l,t} - r_{l,t}\right) + c_{r,h} r_{h,t} + c_{w,h}\left(n^E_{h,t} - a_{h,t} - r_{h,t}\right), \tag{3.2}$$

which is the sum of non-ICU admission and waiting costs for the two classes of patients. Then, the hospital's per period utility can be written as,

$$U(s_t, d_t, \varepsilon_t) = -c(s_t, d_t) + \varepsilon_t(d_t), \tag{3.3}$$

where $\varepsilon_t(d_t)$ is the idiosyncratic utility component associated with action d_t, which is observed by the hospital when making the decision, but not the researcher. The additively separable form (3.3) is similar to the assumption in Rust (1987) and numerous works in the structural estimation of dynamic discrete choice literature.

At the beginning of period t, the hospital observes the system state, s_t and the idiosyncratic utility component, ε_t, then they choose the optimal action d_t that solves following infinite horizon utility maximization problem:

$$\sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t\right\}. \tag{3.4}$$

The discount factor $\beta \in (0,1)$ captures the trade-off between current and future utility, which is the focus of our study. The expectation is taken over both the random component ε_t and the transitions of the system – i.e., the arrivals and departures of patients in each period. In addition, note that the expectation in (3.4) is conditional on both s_t and ε_t, as the random component is observable to the hospital before making decision in period t. We define the value function as the objective in (3.4) given the optimal action sequence, i.e.,

$$V(s_t, \varepsilon_t) = \sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{ \sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t \right\}. \tag{3.5}$$

As noted earlier, the costs in the structural model capture many aspects that influence hospital's decisions, including, but not limited to, patient outcomes, financial costs, and operational constraints. The optimization problem here is intended to model hospital's decision making and its relationship to the observed flows of patients. Thus, it does not require a precise definition of what the costs are capturing. The implicit assumption is that the hospital is acting "rationally" by taking the optimal action that is consistent with the fixed costs and discount factor in the model. In other words, the assumption imposed by the model is the structure of the hospital's objective only. As the "true" objective of the hospital is unknown, the costs and discount factor can be perceived as behavioral parameters that capture how the hospital is balancing between near-term versus long-term costs.

After the hospital chooses an action in period t, the system state evolves as follows. The number of ED patients of class i becomes $n^E_{i,t} - a_{i,t} - r_{i,t}$, and the number of patients in the ICU is $n^I_t + a_{l,t} + a_{h,t}$. We define an "intermediary" state $\varphi(s_t, d_t)$ after the action d_t is taken, which is given by

$$\varphi(s_t, d_t) = \left(n^E_{l,t} - a_{l,t} - r_{l,t}, n^E_{h,t} - a_{h,t} - r_{h,t}, n^I_t + a_{l,t} + a_{h,t} \right), \tag{3.6}$$

and describes the impact of action d_t on the system.

The system then evolves according to the following two steps. First, $A_{i,t}$ new patients of class i arrive to the ED, and E_t patients arrive to the ICU through non-ED channels–i.e., the external arrivals. If the ED or ICU is full, new arrivals cannot be accepted. Thus, the total accepted ED and ICU arrivals are given by

$$A^{acc}_{i,t} = \max\left\{ A_{i,t}, Q_i - (n^E_{i,t} - a_{i,t} - r_{i,t}) \right\}$$

and

$$E_t^{acc} = \max\left\{E_t, B - (n_t^I + a_{l,t} + a_{h,t})\right\},$$

respectively. Second, D_t patients leave the system as they complete their service in the ICU. This completes the system transition for period t.

The system state at the beginning of period $t+1$ is:

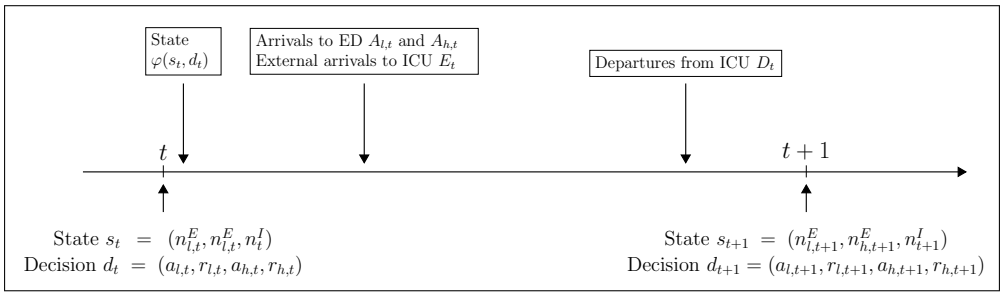
$$s_{t+1} = (n_{l,t+1}^E, n_{h,t+1}^E, n_{t+1}^I),$$

with

$$n_{i,t+1}^E = n_{i,t}^E - a_{i,t} - r_{i,t} + A_{i,t}^{acc} \quad \text{and} \quad n_{t+1}^I = n_t^I + a_{l,t} + a_{h,t} + E_t^{acc} - D_t. \tag{3.7}$$

It is clear from the above description that the transition of s_t is Markovian, and its distribution only depends on s_t and d_t, but not ε_t. The timeline of system transition is summarized in Figure 3.3.

Figure 3.3: Timeline of system evolution: depiction of how the state evolves within a single time-slot.



Both the current period's utility and how a specific action changes the system state, which in turn can impact future payoffs, may influence the hospital's decision. Thus, the hospital chooses the action that maximizes

$$d_t = \underset{d \in \Pi(s_t)}{\arg\max}(-c(s_t, d) + \varepsilon_t(d) + \beta \mathsf{E}[V(s_{t+1}, \varepsilon_{t+1})]),$$

where the function $V(s, \varepsilon)$ is defined in (3.5). The last term in the right-hand side is the expectation of the future value function after the current action is taken. Thus, the optimal value function $V(s_t, \varepsilon_t)$ solves the following Bellman's equation

$$V(s_t, \varepsilon_t) = \max_{d \in \Pi(s_t)} (-c(s_t, d) + \varepsilon_t(d) + \beta \mathsf{E}[V(s_{t+1}, \varepsilon_{t+1})]),$$

where the expectation is taken over both the system transition to s_{t+1} and the random component ε_{t+1}.

The above Bellman's equation is hard to evaluate due to the infinite state space associated with ε_t. Thus, we simplify the model by making the same conditional independence assumption (CI) as in Rust (1987).

Assumption 1 (CI)**.** *The transition probabilities of the controlled process* (s_t, ε_t) *can be factored as*

$$Pr(s_{t+1}, \varepsilon_{t+1}|s_t, \varepsilon_t, d_t) = q(\varepsilon_{t+1}|s_{t+1})g(s_{t+1}|\varphi(s_t, d_t)), \tag{3.8}$$

where $\varphi(s_t, d_t)$ denotes the intermediate state (3.6) after action d_t is taken; the transition probability $g(s_{t+1}|\varphi(s_t, d_t))$ captures the random arrivals and departures shown by (3.7). Assumption (CI) states that s_{t+1} is sufficient to determine the distribution of ε_{t+1}. In other words, the random component $\{\varepsilon_t\}$ is superimposed on the state process $\{s_t\}$. Finally, we assume that the random component ε_t is independent and identically distributed (i.i.d.) and follows type I extreme value distribution for each action $d \in \Pi(s_t)$. Thus, the state s_t impacts the distribution of ε_t only through the number of admissible actions. As shown in Rust (1987), this assumption leads to a closed-form expression of the conditional choice probability for action d_t given state s_t, as denoted by $f(d_t|s_t)$.

Proposition 1. *With the above set-up, the conditional choice probability for action* d_t *given state* s_t *has the following closed-form representation:*

$$f(d_t|s_t) = \frac{\exp\big(-c(s_t, d_t) + \beta\tilde{V}(\varphi(s_t, d_t))\big)}{\sum_{d\in\Pi(s_t)} \exp\big(-c(s_t, d) + \beta\tilde{V}(\varphi(s_t, d))\big)}, \tag{3.9}$$

where function $\varphi(s_t, d_t)$ *is given by (3.6). The function* $\tilde{V}(s)$ *is defined as*

$$\tilde{V}(s) = \sum_{s'} \int_{\varepsilon'} V(s', \varepsilon')\ g(s'|s)q(\varepsilon'|s')d\varepsilon'. \tag{3.10}$$

The explicit expression for $g(s'|s)$*, i.e. the transition probability to state* s' *given* s *(the system state after the action is taken but before the random arrivals and departures take place), is provided in Appendix C.2.1. The function* $\tilde{V}(s)$ *is the unique fixed point to the following functional equation*

$$\tilde{V}(s) = \sum_{s'} \ln\left\{\sum_{d'\in\Pi(s')} \exp\big(-c\,(s', d') + \beta\tilde{V}(\varphi(s', d'))\big)\right\} g(s'|s). \tag{3.11}$$

In the above proposition, we drop the dependence on model parameters to simplify the notation. As in most dynamic discrete choice models, the choice probability (3.9) has a closed-form logit representation and the value function $\tilde{V}(s)$ solves the functional equation (3.11). By (3.10), the new value function $\tilde{V}(s)$ represents the expected future utility given s, which represents the current state after the action has been taken. In the proposition below, we show it is monotonically non-increasing in the number of ICU patients.

Proposition 2. *For two intermediary states (after actions are taken) s and s' with $\left(n_i^E\right) = \left(n_i^E\right)'$ for $i \in \{l, h\}$ and $\left(n^I\right) \leq \left(n^I\right)'$, we have*

$$\tilde{V}(s) \geq \tilde{V}(s').$$

That is, for any given number of ED patients, the function $\tilde{V}(s)$ is monotonically non-increasing in the number of ICU patients.

Proof: See Appendix C.4. □

This proposition shows that the intertemporal externalities indeed exist for the ICU admission decisions: As the hospital admits more patients, thereby increasing the ICU occupancy, the future expected utility decreases. In turn, this can result in a decrease in the likelihood of ICU admission as the ICU occupancy increases. Thus, the hospital must consider both current and future utilities when making ICU admission decisions.

3.4.2 Identification of Discount Factor

In this section, we discuss why identification is challenging and how our model is empirically identified from observed data. We first develop the general identification result for our model using recent developments in the econometric literature. Then, we use two simple examples to illustrate how some aspects of our model can be directly identified from observed choice probabilities.

In this study, we identify the discount factor and cost parameters jointly from observed data. Although the choice probability has a closed-form expression (3.9) in the dynamic choice model, it is not possible to identify the discount factor and costs parameters jointly without further restriction on the dynamic model (see, e.g. Lemma 3.3 in Rust (1994) or

Proposition 2 in Magnac and Thesmar (2002)). Thus, most empirical studies assume the discount factor is known, and then estimate the cost parameters. However, the prespecified discount factors usually lack empirical support and economic justifications. Indeed, the implied discount rate can vary substantially across different settings (Frederick et al., 2002). In our study, identifying the discount factor from observed data is crucial to understanding how much hospitals react to the intertemporal externalities of the ICU admission decisions.

The reason that the discount factor and cost parameters cannot be jointly identified in general is because there exists observationally equivalent structures – i.e., different combinations of discount factor and cost parameters – that lead to the same choice probabilities for all states and actions. Thus, an agent's actions can be rationalized for different choices of discount factor. For example, in the ICU admission context, if we observe the hospital admits patients "aggressively" (e.g. with high probability), this may be because either the hospital is focused on the near-term – i.e., a small β – or the waiting and non-ICU admission costs are large – i.e., high c_w and c_r. Without further restriction on model primitives (e.g., cost and utility), we cannot differentiate between such cases to identify the true discount factor.

There have been positive identification results for the discount factor in dynamic models. Magnac and Thesmar (2002) suggests that exclusion restriction can be used to identify the discount factor. The exclusion restriction they use is that there exists some state and action pairs for which the single period utilities are identical, but the future period utilities differ. This idea is further elaborated and applied in some empirical contexts (Dubé et al. (2014), Wang (2014), and Ching and Osborne (2017)). However, the exclusion restriction is abstract and hard to verify in practice (Magnac and Thesmar, 2002; Abbring and Daljord, 2019). Thus, despite this abstract methodology, there are only two recent papers which estimated the discount factor empirically. Both focus on special cases and provide important contributions in demonstrating identification in those settings. This is a common approach for other theoretical econometric methodologies (e.g. instrumental variables), which require careful application in each specific setting to appropriately achieve identification.

To identify the discount factor in our parametric model, we leverage the recent identification results developed in Komarova et al. (2018). They prove the identification for the

discount factor using an empirical model that is linear in the cost parameters conditional on the discount factor. In particular, they construct a one-dimensional criterion function that can be used for identification as well as estimation, which exploits the conditional linear structure and reduces the nonlinear problem to a one-dimensional grid search for the discount factor. For this approach to work in our setting, we must appropriately adapt it.

The main identification results of Komarova et al. (2018) proceed as follows. They consider an empirical model with linear structure. The choice and transition probabilities are nonparametrically identified. For a given value of the discount factor, they first construct estimates for the cost parameters following the standard two-step estimation procedure pioneered by Hotz and Miller (1993). The estimator minimizes the distance between the value functions observed from data and those directly implied by the empirical model. Then, they reduce the identification problem to a one-dimensional search for $\beta \in (0, 1)$: If there is a unique value of β, together with the corresponding cost estimates, that minimizes the distance objective function, then the model can be identified under some rank condition. This criterion also provides a natural way to estimate the model with observed data.

We now show how the identification results in Komarova et al. (2018) are applied in our setting. First, by construction, our model satisfies the basic assumptions in Komarova et al. (2018), i.e., additive separability of utility, conditional independence of transition and finite state space. Additionally, by (3.2) and (3.3), it is also clear that the deterministic part of the per-period utility (i.e., $-c(s_t, d_t)$) is linear in the cost parameters $c_{w,l}$, $c_{r,l}$, $c_{w,h}$, and $c_{r,h}$. Thus, the linear-in-parameter assumption is also satisfied in our setting. More details of the assumptions in Komarova et al. (2018) and how they apply to our setting are included in Appendix C.2.2. We define the base action as admitting every ED patient whenever there are ICU beds available. As we assume ICU admission costs are zero for both classes of patients, the base action brings zero cost for all states as long as it is admissible. Thus, we can apply Theorem 1 of Komarova et al. (2018) in our setting to identify the discount factor and cost parameters jointly in our dynamic model.

We note that Komarova et al. (2018) assumes the same admissible action set for all states while our setting has state dependent action sets. This creates additional challenges in the identification of the discount factor as the denominator for the choice probability in

(3.9) is also state-dependent. Thus, even if we could verify all the abstract assumptions in Komarova et al. (2018), for the identification of the discount factor to be achieved in our setting, a direct application is still not possible. However, we note that given each state, the admissible action set can be fully determined by (3.1). Thus, we can verify the derivation leading to Theorem 1 in Komarova et al. (2018) still applies to our setting, where we plug in the admissible action set for each state according to (3.1). Additionally, in next section, we use two simple examples to show how the state-dependence property of the admissible action set can be handled by constructing proper state-action pairs that identify certain aspects of our model.

We can construct a one-dimensional criterion similar to the one in Komarova et al. (2018) based on the maximum likelihood estimator, which is asymptotically equivalent to the estimator employed in Komarova et al. (2018): For each candidate β, we estimate the cost parameters that maximize the choice likelihood. Then we conduct a one-dimensional search over $\beta \in (0, 1)$. The model can be identified if there is a unique β (together with the cost estimates) that maximizes the likelihood, and that the rank condition in Theorem 1 of Komarova et al. (2018) is satisfied. Using data from 22 hospitals, we find the optimal discount factor that maximizes the likelihood function is unique in all circumstances. Moreover, in most cases, the likelihood function monotonically decreases as the discount factor moves away from the optimal level, which further strengthens the identification of the discount factor. More details of the algorithmic approach can be found in the next section. We provide some illustrative examples of the likelihood versus discount factor in Figure C.1 of Appendix C.2.3. We also show by simulation that the rank condition is satisfied. The results suggest that the discount factor and cost parameters can be jointly identified in our empirical model.

3.4.2.1 Two Illustrative Examples for Identification

In this section, we use two simple examples to illustrate how some aspects of our dynamic model can be directly identified from observed choice probabilities. While the formal identification results rely on the one-dimensional criterion established above, the two examples below provide some insights about how the observed choice probabilities can be used

for identification. Taking log on both sides of the choice probability (3.9), we have

$$\ln f(d_t|s_t) = -c(s_t, d_t) + \beta\tilde{V}(\varphi(s_t, d_t)) - C(s_t), \tag{3.12}$$

where function $C(s)$ is given by

$$C(s) = \ln\left\{\sum_{d\in\Pi(s)} \exp\left(-c(s,d) + \beta\tilde{V}(\varphi(s,d))\right)\right\}.$$

We see the log choice probability can be decomposed to three terms in (3.12): the first term $-c(s_t, d_t)$ is the negative of the per-period cost, which directly depends on the cost parameters; the second term $\beta\tilde{V}(\varphi(s_t, d_t))$ is related to the discount factor and value function, which captures the impact of the decision on future utility via the change of system state; the last term $C(s_t)$ is a function of system state s_t, but it does not depend on the decision d_t. Note the discount factor and cost parameters are implicitly captured in $\tilde{V}(\cdot)$ and $C(\cdot)$.

As shown by (3.12), the discount factor and cost parameters are captured in the choice probabilities in a complicated way. However, we will illustrate that certain aspects of the model can be directly identified by constructing appropriate state and action pairs. We provide two examples below. In the first example, we show the non-ICU admission cost difference, i.e., $c_{r,h} - c_{r,l}$, can be identified by cancelling out the second and third terms related to $\beta\tilde{V}(\cdot)$ and $C(\cdot)$ in (3.12). In the second example, we show certain linear combinations of $\beta\tilde{V}(\cdot)$ can be identified by removing the first and third terms related to $c(s_t, d_t)$ and $C(\cdot)$ in (3.12). The proofs are provided in Appendix C.5.

Lemma 1. *Denote two states $s_1 = (1, 0, 0)$ and $s_2 = (0, 1, 0)$. That is, there is one patient in the ED from the low and high severity class in s_1 and s_2, respectively, and no patient in the ICU. Then the difference in non-ICU admission costs can be identified as*

$$c_{r,h} - c_{r,l} = \ln\left(\frac{Pr(r|s_1)}{Pr(a|s_1)}\right) - \ln\left(\frac{Pr(r|s_2)}{Pr(a|s_2)}\right), \tag{3.13}$$

where $Pr(a|s_1)$, $Pr(r|s_2)$, $Pr(a|s_1)$, and $Pr(r|s_2)$ denote the probabilities of admitting the patient to ICU and non-ICU units under the two states, respectively.

This result can be interpreted as follows. The ratio of non-ICU to ICU admission probabilities, i.e., $\ln(\mathsf{Pr}(r|s_i)/\mathsf{Pr}(a|s_i))$ for $i = 1, 2$, is negatively related to the relative cost of

admitting the patient to non-ICU unit compared with admitting to the ICU. As the ICU admission cost is assumed to be zero for both patient classes, higher non-ICU to ICU admission ratio implies a smaller non-ICU admission cost. Thus, the difference in the log ratios can be used to identify the difference in non-ICU admission costs of the two classes of patients. A larger difference on the right-hand side of (C.19) suggests the patients from the high severity class are less likely to be admitted to non-ICU unit compared with those from the low severity class, which implies a larger difference in their non-ICU admission costs.

Lemma 2. *Consider two states with* $s_1 = (0, 1, 1)$ *and* $s_2 = (0, 1, 0)$. *That is, there is one patient in the ICU in the first state, and one high severity patient in the ED in both states. Let* $\tilde{V}_k = \tilde{V}((0, 0, k))$ *denote the value function of the state with* k *patients in the ICU and no patients in the ED. Then we have*

$$\beta\left[(\tilde{V}_2 - \tilde{V}_1) - (\tilde{V}_1 - \tilde{V}_0)\right] = \ln\left(\frac{Pr(a|s_1)}{Pr(r|s_1)}\right) - \ln\left(\frac{Pr(a|s_2)}{Pr(r|s_2)}\right). \tag{3.14}$$

where $Pr(a|s_1)$, $Pr(r|s_2)$, $Pr(a|s_1)$, *and* $Pr(r|s_2)$ *denote the probabilities of admitting the high severity patient to ICU and non-ICU units under the two states, respectively.*

The left-hand side of (C.22) is a linear combination of value functions multiplied by the discount factor β. Specifically, $\tilde{V}_2 - \tilde{V}_1$ (resp. $\tilde{V}_1 - \tilde{V}_0$) measures the impact on the future payoff from admitting one more patient to ICU when the ICU currently has one (resp.zero) patient. Thus, the left-hand side of (C.22) actually measures the change in the impact on the future payoff when adding one more patient to the ICU in its current state. Accordingly, it can be identified by the change in log ratios of ICU and non-ICU admission probabilities as the ICU state moves from s_1 to s_2, which have one and zero patients respectively.

3.4.3 Algorithmic Approach

We now document the details for how the dynamic discrete choice model is estimated from data. We employ the nested fixed-point algorithm in Rust (1987) to estimate the cost parameters (conditioning on discount factor) by maximizing the likelihood of observed choices. As shown in the literature (e.g., Pesendorfer and Schmidt-Dengler (2008) and Miessi Sanches et al. (2016)), the ordinary least-squares estimator, as employed in Komarova

et al. (2018) to minimize the Euclidean norm of the value function, is asymptotically equivalent to the maximum likelihood estimator used by us. In particular, given the model can be identified, the set of parameters that leads to the unique maximum likelihood function of observed choices also produces a zero Euclidean norm in Komarova et al. (2018) in the asymptotic limit. Thus, given our model can be identified from observed data, the maximum likelihood based criterion is asymptotically equivalent to the ordinary least-squares estimator in Komarova et al. (2018) (see e.g., Pesendorfer and Schmidt-Dengler (2008) and Miessi Sanches et al. (2016)). We also verify the full rank condition of matrix B in Komarova et al. (2018) holds by simulation with the estimated parameters[2].

First, the arrival and departure rates, as well as the ED and ICU capacities, are estimated directly from data – outside of the structural model. We estimate the ED arrival rates $\lambda_{Q,i}$ and maximum arrival number M_{A_i} for $i \in \{l, h\}$ using the average and maximum number of arrivals to the ED for the two classes in each period. We estimate the ICU external arrival rate λ_E using the average number of patients admitted to ICU in each time slot who are not included in our low and high severity ED groups. The departure probability μ_I is estimated as the ratio of total number of departures to the total periods of ICU stay across all ICU patients. The ICU capacity B is set to be the maximum number of patients in the ICU observed from data. It includes both medical and surgical, emergency and elective patients, to reflect the true maximum ICU occupancy. We also tested other choices of B to show the robustness of our estimation results.

Note that our data captures the number of patients admitted to the hospital from the ED, but does not include any patients who are discharged from the ED (e.g. to home or to a skilled nursing facility), but who inevitably utilize ED resources. Thus, it is difficult to accurately determine the maximum number of admitted patients allowable in the ED, i.e. the ED capacity in our model. Given this challenge, we set ED capacities Q_i using the

[2]With the estimated parameters, we compute the choice probabilities of each state and action by Proposition 1. Then, we approximate the conditional expectations in Tables A and B of Komarova et al. (2018) by simulation for each state and action. For each conditional expectation involved, we use 100 simulation trials with each run having the same number of periods as in the data plus an additional three-month warm up period. Finally, we compute the matrix $\mathbf{B}(\beta_0)$ in Theorem 1 of Komarova et al. (2018) with the simulated values for its components and verify it has full rank for all hospitals in our study.

following heuristic:

$$Q_i = M_{Q_i} + \left\lfloor \sqrt{M_{A_i}} \right\rfloor, \tag{3.15}$$

where M_{Q_i} is the maximum number of patients in the ED observed in the data; M_{A_i} is the maximum number of arrivals in each period; and $\lfloor \cdot \rfloor$ denotes the floor function. We introduce the square root term $\sqrt{M_{A_i}}$ as a "safety buffer" to ensure we have ample ED capacity to avoid balking upon arrival to the ED, as patients are rarely turned away from the hospital at this stage. For appropriately loaded queueing systems, it is well known that stochastic fluctuations of the queue length are on the order of the square root of the average offered load (see e.g., Halfin and Whitt (1981)). In our setting, we could approximate the average offered load by the arrival rates $\lambda_{Q,i}$ since the ED waiting time is generally very short. We take a more conservative approach and use the square root of the maximum number of arrivals M_{A_i} as the "safety buffer". We verify by simulation that the ED rarely reaches its full capacity (3.15) in our structural model; thus, while the ED has a finite capacity, patients rarely balk. We also find that the choice probabilities are very robust to alternative specifications of the ED capacity, which is not the case when varying the ICU capacity B. This suggests that our structural model primarily captures the interplay between ICU congestion and the importance the hospital places on intertemporal externalities when making admission decisions.

The remaining parameters – the discount factor, waiting and non-ICU admission costs for the two classes, i.e., $\theta = \{\beta, c_{r,l}, c_{r,h}, c_{w,l}, c_{w,h}\}$ – are estimated within the structural model using the observed states and actions. Given the observed state and action sequences $\{s_t, d_t\}$ for $t = 1, 2, ..., T$, the likelihood for a fixed set of parameters, θ, is given by

$$l^f(s_1, ..., s_T, d_1, ..., d_T|\theta) = \prod_{t=1}^{T} f(d_t|s_t, \theta) g(s_{t+1}|\varphi(s_t, d_t)), \tag{3.16}$$

where $f(d_t|s_t, \theta)$ denotes the choice probability in (3.9) given parameter θ. The state transition probability $g(s_{t+1}|\varphi(s_t, d_t))$ is explicitly given in the Appendix C.2.1. The likelihood l^f can be decomposed into two parts:

$$l^f(s_1, ..., s_T, d_1, ..., d_T|\theta) = l^d(s_1, ..., s_T, d_1, ..., d_T|\theta) \cdot l^s(s_1, ..., s_T, d_1, ..., d_T),$$

where l^d is the part of l^f associated with the choice probabilities, given by

$$l^d(s_1, ..., s_T, d_1, ..., d_T|\theta) = \prod_{t=1}^{T} f(d_t|s_t, \theta), \tag{3.17}$$

and l^s is the part of l^f from the state transition, i.e.,

$$l^s(s_1, ..., s_T, d_1, ..., d_T|\theta) = \prod_{t=1}^{T} g(s_{t+1}|\varphi(s_t, d_t)).$$

We see that the structural parameter θ is only involved in the likelihood function l^d for the choice probabilities, but not the likelihood function l^s for the state transitions, which only depends on the arrival and departure rates.

Our estimation approach is based on the nested fixed point algorithm in Rust (1987). The estimation procedure consists of two loops: The "inner" loop computes the function $\tilde{V}$ for a fixed θ, and the "outer" loop searches for the value of θ that maximizes the log-likelihood $\ln l^f$ in (3.16). Since the partial likelihood l^d in (3.17) is the only part in l^f that involves θ, maximizing the full log-likelihood is equivalent to finding the value of θ that maximizes $\ln l^d$. In the "inner" loop, the unknown function $\tilde{V}$ is computed by value iteration on the functional equation (3.11). In the "outer" loop, we use a gradient descent algorithm to find the optimal parameter θ. To reduce the computational burden, we restrict the potential values of the discount factor to a discrete grid $\beta = \{0.1, 0.2, ..., 0.9\}$. While coarse, this discrete grid is granular enough to measure how the hospital internalizes the intertemporal externalities, and enables us to determine whether they are more cognizant of the near versus longer-term when making decisions. To summarize, for each candidate value of β, we estimate the cost parameters $\{c_{r,l}, c_{r,h}, c_{w,l}, c_{w,h}\}$ that maximizes the log-likelihood $\ln l^d$. Then, we choose the discount factor and its associated cost estimates that lead to the largest likelihood among all candidate β.

With the estimated $\hat{\theta}$, the standard deviation of the parameters are computed as

$$\mathrm{Sd}(\hat{\theta}_i) = 1/\sqrt{\mathcal{I}(\hat{\theta})_{ii}},$$

where $\mathcal{I}(\hat{\theta})$ is the Fisher's information matrix

$$\mathcal{I}(\hat{\theta}) = \mathsf{E}\left[\frac{\partial \ln l^d(\hat{\theta})}{\partial \hat{\theta}} \left(\frac{\partial \ln l^d(\hat{\theta})}{\partial \hat{\theta}}\right)^{\top}\right].$$

To examine the proportion of variation explained by our structural model, we compute the McFadden's pseudo R-squared as

$$\text{Pseudo } R^2 = 1 - \frac{\ln l^d(\hat{\theta})}{\ln l^{null}}, \tag{3.18}$$

where l^{null} is the "null" likelihood from a multinomial logistic regression model with only an intercept term and hospital fixed-effects, i.e., the action probabilities do not depend on system states (See Appendix C.1 for details of the null model).

3.5 Estimation Results

We present the estimation results from the structural model in this section. Before we discuss the detailed results, we first describe how we apply the structural model to the data in our study setting and provide some preliminary findings using the system summary statistics calculated from the data and the parameters estimated outside the structural model.

We define each period to be a two hour time interval. This granularity provides a reasonable amount of time for transferring the patient from one unit to the next after the admission request is issued. We discuss this choice of time interval in Appendix C.3.

Recall that only 12% of all ED patients are admitted to the ICU; there are some patients (e.g. those with very low LAPS2 scores) who are likely to be admitted to non-ICU units (e.g., ward) regardless of ICU bed availability. Thus, in order to understand the impact of intertemporal externalities on the ICU admission decision, we must identify a group of high severity patients with sufficiently high likelihood of ICU admission. Additionally, we require enough observations in the high severity class to effectively estimate the non-ICU admission and waiting costs, $c_{r,h}$ and $c_{w,h}$, as well as to be able to identify the discount factor. We partition the ED patients into two classes by their LAPS2 score, as it has the highest correlation with the ICU admission decision among all severity scores. We define the low severity class as patients with LAPS2 score in the range of $[0, 113]$, corresponding to those below the 85th percentile of the LAPS2 score distribution, and the high severity class as patients with LAPS2 score in the range $(113, 294]$, corresponding to those above

the 85th percentile. We note that this means any medical patient admitted via the ED is included in our structural model as either a low or high severity patient. While we expect the ICU admission rate for low severity patients to be lower, some of these patients will be admitted to the ICU possibly due to factors that are unobservable in the data, thereby introducing intertemporal externalities on the system. Indeed, we find that while 8% of low severity patients are admitted to the ICU (compared to 34% of high severity patients), they contribute more (58%) to the total number of medical patients admitted from ED than the high severity class (42%). This highlights the importance of including the costs of both severity classes in our model, instead of just the high severity class. If we ignored the low severity patients, we would significantly underestimate the impact of the ICU admission decision on future patients.

Recall that our structural model considers the admission decisions observed at each hospital. Within each hospital, multiple physicians and administrators cover the ED and ICU over the study period. Due to limitations in data availability, we only know the hospital where the decision is made, but have no information on any individuals or specific system constraints which potentially impact the decision. Thus, we can only estimate our model at the hospital level even though there may be many different factors and systems involved in the decision making process in practice. Therefore, the estimation results will reflect the 'average' behavior within each hospital. We note that the hospital level result will tend to underestimate the actual variation in the systems' behaviors, as it ignores the potential heterogeneity among systems (e.g. night versus day, weekday versus weekend, ability to flex beds, etc.) within the same hospital. Thus, if we see large heterogeneity across hospitals, this would imply there is likely even larger heterogeneity when considering different system level factors, which we do not observe in the data.

In Table 3.2, we provide the system summary statistics of each hospital, including the ED and ICU capacities, average ICU occupancy, arrival and departure rates, and overall ICU admission probabilities for two patient classes. The ICU admission probability is the proportion of ED patients who are eventually admitted to the ICU regardless of their waiting time. As we already restrict our final study cohort to the patients whose next unit is either the ICU or ward (including the TCU if the hospital has one), the admission probability can

be computed by $N_{ICU}/(N_{ICU}+N_{nonICU})$, where N_{ICU} (N_{nonICU}) denotes the total number of patients that are admitted to ICU (non-ICU units). All the statistics in Table 3.2 are estimated directly from data outside of the structural model.

Table 3.2: System summary statistics by hospital

Hosp	Q_l	Q_h	B	ICUOccu	$\lambda_{Q,l}$	$\lambda_{Q,h}$	λ_E	μ_I	$\mathsf{Pr}(a_l)$	$\mathsf{Pr}(a_h)$
1	11	5	21	0.67	1.230	0.221	0.252	0.035	0.12	0.41
2	14	7	26	0.76	1.429	0.227	0.268	0.026	0.11	0.36
3	8	4	12	0.49	0.859	0.146	0.101	0.030	0.05	0.22
4	13	7	31	0.71	1.529	0.327	0.519	0.031	0.05	0.29
5	7	3	11	0.65	0.633	0.098	0.108	0.030	0.11	0.41
6	9	5	21	0.58	1.123	0.208	0.278	0.036	0.08	0.36
7	7	4	11	0.71	0.583	0.111	0.188	0.030	0.05	0.18
8	10	6	16	0.67	0.868	0.185	0.158	0.033	0.14	0.40
9	15	5	22	0.71	1.807	0.264	0.354	0.036	0.07	0.31
10	8	4	12	0.52	0.714	0.092	0.228	0.048	0.06	0.28
11	6	4	7	0.55	0.284	0.045	0.056	0.029	0.12	0.46
12	12	6	24	0.69	1.256	0.173	0.299	0.026	0.06	0.33
13	9	4	16	0.50	0.817	0.182	0.110	0.028	0.07	0.30
14	11	6	36	0.72	1.332	0.257	0.493	0.026	0.07	0.33
15	8	4	16	0.43	0.897	0.111	0.114	0.031	0.07	0.31
16	8	4	13	0.44	0.752	0.133	0.108	0.033	0.06	0.26
17	7	4	9	0.62	0.418	0.099	0.066	0.024	0.09	0.32
18	9	6	32	0.58	0.796	0.122	0.557	0.034	0.06	0.26
19	8	4	25	0.46	0.966	0.170	0.196	0.031	0.10	0.37
20	8	4	11	0.34	0.545	0.087	0.065	0.035	0.07	0.29
21	8	4	13	0.68	0.451	0.076	0.139	0.025	0.10	0.44
22	8	4	16	0.62	0.422	0.090	0.182	0.028	0.15	0.45

System summary statistics for each hospital: Q_i for $i \in \{l,h\}$ is the ED capacity for the two classes of patients; B is the ICU capacity; ICUOccu is the average ICU occupancy level; $\lambda_{Q,i}$ for $i \in \{l,h\}$ is the ED arrival rate; λ_E is the external arrival rate to ICU; μ_I is the ICU departure rate; $\mathsf{Pr}(a_i)$ for $i \in \{l,h\}$ is the overall admission probability for the ED patients.

We note the following observations from Table 3.2. First, the ICUs in the hospitals are generally congested. The average ICU occupancy in most hospitals is higher than 50%. For some, this number is even higher than 70%. Second, the ICU admission probability for the high severity class is above 30% for most hospitals, and is usually three or four times larger than that for the low severity class. This implies the costs parameters of the two classes should be very different, which is indeed captured in our structural model. Finally, we see the hospitals have very different sizes, work loads, and admission behaviors. For example, large hospitals have more than 30 beds in their ICU, while small hospitals have fewer than

10. Additionally, the ICU admission probabilities can be very different across hospitals even for the same severity class. With such large heterogeneity in the system statistics, hospitals may also behave very differently when they make admission decisions. We will quantify this heterogeneity with empirical evidence from our structural model.

3.5.1 Intertemporal Externalities

3.5.1.1 Main Estimation Results

In this section, we provide the estimation results of our structural model, i.e., the estimated discount factor and costs. Before estimating the model for each hospital separately, we first show the estimation results for all hospitals combined. That is, we estimate one set of parameters that maximizes the sum of log-likelihood from all hospitals. The results are summarized in the table below.

Table 3.3: Estimation results of structural model: All hospitals combined ($N = 154,140$ hospital-periods)

Discount factor	Low Severity		High Severity		
$\hat{\beta}$	$\hat{c}_{w,l}$	$\hat{c}_{r,l}$	$\hat{c}_{w,h}$	$\hat{c}_{r,h}$	R^2
0.3^{***}	0.071^{***}	-1.950^{***}	0.932^{***}	-0.671^{***}	0.14
(0.003)	(0.007)	(0.007)	(0.015)	(0.012)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

The estimated discount factor is $\hat{\beta} = 0.3$, and is significantly different from adjacent levels 0.2 and 0.4 at the 0.1% level. The estimated waiting costs $\hat{c}_{w,l}$ and $\hat{c}_{w,h}$ are both significantly positive, while the non-ICU admission costs $\hat{c}_{r,l}$ and $\hat{c}_{r,h}$ are both significantly negative. The McFadden's pseudo R^2 from the structural model is 0.14, which is comparable to the level ($R^2 = 0.16$) from the comprehensive multinomial logit regression in Table C.1.

At first glance, the estimated $\hat{\beta}$ is quite surprising. In most of the empirical literature, the discount factor is assumed to be relatively large, e.g., 0.90 or 0.95. However, we see here the estimated $\hat{\beta}$ is much smaller than these levels. This provides additional evidence that the level of discount factor may vary dramatically in different empirical settings (Frederick et al. (2002)). Thus, it is crucial to identify the discount factor using real data instead of assuming a pre-specified value. In our model, the relatively small value of $\hat{\beta}$ implies the

hospitals are not very forward-looking when making ICU admission decisions. Given each period in our model is a two hour interval, the result suggests that the hospitals barely consider the impact of their decisions on the system beyond the next six hours (after three 2-hour periods, $0.3^3 \approx 0.03$). With 12-hour shifts at KPNC, this suggests that while the hospitals indeed account for the future when making decisions, they mostly internalize the impact of their decisions on the system state within their own shifts and/or slightly into the immediately following shift.

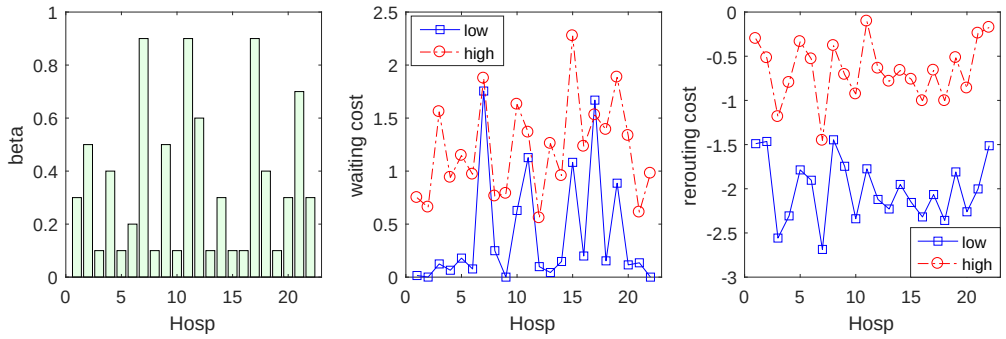
Unlike the waiting costs, which are constrained to be non-negative, the non-ICU admission costs are negative and significant for both the low and high severity classes. The negative non-ICU admission costs suggest that it is more desirable to admit a large fraction of patients to the ward. Indeed, this seems reasonable for low severity patients for whom ICU care is expensive, and (likely) unnecessary. This turns out to also be the case for some of the patients in the high severity group, which includes all patients with a LAPS2 score higher than the 85th percentile. This is also supported by the observed admission probability $\mathsf{Pr}(a_h)$ in Table 3.2: The admission probability for the high severity class is less than 50% in all hospitals, which means that the majority of these patients are admitted to the ward.

Finally, both the waiting and non-ICU admission costs are significantly higher for the high severity class than for the low severity class. As expected, it is more costly (financially, operationally, and clinically) for the high severity patients to wait in the ED, i.e., $\hat{c}_{w,h} > \hat{c}_{w,l} > 0$, and the low severity class on average benefits more from being admitted to the ward, i.e., $\hat{c}_{r,l} < \hat{c}_{r,h} < 0$. The apparent differences in cost parameters highlights the importance of differentiating between the two severity classes.

The estimated $\hat{\beta} = 0.3$ represents the average behavior across all hospitals; however, it does not provide information about potential differences across hospitals. If all hospitals behave quite similarly when internalizing the intertemporal externalities on admission decisions, most of them should have a discount factor close to the average level 0.3. On the other hand, if there is large heterogeneity in the discount factors, e.g., some hospitals with $\hat{\beta} = 0.9$ and some with $\hat{\beta} = 0.1$, this would imply that some hospitals are quite concerned with the longer-term while others focus more on the near-term. To address this question, we

estimate the structural parameters $\theta = \{\beta, c_{r,l}, c_{r,h}, c_{w,l}, c_{w,h}\}$ for each hospital individually. The results are summarized in Table 3.4. We also plot the estimated discount factors and costs in Figure 3.4.

Figure 3.4: Comparison of estimated discount factors and costs across the 22 individual hospitals.



Note: The left panel shows the estimated discount factor, the middle and right panels show the estimated waiting and non-ICU admission costs, respectively, for the low (blue solid line) and high (red dotted line) severity patients.

We see substantial heterogeneity in the estimated discount factors across hospitals. In particular, we have 13 out of 22 hospitals with relatively small estimated discount factors $\hat{\beta} \in \{0.1, 0.2, 0.3\}$, five with medium discount factors $\hat{\beta} \in \{0.4, 0.5, 0.6\}$, and the other four with relatively large discount factors $\hat{\beta} \in \{0.7, 0.8, 0.9\}$. All $\hat{\beta}$ estimates are significantly different from adjacent levels at the 0.1% level. Thus, ICU admission dynamics are very different across hospitals: Some of them have relatively small, near-term discount factors – focusing primarily on the individual patient in front of them – while others have relatively large, longer-term discount factors – accounting for the impact of their current decisions on the ability to treat other patients later. Such heterogeneity in the discount factor reflects the behavioral variation across hospitals.

As we have discussed before, there is much debate on how much hospitals internalize the intertemporal externalities when making admission decisions, and supporting evidence exists for both near versus longer term aspects of their behaviors. On one hand, physicians are trained to provide timely and appropriate care to their patients. On the other hand, hospitals need to manage the occupancy level in the often congested ICU to reserve enough

Table 3.4: Estimation results of structural model by individual hospital

Hosp	Num. of periods	Discount Factor $\hat{\beta}$	Low Severity $\hat{c}_{w,l}$	Low Severity $\hat{c}_{r,l}$	High Severity $\hat{c}_{w,h}$	High Severity $\hat{c}_{r,h}$	R^2
1	8,016	0.3^{***}	0.015	-1.490^{***}	0.749^{***}	-0.301^{***}	0.20
		(0.012)	(0.021)	(0.022)	(0.048)	(0.043)	
2	6,012	0.5^{***}	0.001	-1.465^{***}	0.659^{***}	-0.519^{***}	0.24
		(0.010)	(0.017)	(0.024)	(0.047)	(0.051)	
3	8,016	0.1^{***}	0.124^{*}	-2.558^{***}	1.560^{***}	-1.185^{***}	0.13
		(0.019)	(0.051)	(0.040)	(0.128)	(0.065)	
4	8,016	0.4^{***}	0.065^{**}	-2.307^{***}	0.940^{***}	-0.798^{***}	0.21
		(0.010)	(0.024)	(0.029)	(0.046)	(0.039)	
5	6,924	0.1^{***}	0.179^{***}	-1.786^{***}	1.147^{***}	-0.336^{***}	0.13
		(0.023)	(0.046)	(0.037)	(0.107)	(0.072)	
6	8,016	0.2^{***}	0.077^{*}	-1.904^{***}	0.970^{***}	-0.531^{***}	0.18
		(0.014)	(0.030)	(0.027)	(0.063)	(0.046)	
7	6,948	0.9^{***}	1.755^{***}	-2.687^{***}	1.876^{***}	-1.453^{***}	0.09
		(0.016)	(0.047)	(0.075)	(0.093)	(0.095)	
8	7,848	0.1^{***}	0.249^{***}	-1.446^{***}	0.764^{***}	-0.380^{***}	0.15
		(0.020)	(0.034)	(0.026)	(0.062)	(0.048)	
9	6,180	0.5^{***}	0.000	-1.745^{***}	0.789^{***}	-0.707^{***}	0.25
		(0.009)	(0.017)	(0.025)	(0.048)	(0.048)	
10	7,320	0.1^{***}	0.629^{***}	-2.339^{***}	1.630^{***}	-0.929^{***}	0.12
		(0.027)	(0.067)	(0.046)	(0.160)	(0.081)	
11	8,016	0.9^{***}	1.127^{***}	-1.774^{***}	1.366^{***}	-0.104	0.08
		(0.017)	(0.036)	(0.064)	(0.088)	(0.106)	
12	4,668	0.6^{***}	0.099^{***}	-2.121^{***}	0.558^{***}	-0.637^{***}	0.22
		(0.010)	(0.023)	(0.040)	(0.052)	(0.068)	
13	8,016	0.1^{***}	0.044	-2.229^{***}	1.260^{***}	-0.785^{***}	0.14
		(0.018)	(0.043)	(0.035)	(0.088)	(0.052)	
14	8,016	0.3^{***}	0.147^{***}	-1.951^{***}	0.956^{***}	-0.662^{***}	0.19
		(0.012)	(0.027)	(0.027)	(0.054)	(0.042)	
15	6,912	0.1^{**}	1.082^{***}	-2.155^{***}	2.275^{***}	-0.762^{***}	0.14
		(0.033)	(0.074)	(0.042)	(0.195)	(0.075)	
16	8,016	0.1^{***}	0.199^{***}	-2.319^{***}	1.233^{***}	-1.005^{***}	0.12
		(0.020)	(0.050)	(0.039)	(0.108)	(0.063)	
17	6,588	0.9^{***}	1.669^{***}	-2.065^{***}	1.530^{***}	-0.660^{***}	0.09
		(0.021)	(0.049)	(0.067)	(0.077)	(0.085)	
18	8,016	0.4^{***}	0.154^{***}	-2.358^{***}	1.391^{***}	-1.005^{***}	0.14
		(0.013)	(0.033)	(0.039)	(0.099)	(0.070)	
19	6,576	0.1^{***}	0.886^{***}	-1.809^{***}	1.885^{***}	-0.517^{***}	0.15
		(0.029)	(0.057)	(0.035)	(0.126)	(0.058)	
20	8,004	0.3^{***}	0.116^{**}	-2.260^{***}	1.335^{***}	-0.862^{***}	0.10
		(0.017)	(0.041)	(0.042)	(0.115)	(0.078)	
21	4,008	0.7^{***}	0.135^{***}	-2.002^{***}	0.612^{***}	-0.240^{*}	0.13
		(0.012)	(0.030)	(0.064)	(0.071)	(0.111)	
22	4,008	0.3^{***}	0.000	-1.513^{***}	0.979^{***}	-0.175	0.11
		(0.025)	(0.047)	(0.051)	(0.116)	(0.099)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. The second column reports the number of periods (two-hour snapshots) in each hospital, and the last column provides the McFadden's pseudo R^2.

capacity for future, perhaps more severe patients. Moreover, effectively accounting for the future can be difficult as members of the hospitals (e.g. physicians and administrators) are rarely trained to think about the complex system dynamics that arise in hospitals. Our findings show that both immediate and future considerations can influence the hospital's behavior, and the overall effect can vary substantially across hospitals. This finding reveals an important behavioral perspective of the decision-making process for ICU admission and provides additional explanation for observed practice variation, which has received much attention from medical professionals in recent literature (Westert et al., 2018; Corallo et al., 2014). Indeed, a number of studies suggest that identifying the patients that will benefit from ICU care is highly subjective and depends on a physician's own training and experience (Fisher et al. (2004), Mullan (2004), O'Connor et al. (2004), Chen et al. (2012)); our work suggests that some of the variation observed in practice may also be due to variations in how hospitals internalize intertemporal externalities.

The pattern of estimated non-ICU admission and waiting costs in Table 3.4 are very similar to that for all hospitals combined. Most of the waiting costs are significant and positive, including 17 out of 22 estimates for $\hat{c}_{w,l}$ and all estimates for $\hat{c}_{w,h}$. The non-ICU admission costs are all significant and negative, except for $\hat{c}_{r,h}$ of Hospital 11 and 22. In most cases, the non-ICU admission and waiting costs for high severity class are significantly higher than those for low severity class. The costs estimates vary substantially across hospitals. Recall these parameters represent the average costs measured relative to the ICU admission decision in each hospital. Such variation suggests that there is large heterogeneity across hospitals in their medical resources and the degree of severity of their patient population.

Using the estimated discount factors and cost parameters, we can show our structural model is able to capture the relationship between ICU admission probability and ICU occupancy observed in Figure 3.1. As an illustrative example, we consider a representative ED state with $n^E_{l,t} = \lfloor Q_l/2 \rfloor$ and $n^E_{h,t} = 1$, i.e., several low severity class patients and one high severity patient[3]. We focus on the ICU admission probability for the high severity class patient, who is more likely to be admitted to the ICU. We compute the reduction in her

[3]We find that the admission probability is practically insensitive to the particular choice of $n^E_{l,t}$.

admission probability as the ICU occupancy moves from half full to almost full; that is,

$$\text{AdmDrop} = \mathsf{Pr}\left(a_{h,t} = 1 | n_t^I = \lfloor B/2 \rfloor\right) - \mathsf{Pr}\left(a_{h,t} = 1 | n_t^I = B - 1\right). \tag{3.19}$$

We also compute the relative drop as

$$\text{Rel.AdmDrop} = \frac{\mathsf{Pr}\left(a_{h,t} = 1 | n_t^I = \lfloor B/2 \rfloor\right) - \mathsf{Pr}\left(a_{h,t} = 1 | n_t^I = B - 1\right)}{\mathsf{Pr}\left(a_{h,t} = 1 | n_t^I = \lfloor B/2 \rfloor\right)}. \tag{3.20}$$

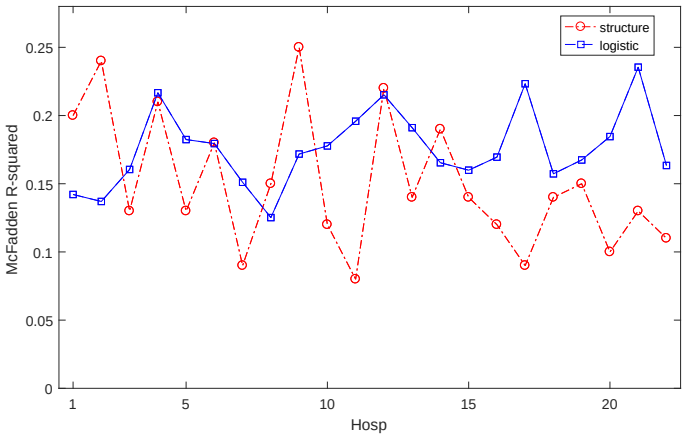
We find the ICU state has a substantial impact on the ICU admission decision. As summarized in Table C.4 in Appendix C.2.3, on average, the hospitals see a 11% relative drop in the ICU admission probability as the ICU becomes almost full. Such drop can be as large as 17% for some hospitals. We also find the relative drop and estimated $\hat{\beta}$ are highly correlated. Their correlation is 0.557, which is statistically significant at the 1% level. Thus, hospitals with larger discount factors are indeed more sensitive to the ICU occupancy level when admitting patients to the ICU. This relationship confirms that the discount factors in our structural model capture how the hospitals balance near versus longer term considerations in ICU admission decisions.

3.5.1.2 Goodness-of-Fit for Structural Model

While the structural model can help bring behavioral insights into the hospitals' decisions, it is important to check whether the estimated model fits the data well. In this section, we show our structural model provides good estimates to both the hospitals' decisions and the system states.

First, we compare the explanatory power for the hospitals' decisions, as measured by the McFadden's pseudo R^2 in (3.18), from our structural model with that from the reduced-form multinomial logistic regression model in Appendix C.1. The results are shown in Figure 3.5. Our structural model has comparable or higher McFadden's pseudo R^2 to the comprehensive multinomial logistic model for most hospitals. The average McFadden's pseudo R^2 in the structural model is also close to the combined hospital multinomial logistic model (0.14 versus 0.16). The R^2 values are not very high in both models. This is not too surprising as the hospitals consider many factors that are not recorded in the data when making admission decisions; that is, the decisions appear very "noisy" in the data. We note that the multinomial

Figure 3.5: Comparison of McFadden's pseudo R^2 from structural and multinomial models



Note: The red dotted and blue solid lines report the McFadden's pseudo R^2 for each hospital from the structural model and the multinomial model, respectively.

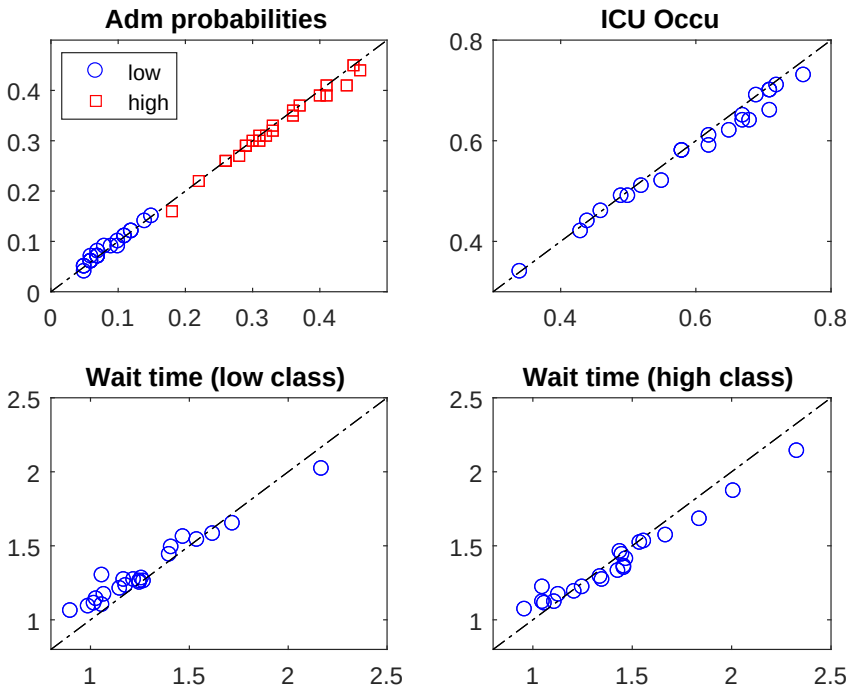
model contains most variables one could expect to influence admission decisions, including patient's characteristics (gender, age, and three severity scores), system states (ED and ICU occupancy, departures in previous period, and average ICU severity level), as well as various dummies for seasonality fixed effects (time of day, day of week, and month of the sample). On the other hand, our structural model is quite parsimonious with only five free parameters (discount factor and costs parameters), but is still able to explain similar (or more) variation in the data. Thus, our structural model appears to have reasonable explanatory power in capturing the hospitals' decisions. Of course, we acknowledge that there is still quite a bit of variation that our data cannot capture (e.g. perhaps due to availability of system-level interventions to increase capacity or stabilize and treat patients outside of the ICU.

With a small number of parameters, overfitting is unlikely to be an issue for our structural model. To further address goodness-of-fit, we divide the sample to first and second halves for each hospital. We then estimate the structure model with all hospitals combined using the first half sample, and do out-of-sample prediction on the second half sample. The McFadden's pseudo R^2 from the out-of-sample prediction is very similar to the in-sample estimation (0.14 vs 0.13). Moreover, it is close to the level estimated from the full sample (0.14) in Table 3.3.

Next, we show by simulation that our structural model produces system statistics close

to those observed in the data. As the arrival and departure rates are directly calibrated from data, we expect the average number of arrivals and departures in each period of our structural model to be close to that observed in the data. Therefore, we focus on other important system statistics including the average ICU occupancy, overall proportion admitted to the ICU (i.e., $N_{ICU}/(N_{ICU} + N_{nonICU})$), as well as the ED waiting times of each patient class. The statistics estimated from our structural model are averaged over 100 simulation runs. Each run contains the same number of periods as in the data plus an additional three-month warm up period to allow the system to reach steady-state. The warm up period is dropped when computing the system statistics. The comparison of the system statistics is shown in Figure 3.6.

Figure 3.6: Comparison of system statistics from structural model and real data



Note: The figure compares the system statistics simulated from the structural model (y-coordinate) and observed from the real data (x-coordinate).

In each panel, each point represents a hospital in our study; its x-coordinate (y-coordinate) corresponds to the observed (simulated) value of the system statistic. We plot the 45-degree line in each panel, which represents a perfect fit. As we can see, most points fall close to the

identity line, implying that our estimated structural model produces system statistics that are close to the observed data. Although our structural model is trained to fit the choice probabilities of the hospital's actions, it also leads to system dynamics that fit a number of observed metrics very well. This further supports its effectiveness in modeling the admission process for ED patients.

3.5.1.3 Heterogeneity in Discount Factors

We have seen in the previous section there is large heterogeneity in the discount factors across hospitals, i.e., some hospitals appear to be relatively focused on the near-term in making admission decisions, while others appear to consider more longer-term dynamics. Such heterogeneity in the perceived discout factor is important for understanding system-level admission decisions. There are many potential reasons for this heterogeneity, such as the different physicians and administrators in the hospital as well as system-level factors (e.g. the ability to flex capacity and/or the amount of demand from other sources) which differentially influence the occupancy challenges at each hospital. In this section, we look into the heterogeneity in discount factors in more depth. We identify possible reasons that explain the heterogeneity, and examine how the heterogeneity impacts system performance.

We start by computing the correlations between the estimated discount factors and other system statistics across the 22 hospitals. The results are shown in Table 3.5. We show the correlations between $\hat{\beta}$ and the observed ICU departure rate μ_I, average ICU occupancy level, average ED waiting time EDWait_i for $i \in \{l, h\}$, and the increase in waiting time for admitted patients due to ICU congestion, i.e., $\Delta\text{EDWait\_Adm}_i$. This increase is defined as the difference between average ED waiting time of patients admitted to the ICU when the occupancy level is below the 70th percentile versus when it is above the 95th percentile.

It is not obvious ex-ante in which direction ICU congestion impacts the hospitals' admission behaviors. On one hand, when the ICU is busy, hospitals have to be more judicious when making bed allocation decisions as access issues could substantially jeopardize quality of care, which suggests they should be more forward-looking when making decisions. For example, if the hospital expects the ICU will be highly congested, they may choose to save some ICU beds for future (potentially sicker) patients by delaying the admission of current

Table 3.5: Correlations between estimated discount factor, $\hat{\beta}$, and system statistics from data

μ_I	ICUOccu	EDWait$_l$	EDWait$_h$	ΔEDWait_Adm$_l$ with ICU congestion	ΔEDWait_Adm$_h$ with ICU congestion
-0.428^{*}	0.445^{*}	0.545^{**}	0.581^{**}	0.475^{*}	0.566^{**}

$^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$. Correlations between estimated discount factor and the observed ICU departure rate (μ_I), ICU occupancy (ICUOccu), average ED waiting time (EDWait$_i$), and increase in ED waiting time for admitted patients under ICU congestion (ΔEDWait_Adm$_i$).

ED patients. On the other hand, busier ICUs also tends to have more external arrivals, such as non-ED surgical patients. This introduces increased competing demand for ICU beds from external arrivals, which may reduce the hospital's motivation or ability to save ICU beds for future ED patients (the cohort for which our model captures) as the beds will likely be utilized by external arrivals. Our results suggest that, on average, the first effect dominates the second: hospitals with lower ICU departure rates or higher ICU occupancy levels tend to have larger discount factors and these hospitals tend to be more congested. One possible explanation is that the majority of the external arrivals ($> 70\%$) are surgical patients whose ICU beds are often reserved in advance to accommodate their post-surgery recovery, so hospitals may not necessarily account for the competing demand from external arrivals when making ICU admission decisions.

We also find that the discount factors are positively correlated with the average ED waiting time of both classes. This makes sense as hospitals who account more for the longer-term are more likely to save beds for future patients, thereby increasing the waiting time of some ED patients.

The discount factors are also positively correlated with the increase in ED waiting time for admitted patients when the ICU is congested. This is not surprising because hospitals with longer-term discounting behaviors tend to be more sensitive to ICU congestion when making admission decisions and are more likely to delay patients when the unit is congested. These results are consistent with the relationship between $\hat{\beta}$ and the relative admission probability drop (3.20) discussed at the end of Section 3.5.1.1. These results lend further support that our structural model is able to capture the intertemporal externalities on

admission decisions.

3.5.1.4 Robustness Check: Estimation with Stratified Sample

In the above estimation, we assume the model parameters are constant for the entire sample. To address potential seasonality issues, we also estimate the structural model with stratified sample based on flu vs non-flu seasons, as well as day and night periods. The flu season ranges from November to March, and the non-flu season includes the rest. The day periods include the twelve hours from 7AM to 7PM. We re-estimate the ED arrival rates $\lambda_{Q,l}$ and $\lambda_{Q,h}$, external ICU arrival rate λ_E, and ICU departure rate μ_I for each stratified sample. The ED and ICU capacities Q_l, Q_h, and B are set as those for the full sample in Table 3.2.

We first estimate the structural model for all hospitals combined with the stratified samples. The results are shown in Table 3.6. We can see the estimated parameters are close to those estimated by the full sample in Table 3.3. In particular, the estimated discount factor is 0.4 for the flu season and 0.2 for the non-flu season, which are close to $\hat{\beta} = 0.3$ of the full sample. Additionally, the discount factor is estimated to be 0.4 for the day sample, which is also close to that of the full sample. The discount factor is estimated to be lower ($\hat{\beta} = 0.1$) for the night sample. This can be explained as the night time has much lower arrival and departure rates, as well as fewer staffs than the day time. Thus the hospitals may be more near-term focused and admit patients more quickly compared with the day time.

We then estimate the structural model for each hospital separately with the stratified data. We find the heterogeneity in discount factors estimated from the stratified data is consistent with that from the full sample given in Table 3.4. In particular, the across-hospital correlations between the estimated $\hat{\beta}$ from stratified and full sample are significantly positive, which are given in Table 3.7. Importantly, the the main, high level takeaway from the structural model is robust to accounting for these potential temporal variations: 1) hospitals on average are not very forward-looking when making decisions; 2) there is substantial heterogeneity in the discount factors across hospitals.

Table 3.6: Estimation of structural model with stratified data: All hospitals combined

	Size	Discount factor	Low Severity		High Severity		
	N	$\hat{\beta}$	$\hat{c}_{w,l}$	$\hat{c}_{r,l}$	$\hat{c}_{w,h}$	$\hat{c}_{r,h}$	R^2
Flu	60,031	0.4^{***}	0.028^{***}	-1.936^{***}	0.755^{***}	-0.674^{***}	0.14
		(0.004)	(0.010)	(0.012)	(0.021)	(0.019)	
Non-flu	108,737	0.2^{***}	0.058^{***}	-1.960^{***}	1.086^{***}	-0.670^{***}	0.13
		(0.005)	(0.010)	(0.009)	(0.023)	(0.016)	
Day	84,372	0.4^{***}	0.001	-2.006^{***}	0.806^{***}	-0.655^{***}	0.13
		(0.004)	(0.009)	(0.011)	(0.020)	(0.018)	
Night	84,396	0.1^{***}	0.045^{***}	-1.875^{***}	0.971^{***}	-0.684^{***}	0.14
		(0.005)	(0.011)	(0.009)	(0.024)	(0.016)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table 3.7: Correlation between $\hat{\beta}$ from stratified and full sample

Sample	Flu	Non-flu	Day	Night
Correlation	0.87^{***}	0.67^{***}	0.67^{***}	0.46^{*}

$^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

3.6 Counterfactual Simulations

One of the most important advantages of using the structural estimation approach is the ability to conduct counterfactual studies via simulation. This allows us to evaluate the effect of different policies or interventions that cannot be directly observed in the data or through reduced-form regression analyses, which – at best – only provides local treatment effects. We consider operational and behavioral counterfactual studies. The counterfactual results can provide insights into the impact from various interventions on medical and economic outcomes and system performance metrics.

In the first operational counterfactual, we measure the impact of adding one bed in ICU while assuming the hospitals' behaviors remain unchanged. That is, we increase the ICU capacity by one in each hospital, and use the estimated structural model (discount factors and costs parameters) to predict the hospitals' decisions. This enables us to quantify the change in system statistics from adding one bed in ICU. Note that for small hospitals, adding a single ICU bed can increase capacity by up to 14%, which is a large enough change

that it is unreasonable to expect that the local, marginal effects captured by a reduced form regression analysis are sufficient to make projections on how this extra bed will impact the performance measures of interest. While increasing ICU capacity naturally reduces ICU congestion, such a change requires substantial capital investments (e.g. up to $1.1 million per year estimated from the $4302 daily expense in Franzini et al. (2011)). In the second operational study, we keep the ICU capacity unchanged and decrease external arrivals to the ICU by 5% or 10%. As more than 60% of external arrivals to ICU are surgical patients, a decrease in external arrival rates implies fewer surgical cases performed by the hospital. This can be very costly to implement, as surgical cases can contribute up to a half of hospitals' revenue (McDermott et al., 2017). Thus a 10% decrease in external arrival rates would translate to 5% loss in hospital's total revenue, which can be crucial as many hospitals in US are financially strained.

In the last counterfactual study, we focus on a behavioral change. We keep the ICU capacity and external arrival rate unchanged and consider the hypothetical situation where the hospital is made to incorporate longer-term impacts of their decisions when making their admission decisions. In particular, we quantify the impact on system statistics as the hospitals increase their discount factors from the current estimated levels to $\beta = 0.9$ without introducing any change in the medical resources and system workloads. We measure such impact relative to that from adding an ICU bed or decreasing external arrival rate.

Finally, while increasing the discount factor seems abstract in practice, we discuss potential ways for hospitals to implement such change. Moreover, we show the benefit of increasing discount factor can be largely obtained by a simple heuristic policy that mimics the actions from the structural model.

In all the counterfactual studies, the statistics are computed from the average of 100 simulation runs. Each run has the same number of periods as in the data plus a three-month warm up period. The system statistics we are interested in are related to the impact of high ICU congestion. In particular, we consider the probability of high ICU congestion, which is defined by

$$\mathsf{Pr}(\text{HighCgstn}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{n_t^I \geq B - 1\right\}; \tag{3.21}$$

i.e, the proportion of periods with only one or no empty bed(s) in the ICU. High ICU congestion is known to be associated with worse medical outcomes such as higher mortality, longer hospital length-of-stay, and higher risk from postoperative complications (Gattinoni et al. (2004), Hugonnet et al. (2007), and Gabler et al. (2013)); this may be due, for example, to reduced likelihood of ICU admission (Kim et al., 2015) or increased likelihood of demand-driven discharges (Kc and Terwiesch, 2012).

We also examine the probability of the ICU being full. This impacts the likelihood of external arrivals balking upon arrival because there are no ICU beds available.

$$\mathsf{Pr}(\text{Balk}|\text{External Arrival}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{n_t^I = B\right\}. \tag{3.22}$$

We estimate the absolute and relative impact of each intervention on the probability of high ICU congestion and the probability of external arrivals balking. Recall that each period in our structural model corresponds to a two hour interval. We provide a conservative estimate for the number of ICU patients who spend their ICU stay under highly congested states by

$$\text{Pats HighCgstn} = \mathsf{Pr}(\text{HighCgstn}) \times 365 \times 12 \times \mu_I \times (B-1).$$

The first three components provide an estimate for the total number of high congestion periods in a year. We divide this by the average LOS of each patient, $1/\mu_I$ periods, to estimate the number of patients each bed can serve during the high congestion periods in a year. Since there is at most one bed available during the high congestion periods, putting everything together gives an estimate for the number of patients that are exposed to high congestion periods. Using a similar argument, the number of external arrivals who balk is estimated by

$$\text{Pats Balk} = \mathsf{Pr}(\text{Balk}|\text{External Arrival}) \times 365 \times 12 \times \lambda_E, \tag{3.23}$$

where λ_E is the external arrival rate in each period.

3.6.1 Adding One ICU Bed

We start by studying the impact of adding one more bed in ICU. In particular, we focus on the impact on ICU congestion when keeping the arrival and service rates fixed. Due to

the size of the ICUs in our study, the addition of one bed has very limited impact on the average ICU occupancy level. The results are shown in Table 3.9. We see adding one bed in the ICU indeed leads to substantial reductions in ICU congestion. The relative drops in high ICU congestion and balking probabilities are greater than 30% in most cases. While not reported here, the standard deviations computed from simulation confirm all the differences are statistically significant at the 1% level. For some hospitals, such reduction translates to 20 fewer days and 50 fewer patients exposed to high ICU congestion, as well as 10 more external arrivals admitted in each year. As the patients who require ICU care are usually very unstable, the observed effects have the potential to lead to large improvements for their medical outcomes.

Next, we look into how the benefit of adding one bed in ICU varies across hospitals. Table 3.8 reports the correlations of the three measures (ΔDays HighCgstn, ΔPats HighCgstn, and ΔPats Balking in Table 3.9) with the average ICU occupancy levels and estimated discount factors across 22 hospitals. All three effect measures are positively correlated with both the ICU occupancy levels and discount factors. This relationship also holds when we perform a simple linear regression for the three effect measures with both estimated discount factor and average ICU occupancy level as independent variables. The results suggest the benefit from adding one ICU bed is more significant for ICUs with higher occupancy and/or for hospitals that are more longer-term looking.

Table 3.8: Correlations between impact of one bed on ICU congestion and ICU occupancy or $\hat{\beta}$

	ΔDays HighCgstn	ΔPats HighCgstn	ΔPats Balk
ICUOccu	0.354	0.737^{**}	0.732^{**}
$\hat{\beta}$	0.700^{***}	0.520^{*}	0.602^{**}

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

3.6.2 Decreasing External Arrivals to the ICU

We next study the impact of decreasing the external arrival rate to the ICU. We keep the ICU capacity unchanged and decrease the external arrival rate λ_E by 5% and 10%. Then

Table 3.9: Counterfactual estimates of impact when adding one bed in ICU

Hosp	$\hat{\beta}$	$\Delta\mathsf{Pr}$(HighCgstn) (in % points)	Relative $\Delta\mathsf{Pr}$(HighCgstn)	$\Delta\mathsf{Pr}$(Balk) (in % points)	Relative $\Delta\mathsf{Pr}$(Balk)	ΔDays HighCgstn (in # days)	ΔPats HighCgstn (in # patients)	ΔPats Balk (in # patients)
1	0.3	1.72	0.35	0.61	0.39	6.27	52.26	6.73
2	0.5	1.86	0.25	0.72	0.27	6.78	52.78	8.43
3	0.1	1.66	0.54	0.48	0.56	6.06	23.82	2.11
4	0.4	1.16	0.31	0.38	0.31	4.22	47.08	8.63
5	0.1	5.13	0.38	1.87	0.39	18.74	68.04	8.84
6	0.2	0.75	0.42	0.20	0.37	2.73	23.77	2.42
7	0.9	6.36	0.35	2.57	0.38	23.23	84.04	21.20
8	0.1	2.99	0.38	1.14	0.42	10.90	64.88	7.92
9	0.5	2.36	0.32	0.88	0.35	8.60	78.61	13.57
10	0.1	2.25	0.53	0.76	0.60	8.23	51.80	7.56
11	0.9	7.39	0.49	2.70	0.54	26.98	56.65	6.61
12	0.6	1.66	0.30	0.59	0.31	6.06	43.17	7.67
13	0.1	0.64	0.51	0.20	0.57	2.34	11.65	0.95
14	0.3	0.65	0.24	0.24	0.28	2.36	25.90	5.26
15	0.1	0.23	0.60	0.06	0.61	0.85	4.77	0.30
16	0.1	0.69	0.56	0.20	0.56	2.52	12.07	0.95
17	0.9	5.90	0.40	2.28	0.45	21.54	50.58	6.60
18	0.4	0.18	0.44	0.05	0.42	0.65	8.24	1.16
19	0.1	0.02	0.41	0.01	0.56	0.08	0.75	0.09
20	0.3	0.26	0.63	0.05	0.52	0.94	3.91	0.15
21	0.7	4.55	0.40	1.97	0.47	16.61	59.02	11.97
22	0.3	2.26	0.39	0.95	0.46	8.27	42.31	7.59

Note: Counterfactual simulation result from adding one ICU bed: The third and fourth columns report the absolute and relative drops in high congestion probability in (3.21); the fifth and sixth columns report the absolute and relative drops in balking probability in (3.22). The last three columns report the equivalent numbers of days and patients affected in a year.

we use counterfactual simulation to estimate the reductions in the probabilities of high ICU congestion (3.21) and balking (3.22).

The estimation results on ICU congestion reduction from decreasing λ_E are reported in Table 3.10 (5% decrease). The results for a 10% decrease are included in Table C.5 of Appendix C.2.3. Decreasing external arrival rates can significantly reduce ICU congestion. The average relative drop in high ICU congestion probability is 19% for when λ_E decreases by 5%, and 34% when λ_E decreases by 10%. For some hospitals, such reduction translates to 10 fewer days and 50 fewer patients exposed to high ICU congestion per year. The reduction in numbers of external arrivals who balk is also substantial: a 5% (resp. 10%) decrease in external arrival rate can lead to 10 (resp. 20) fewer patients who cannot be admitted due to full ICU. This reduction is due to both a smaller λ_E that reduces the total number of external arrivals as well as a lower balking probability Pr(Balk|External Arrival) as the ICU is less congested.

3.6.3 Increasing the Discount Factor

The operational interventions discussed above can significantly reduce congestion, but are very costly to implement and may lead to substantial financial strain. We now consider the potential impact of modifying the hospital's behavior. As discussed earlier, hospitals are typically not trained to manage their patients with a system-level view; our estimation results suggest they tend to be more near-term focused when making ICU admission decisions. That said, we found that some hospitals were more forward looking. Hence, we wish to estimate what might happen if the hospitals were provided with the right support and information to enable them to account more for the longer-term when making admissions decisions. While it might not be possible to alter the discount factor of all hospitals, this experiment provides some insight into what might be possible if the discount factor could be nudged higher as well as what could happen if the hospitals had more sophisticated strategies to deal with congestion. Additionally, we will also discuss some practical ways to potentially achieve these savings.

The specific counterfactual which we consider is to keep the ICU capacity and external

Table 3.10: Counterfactual estimates of impact when λ_E decreases by 5%

Hosp	$\hat{\beta}$	$\Delta\mathsf{Pr}$(HighCgstn) (in % points)	Relative $\Delta\mathsf{Pr}$(HighCgstn)	$\Delta\mathsf{Pr}$(Balk) (in % points)	Relative $\Delta\mathsf{Pr}$(Balk)	ΔDays HighCgstn (in # days)	ΔPats HighCgstn (in # patients)	ΔPats Balk (in # patients)
1	0.3	0.79	0.16	0.32	0.20	2.90	24.15	4.19
2	0.5	1.02	0.14	0.34	0.13	3.72	28.96	5.27
3	0.1	0.70	0.23	0.21	0.24	2.55	10.02	1.07
4	0.4	1.06	0.29	0.35	0.29	3.88	43.28	8.94
5	0.1	1.29	0.10	0.57	0.12	4.71	17.11	3.71
6	0.2	0.38	0.21	0.12	0.23	1.38	12.07	1.76
7	0.9	2.59	0.14	1.09	0.16	9.47	34.24	11.30
8	0.1	0.96	0.12	0.36	0.13	3.52	20.95	3.30
9	0.5	1.54	0.21	0.58	0.23	5.61	51.25	10.44
10	0.1	0.88	0.21	0.28	0.22	3.22	20.30	3.25
11	0.9	1.71	0.11	0.72	0.15	6.23	13.08	2.29
12	0.6	1.16	0.21	0.43	0.22	4.25	30.24	6.59
13	0.1	0.27	0.22	0.07	0.21	1.00	4.98	0.42
14	0.3	0.93	0.35	0.32	0.36	3.40	37.44	7.52
15	0.1	0.07	0.19	0.00	0.00	0.27	1.50	0.03
16	0.1	0.24	0.19	0.07	0.20	0.87	4.14	0.41
17	0.9	1.21	0.08	0.69	0.14	4.41	10.35	2.64
18	0.4	0.17	0.42	0.05	0.48	0.62	7.79	1.40
19	0.1	0.01	0.09	0.01	0.31	0.02	0.17	0.05
20	0.3	0.02	0.04	0.01	0.05	0.06	0.27	0.03
21	0.7	1.40	0.12	0.82	0.19	5.09	18.10	6.01
22	0.3	1.43	0.24	0.69	0.33	5.22	26.72	6.02

Note: Counterfactual simulation result from decreasing external arrival rates λ_E by 5%: The third and fourth columns report the absolute and relative drops in high congestion probability in (3.21); the fifth and sixth columns report the absolute and relative drops in balking probability in (3.22). The last three columns report the equivalent numbers of days and patients affected in a year.

arrival unchanged, and increase the discount factors from the current estimated levels to 0.9. Practically, this corresponds to the hospital caring about the decision costs and system states over the next one to two days ($0.9^{12} \approx 0.28$), rather than the next six hours, which was the average behavior implied by Table 3.3. Note that for some hospitals, the estimated $\hat{\beta}$ is already 0.9, so this counterfactual has no impact on their system dynamics.

We summarize the impact of increasing β on ICU congestion in Table 3.12. We see increasing β alone can have substantial effect on reducing ICU congestion. For some hospitals, the relative drops in the probability of high ICU congestion and balking are more than 20% as the hospital shifts towards longer-term discounting. This translates to significant improvement in terms of the frequency of high ICU congestion as well balking of external arrivals.

To facilitate comparisons, in Table 3.11, we summarize the reductions in ICU congestion for Hospitals 1, 2, 5, 8, 9, and 14 from different types of interventions: increasing $\hat{\beta}$ to 0.9, adding one ICU bed, and decreasing the external arrival rate by 5% and 10%. We notice that for these hospitals, the impact of increasing β can be comparable to adding one ICU bed or decreasing external arrivals by 5 – 10%. This highlights the importance of understanding how the hospitals internalize the intertemporal externalities on ICU admission decisions, and reveals a potential approach for hospitals to reduce ICU congestion, i.e., providing their hospitals more tools and skills to manage and react to congestion.

Table 3.11: Comparison of effects from different counterfactual interventions (select hospitals)

Hosp	ΔDays HighCgstn				ΔPats HighCgstn				ΔPats Balk			
	$\hat{\beta}\uparrow$	$B\uparrow$	$\lambda_E\downarrow$	$\lambda_E\downarrow\downarrow$	$\hat{\beta}\uparrow$	$B\uparrow$	$\lambda_E\downarrow$	$\lambda_E\downarrow\downarrow$	$\hat{\beta}\uparrow$	$B\uparrow$	$\lambda_E\downarrow$	$\lambda_E\downarrow\downarrow$
1	3.85	6.27	2.90	4.93	32.12	52.26	24.15	41.11	4.67	6.73	4.19	6.37
2	5.79	6.78	3.72	7.90	45.04	52.78	28.96	61.43	7.59	8.42	5.27	11.13
5	6.21	18.74	4.71	8.06	22.56	68.04	17.11	29.28	3.38	8.88	3.71	5.78
8	6.27	10.90	3.52	6.57	37.32	64.88	20.95	39.08	5.15	7.94	3.30	5.98
9	5.50	8.60	5.61	10.02	50.25	78.61	51.25	91.62	8.91	13.56	10.44	17.29
14	1.56	2.36	3.40	5.33	17.11	25.90	37.44	58.57	3.02	5.27	7.52	11.58

Note: Reductions in days of high ICU congestion (ΔDays HighCgstn), patients with highly congested ICU stay (ΔPats HighCgstn), and external arrival patients who balk (ΔPats Balk) from four types of interventions: the $\hat{\beta}\uparrow$ denotes increasing $\hat{\beta}$ from estimated level to 0.9; $B\uparrow$ denotes adding one ICU bed; $\lambda_E\downarrow$ and $\lambda_E\downarrow\downarrow$ denote decreasing the external arrival rate by 5% and 10% respectively.

Table 3.12: Counterfactual estimates of impact when β increases from the estimated $\hat{\beta}$ to 0.9

Hosp	$\hat{\beta}$	ΔPr(HighCgstn) (in % points)	Relative ΔPr(HighCgstn)	ΔPr(Balk) (in % points)	Relative ΔPr(Balk)	ΔDays HighCgstn (in # days)	ΔPats HighCgstn (in # patients)	ΔPats Balk (in # patients)
1	0.3	1.06	0.22	0.42	0.27	3.85	32.12	4.67
2	0.5	1.59	0.21	0.65	0.25	5.79	45.04	7.59
3	0.1	0.38	0.12	0.11	0.13	1.38	5.41	0.48
4	0.4	0.34	0.09	0.11	0.09	1.25	13.98	2.45
5	0.1	1.70	0.13	0.71	0.15	6.21	22.56	3.38
6	0.2	0.41	0.23	0.14	0.26	1.49	12.98	1.69
7	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.1	1.72	0.22	0.74	0.27	6.27	37.32	5.15
9	0.5	1.51	0.20	0.58	0.23	5.50	50.25	8.91
10	0.1	0.26	0.06	0.06	0.04	0.96	6.04	0.57
11	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.6	0.62	0.11	0.28	0.15	2.25	16.00	3.70
13	0.1	0.16	0.12	0.02	0.05	0.57	2.82	0.08
14	0.3	0.43	0.16	0.14	0.16	1.56	17.11	3.02
15	0.1	0.06	0.15	0.01	0.07	0.21	1.18	0.03
16	0.1	0.26	0.21	0.08	0.23	0.94	4.48	0.39
17	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.4	0.03	0.08	0.01	0.06	0.12	1.51	0.18
19	0.1	0.01	0.26	0.00	0.00	0.05	0.47	0.00
20	0.3	0.01	0.02	0.01	0.10	0.02	0.10	0.03
21	0.7	0.12	0.01	0.35	0.08	0.45	1.59	2.16
22	0.3	0.34	0.06	0.15	0.07	1.24	6.35	1.22

Note: Counterfactual simulation result from increasing discount factor from current estimated level ($\hat{\beta}$) to 0.9: The third and fourth columns report the absolute and relative drops in high congestion probability in (3.21); the fifth and sixth columns report the absolute and relative drops in balking probability in (3.22). The last three columns report the equivalent numbers of days and patients affected in a year.

While increasing β helps to reduce ICU congestion, it may also lead to longer ED waiting time as hospitals would reduce the likelihood of ICU admission as the unit gets more congested. This potentially undesirable effect can also be quantified by our structural model via simulation. In Table 3.13, we show the increase in average ED waiting time for the two classes of patients as β increases to 0.9, denoted by ΔEDWait_l and ΔEDWait_h respectively. We also compute the difference in average ED waiting time between the two classes at the estimated $\hat{\beta}$ and $\beta = 0.9$. A larger β indeed increases the average ED waiting time for both classes of patients. For example, as β increases from 0.3 to 0.9 in Hospital 1, the average ED waiting time increases by two hours and 36 minutes (0.6 hour) for low and high severity classes, respectively. However, the magnitude of the impact is very different for the two classes. Comparing ΔEDWait_l and ΔEDWait_h, we see the increase in ED waiting time is much more significant for patients in the low severity class than for those in the high severity class. This can be interpreted as follows: As β increases, hospitals tend to reduce the probability of ICU admission and thus increase waiting times. When they do so, they generally prefer to delay the admissions of patients from the low severity class rather than those from high severity class, as the latter on average have higher waiting costs and are more likely to eventually be admitted to the ICU.

Next, we look into the difference in ED waiting times of the two classes, EDWait_l - EDWait_h, evaluated at the estimated $\hat{\beta}$ and $\beta = 0.9$. The average ED waiting times for the two classes are very close at the estimated $\hat{\beta}$. As seen in Section 3.5.1.2, the average waiting time of 1.27 hours (low severity) and 1.42 hours (high severity) produced by our structural model approximates the observed data very well. When β increases to 0.9, the average ED waiting time for the low severity class becomes significantly longer than that for the high severity class in most hospitals. As the hospital accounts more for the longer-term (i.e. the discount factor increases), they tend to actively differentiate across the two classes by admitting high severity patients more quickly and having low severity patients wait longer. Such behavior may partially offset the negative impact of longer ED waiting times, by disproportionately impacting the less severe patients.

We also find there is little impact of β on the overall admission probability, i.e., the proportion of patients eventually admitted to ICU. This is because the overall admission

Table 3.13: Relationship of ED waiting time with β (measured in hours)

Hosp	Estimated $\hat{\beta}$	$\text{EDWait}_l - \text{EDWait}_h$		Change β from $\hat{\beta} \to 0.9$	
		$\beta = \hat{\beta}$	$\beta = 0.9$	ΔEDWait_l	ΔEDWait_h
1	0.3	−0.09	1.33	2.05	0.63
2	0.5	−0.03	1.19	1.84	0.63
3	0.1	0.01	1.31	1.59	0.29
4	0.4	−0.10	0.81	1.33	0.42
5	0.1	−0.05	1.26	1.75	0.44
6	0.2	−0.08	1.11	1.70	0.50
7	0.9	−0.06	−0.06	0.00	0.00
8	0.1	−0.15	0.46	1.36	0.76
9	0.5	−0.04	1.02	1.55	0.49
10	0.1	−0.03	0.43	0.73	0.26
11	0.9	0.01	0.01	0.00	0.00
12	0.6	−0.33	0.10	1.15	0.71
13	0.1	−0.02	1.67	2.08	0.38
14	0.3	−0.09	0.71	1.26	0.46
15	0.1	−0.01	0.24	0.37	0.11
16	0.1	−0.03	1.02	1.48	0.44
17	0.9	−0.17	−0.17	0.00	0.00
18	0.4	0.08	1.22	1.42	0.28
19	0.1	−0.02	0.26	0.45	0.16
20	0.3	0.06	1.63	1.90	0.33
21	0.7	−0.13	0.71	1.33	0.49
22	0.3	0.10	2.86	3.23	0.47

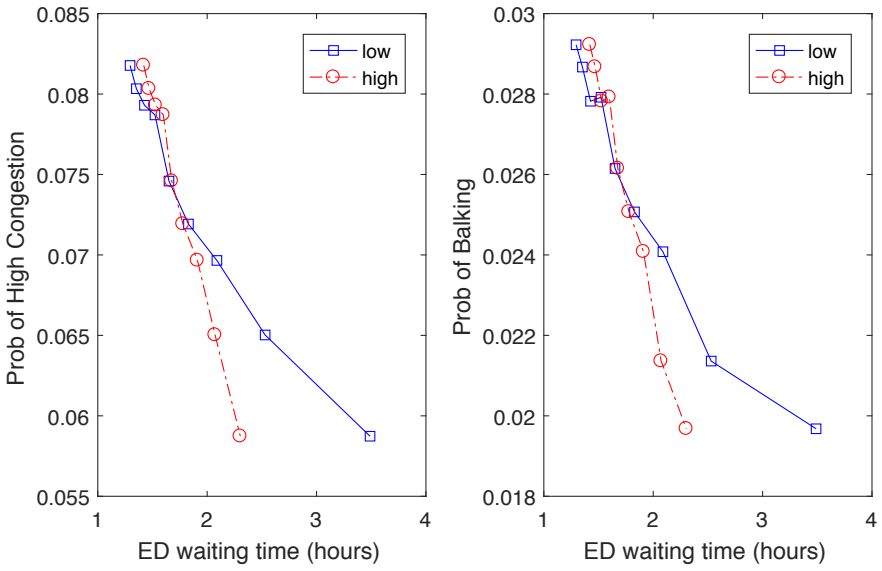
Note: This table reports the impact on ED waiting time from increasing discount factor from current estimated level to 0.9: The third and fourth columns report the difference in ED waiting time between the two classes ($\text{EDWait}_l - \text{EDWait}_h$) at the estimated $\hat{\beta}$ and $\beta = 0.9$; the last two columns (ΔEDWait_l and ΔEDWait_h) report the increase in ED waiting time for the two classes of patients when β increases to 0.9.

probability primarily depends on the non-ICU admission costs and system workload. The non-ICU admission costs measure the average relative effect of admitting the patient to the ward compared with to the ICU, and the system workload determines the long-run bed availability in the ICU. As both are unaffected by changes in β, the overall admission probability also is practically unchanged as β varies. Thus, the primary impact of increasing β is to "delay" admissions to periods when the ICU is less congested. This effectively reduces the likelihood of high ICU congestion states but increases the ED waiting time. In a way, increasing β smooths the load on the ICU.

We use Hospital 2 as an example to further illustrate the trade-off between reduction in

ICU congestion states and longer ED waiting time. In Figure 3.8, we plot the probabilities of high ICU congestion and balking versus the average ED waiting time for the two patient classes, for each $\beta \in \{0.1, 0.2, \ldots, 0.9\}$. The points on each line from left to right represent the results with $\beta = 0.1, 0.2, \ldots, 0.9$ respectively. The downward patterns again illustrate the trade-off of increasing β: Larger β decreases the probabilities of high ICU congestion and balking, but increases ED waiting time for both classes. Moreover, the increase in ED waiting time is more significant for the low severity class, and the two classes are differentiated by their average waiting time when β is large.

Figure 3.7: Counterfactual statistics for Hospital 2 with $\beta = \mathbf{0.1}, \mathbf{0.2}, \ldots, \mathbf{0.9}$ (from left to right)



Note: The left (resp. right) panel shows the high-congestion probability (resp. balking probability) versus ED waiting time for the low (blue solid line) and high (red dotted line) severity patients at Hospital 2 with $\beta = 0.1, 0.2, \ldots, 0.9$ (from left to right).

3.6.4 Practical ways for increasing the discount factor

While the intervention of increasing the discount factor may seem abstract, in some instances there may be actionable approaches the hospitals can take to achieve the desired behavior. Our estimation results demonstrate there are some hospitals that already demonstrate forward-looking behavior (i.e. have large discount factors) when making decisions.

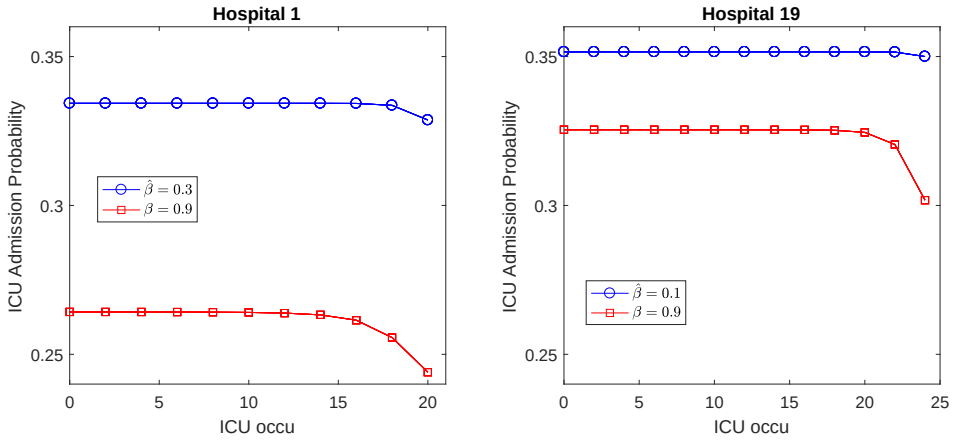
Thus, it may be possible for other hospitals to do so with proper interventions.

There may be multiple reasons a hospital is not forward-thinking in making ICU admission decisions: 1) Conventional training has favored more myopic behaviors that focus on the current patient. 2) Accurate state information is unavailable or not updated in a timely way, e.g., the "occupancy information hurdle" studied in Kim et al. (2019). This makes it impossible to make "correct" forward-looking decisions by taking system state into account. 3) There are other medical and operational constraints that hinder the hospital from acting in a forward-looking manner. For example, shortage in nurse availability and inappropriate levels of care in the ED may force the hospital to admit patients quickly and thus behave "myopically". The corresponding intervention to achieve more forward-looking behavior will depend highly on the root cause of the more near-term looking behavior.

First, the hospital can educate the people involved (e.g., ED physicians, interventionists) about the benefits of forward-looking behaviors and encourage them to act this way. For this purpose, the counterfactual studies by our structural model can provide valuable evidence and insight. Many hospitals lack real-time dashboards that track bed availability. If lack of state information is contributing to near-term focused behavior, the hospital can improve its information system or sharing process to make accurate system state available to people involved in ICU admission decisions. Additionally, the hospital can develop and employ more effective models to forecast the workload of ED and ICU (e.g., Xu and Chan (2016) and Ang et al. (2016)). If other resource limitations drive the behaviors, the hospital can take proactive measures in the ED to stabilize the patients and/or address these staffing limitations. Indeed, it has been observed that providing higher level of care in the ED can decrease the need for ICU and lead to better medical and economic outcomes (see, e.g., Huang (2004); Weingart et al. (2013); Haas et al. (2020))

We use a concrete example of two select hospitals to illustrate how increasing the discount factor changes the admission behaviors. Figure 3.8 shows the ICU admission probability for a single high-severity patient ($n_l^E = 0$ and $n_E^h = 1$) as the ICU occupancy varies from zero to $B - 1$, i.e., only one empty bed, in Hospitals 1 and 9 at their respective estimated discount factor $\hat{\beta}$ (blue) and the intervention level $\beta = 0.9$ (red). As there is only one patient in the ED, it is always permissible to admit the patient to the ICU. Thus, the change in admission

Figure 3.8: Admission probability for a single high-severity patient at different ICU occupancy levels



probability reflects the response to different ICU states. First, the overall probability for ICU admission in each period decreases, e.g., from 33.4% to 26.4% in Hospital 1 and 35.1% to 32.5% in Hospital 19. Moreover, we see the hospitals react more to increases in the ICU occupancy level when the discount factor increases to 0.9. To quantify this relationship, we check the ICU occupancy level where the admission probability drops by one percentage point, i.e.,

$$n^I_{\text{drop}} = \min\{m : \Pr(a_{h,t} = 1 | n^I_t = m) \leq \Pr(a_{h,t} = 1 | n^I_t = 0) - 0.01\}. \tag{3.24}$$

At current estimated $\hat{\beta}$, the level n^I_{drop} is equal to the ICU capacity B for both hospitals (i.e., $n^I_{\text{drop}} = 21$ in Hospital 1 and $n^I_{\text{drop}} = 25$ in Hospital 19). That is, the hospitals will not decrease their ICU admission probability (by up to one percentage point) before the ICU gets full, at which point the admission probability naturally drops to zero. This shows the current behaviors of the two hospitals are indeed near-term focused. On the other hand, with the discount factor increases to 0.9, the level n^I_{drop} becomes $B - 2$ for both hospitals (i.e., $n^I_{\text{drop}} = 19$ in Hospital 1 and $n^I_{\text{drop}} = 23$ in Hospital 19). This means the two hospitals will decrease its ICU admission probability by more than one percentage point when there are still two beds remaining in the ICU. Thus, with larger discount factor, the two hospitals behave more forward-looking by responding to ICU congestion before the the ICU gets full.

3.6.4.1 Mimicking Heuristic

As we observe some hospitals are forward looking, it may be possible to educate the people involved in the ICU decision-making and/or providing higher levels of care in ED and ward in order to achieve such behavior. While this may create an environment where the hospital is open to changing behavior, it still does not provide concrete guidance on how to achieve the desired performance. As such, we propose a simple heuristic policy that mimics the policy from the structural model with $\beta = 0.9$. The heuristic is able to achieve most of the benefits in reducing ICU congestion generated under the 'optimal' policy suggested by the structural model. Crucially, the simplicity of this policy is easy to communicate to clinicians and administrators, thereby helping to facilitate adoption.

For this heuristic policy, we assume the decisions are made independently for each patient in the ED. The decision probabilities are restricted to be independent of the ED state and can only change once with respect to the ICU state. The heuristic policy is implemented as follows. We divide the ICU states to two regimes given a threshold in ICU occupancy level T_{ICU}. Thus, T_{ICU} separates the ICU state into high versus low congestion states. The probabilities for ICU admission, non-ICU admission, and ED waiting for each patient in every period are given by

$$p^{(1)}_{I,\iota},\ p^{(1)}_{N,\iota},\ \text{and}\ p^{(1)}_{W,\iota},\ \text{if}\ n^I < T_{ICU},$$

and

$$p^{(2)}_{I,\iota},\ p^{(2)}_{N,\iota},\ \text{and}\ p^{(2)}_{W,\iota},\ \text{if}\ n^I \geq T_{ICU},$$

where $\iota \in \{l, h\}$ denote the patient severity classes. Note that when some of the actions are not permissible given the ED and ICU states, e.g., not all ED patients can be admitted to the ICU due to the capacity constraint, we normalize the probabilities of other permissible actions to have a sum of one. The constant probabilities $p^{(i)}_{I,\iota}$, $p^{(i)}_{N,\iota}$, $p^{(i)}_{W,\iota}$ for $\iota \in \{l, h\}$ and $i \in \{1, 2\}$, as well as the threshold level T_{ICU} are parameters to be estimated such that the heuristic policy mimics the structural model with $\beta = 0.9$.

We estimate the heuristic policy for each hospital separately. We first use the structural model with $\beta = 0.9$ to generate a large sample of state-action pairs for each hospital with

100 simulation runs; each run contains the same number of periods as in the data plus a three-month warm-up period. For a given threshold T_{ICU}, we estimate the probabilities $p^{(i)}_{I,\iota}$, $p^{(i)}_{N,\iota}$, $p^{(i)}_{W,\iota}$ for $\iota \in \{l, h\}$ and $i \in \{1, 2\}$ by a simple logit model with only intercepts that are allowed to change across the two ICU state regimes (e.g. high versus low congestion). We then search over all possible threshold values to find the one (together with the probabilities) that maximizes the log-likelihood. In the estimation, we drop the periods when ICU is full, as no ICU admissions are permissable in this state.

For illustrative purposes, we provide the estimated heuristic policy for the same six hospitals in Table 3.11. The policies for the other hospitals are similar. The threshold T_{ICU} and the action probabilities for the low ($n^I < T_{ICU}$) and high ICU occupancy regimes ($n^I \geq T_{ICU}$) for each patient class are reported in Table 3.14. We observe the hospitals change their action probabilities only when the ICU occupancy is relatively high, i.e., the threshold T_{ICU} is close to the capacity B. When the ICU occupancy reaches the threshold, the ICU admission probabilities for both types of patients drop. These observations reconcile the patterns we observed in Figure 3.8 for the ICU admission probability when $\beta = 0.9$.

Table 3.14: Heuristic policy for select hospitals

			Low severity patients						High severity patients					
Hosp	B	T_{ICU}	Low ICU: $n^I < T_{ICU}$			High ICU: $n^I \geq T_{ICU}$			Low ICU: $n^I < T_{ICU}$			High ICU: $n^I \geq T_{ICU}$		
			$p^{(1)}_{I,l}$	$p^{(1)}_{N,l}$	$p^{(1)}_{W,l}$	$p^{(2)}_{I,l}$	$p^{(2)}_{N,l}$	$p^{(2)}_{W,l}$	$p^{(1)}_{I,l}$	$p^{(1)}_{N,l}$	$p^{(1)}_{W,l}$	$p^{(2)}_{I,l}$	$p^{(2)}_{N,l}$	$p^{(2)}_{W,l}$
1	21	20	0.05	0.39	0.55	0.03	0.41	0.55	0.26	0.37	0.37	0.19	0.40	0.37
2	26	24	0.05	0.39	0.55	0.036	0.41	0.55	0.22	0.39	0.39	0.18	0.43	0.38
5	11	10	0.05	0.45	0.50	0.041	0.46	0.49	0.30	0.43	0.27	0.25	0.49	0.25
8	16	15	0.07	0.48	0.45	0.05	0.50	0.44	0.25	0.39	0.36	0.19	0.45	0.36
9	22	20	0.04	0.46	0.51	0.03	0.46	0.50	0.20	0.45	0.34	0.17	0.49	0.33
14	36	33	0.04	0.53	0.43	0.03	0.53	0.43	0.23	0.48	0.29	0.21	0.50	0.29

With the heuristic policy for each hospital, we can then measure its performance using simulation. Table 3.15 reports the effects on ICU congestion reduction for the select hospitals from the heuristic policy and those from increasing β to 0.9 as given in Table 3.12. We can see the benefits in reducing ICU congestion from increasing β to 0.9 can be mostly obtained by the heuristic policy. For some hospitals, the effects from the heuristic policy are even slightly larger. Additionally, we note that the overall ICU admission probabilities and ED waiting times for the two classes of patients, as well as the average occupancy level of ICU, are almost unchanged under the heuristic policy compared to those under the structural

model with $\beta = 0.9$. This suggests the hospitals could reduce ICU congestion safely by being forward-thinking through the use of a simple policy that is easy to implement in practice.

Table 3.15: Comparison of effects from the heuristic policy and increasing β to 0.9 (select hospitals)

Hosp	$\hat{\beta}$	Heuristic policy			Structural model with $\beta = 0.9$		
		ΔDaysHighCgstn	ΔPatsHighCgstn	ΔPatsBalk	ΔDaysHighCgstn	ΔPatsHighCgstn	ΔPatsBalk
1	0.3	3.38	28.20	4.27	3.85	32.12	4.67
2	0.5	6.64	50.12	8.73	5.79	45.04	7.59
5	0.1	5.24	19.02	3.11	6.21	22.56	3.38
8	0.1	5.45	32.41	4.62	6.27	37.32	5.15
9	0.5	6.36	58.22	10.56	5.50	50.25	8.91
14	0.3	1.73	19.07	3.54	1.56	17.11	3.02

Note: Reductions in days of high ICU congestion (ΔDays HighCgstn), patients with highly congested ICU stay (ΔPats HighCgstn), and external arrival patients who balk (ΔPats Balk) from the heuristic policy and the structural model with $\beta = 0.9$ for select hospitals.

3.7 Concluding Remarks

Clinical practice aims to provide the best care possible for each individual patient. That said, it has also been well documented that ICU admission behaviors are impacted by ICU congestion. In this work, we aim to understand if some of this practice variation could be explained by perceived discounting behavior. To the best of our knowledge, our work is the first to study how hospitals internalize intertemporal externalities – i.e., the admission decision for a current patient impacts ICU congestion and possible other, future patients – when making the dynamic ICU admission decision. While, on average, the hospitals appear to be more focused on the near-term when making ICU admission decisions, we find that there is large heterogeneity in the degree of forward-looking behavior across hospitals. This suggests that some of the observed practice variation in how hospitals alter admission behaviors in response to ICU congestion may be partially explained by the fact that hospitals appear to internalize the intertemporal externalities very differently.

We use counterfactual simulations to show that if it were possible to increase the hospital's degree of forward-looking behavior alone, this can have a substantial impact on ICU congestion and patient flow. For some hospitals, the effect of this change is comparable to increasing the ICU capacity or decreasing external arrivals, which can be very capital and space expensive. We provide a number of suggestions on how such behavioral change

may be achieved in practice. For instance, we demonstrate a simple heuristic policy which alters admission behavior once the ICU reaches a 'high congestion state' can achieve most of the benefit in ICU congestion reduction that is generated by increasing the discount factor. Despite the practical challenges in altering hospital behavior, our work suggests that how hospitals respond to staffing and capacity shortages could benefit from a better understanding of their forward-thinking behavior. This would also help the hospital better understand the potential impact different strategies – e.g. better forecasting capabilities of demand and/or patient severity – could have on patient flow.

Our study has several limitations that may shed light on future research directions. First, the data we use have no direct information on the ICU admission decision process – we can only observe the resulting outcome of where patients are admitted. Thus, while we posit a structural model to capture various features which influence where patients are admitted and when, there are many factors which are not explicitly included in the model nor do we have data to understand their impact on the decisions. Second, the data is also limited to hospitals within one healthcare system. It is possible that different payment models may impact the manner in which hospitals internalize intertemporal externalities. For instance, it is possible that under a purely Fee-For-Service system (rather than the capitated system of KPNC), hospitals may be even more focused on the near-term. If this is the case, our counterfactual results suggest that other hospital systems may benefit even more from shifting to be more longer-term focused. Finally, all the ICUs in our study cohort are closed, so that the attending intensivist has final say on which patients are admitted. It is not clear how the hospital would internalize the intertemporal externalities in an open ICU.

In order to focus on the identification of the intertemporal externalities – the discount factor – we propose a parsimonious structural model. As ICU patient flow and admission decisions are quite complex, it would be impossible to capture all features and still have a tractable model and so we made a number of simplifications in our model. For example, we assume constant patient arrival rates for ED patients, a homogeneous constant and state-independent departure rate for all ICU patients, and we do not account for possible ICU readmissions. As analysis of queueing systems with such features is an active area of research

which often utilizes approximation approaches (e.g. fluid and/or infinite server models) and/or asymptotic regimes, it would be interesting to see whether it would be possible to incorporate such features into the structural model. This might require assuming that the hospital must heuristically solve the dynamic optimization problem. Finally, this study focuses on one aspect of the hospital's behavior in ICU admission decisions – the degree of internalization of the intertemporal externalities. However, there is evidence that there are other adaptive behaviors – such as demand driven discharges (Kc and Terwiesch, 2012) – the hospital may utilize when managing ICU beds. Such behavior may also affect or be affected by the discount factor of the hospital. Thus, a potential future direction of research is to study the joint impact of these various behavioral aspects of the hospital's decision-making on ICU capacity management.

Chapter 4

Effects of Surgeon's Daily Workload and Implications in Operating Room Scheduling

4.1 Introduction

The relationship between system workload and service performance has drawn increasing attention in the operations management community. Traditional operation models generally assume service time is fixed and independent of the system workload (see e.g., Wolff (1989) and Dallery and Gershwin (1992)). However, other research empirically shows that the service time of human-involved systems can be endogenously impacted by the overall workload (Schultz et al., 1998; Staats and Gino, 2012; Tan and Netessine, 2014). Such impacts are particularly important in the field of healthcare, as medical services need to be provided in a time-sensitive manner with limited resources. The system workload has been shown to affect service time in different healthcare settings, such as intensive care units (Kc and Terwiesch, 2012), patient transportation and cardiac surgery (Kc and Terwiesch, 2009), as well as emergency departments (Kc, 2014; Batt and Terwiesch, 2016). While some studies found the service time can increase with workload levels (Green and Nguyen, 2001; Tan and Netessine, 2014), the opposite pattern, i.e., service time decreasing in workload levels, is also observed (Kc and Terwiesch, 2009; Chan et al., 2012). Different mechanisms have been proposed to explain the observed patterns. For example, the pattern of increasing service time can be attributed to multitasking (Freeman et al., 2017) and mental strain of workers (Kuntz et al., 2015), while the decrease in service time can be explained by workers' behavioral response of "working faster" (Kc and Terwiesch, 2009; Staats and Gino, 2012) and "omitting tasks" (Oliva and Sterman, 2001) when faced with high workloads. Reconciling the two patterns, some recent studies show that the service time can react non-monotonically to

workload levels. These studies have found an inverted-U shaped pattern, i.e., service time first increases and then decreases in the workload measures (see, e.g., Tan and Netessine (2014), Batt and Terwiesch (2016), and Berry Jaeker and Tucker (2017)).

Beyond service time, the effect of hospital's workload on the quality of care has also been investigated in both the operations management and medical community. There is empirical evidence that increased workload can lead to worse medical outcomes, such as higher mortality (Kc and Terwiesch, 2009) and readmission rate (Kc and Terwiesch, 2012; Kc, 2014). This positive linkage is also observed in the medical literature (e.g., Needleman et al. (2011)) . The negative effect of high workload can be explained by the mental strain of healthcare workers (Kuntz et al., 2015) or the delay in treatment received by patients (Chalfin et al., 2007). Moreover, some studies have identified a tipping point in the workload level, after which the service time increases and the quality of care deteriorates (Kuntz et al., 2015; Berry Jaeker and Tucker, 2017). This suggests a workload-related "saturation effect": when the workload is very high, workers become exhausted and system buffers for handling the demand are depleted.

In this chapter, we empirically investigate the impact of workload on service time and quality in the context of cardiac surgeries. We focus on the daily workload of surgeons, e.g., the number of surgeries performed by the surgeon on a given day. In most of the existing literature, workload is measured on a system level, usually as the bed occupancy in different hospital units at the time of patient's admission (e.g., Kc and Terwiesch (2012), Kuntz et al. (2015), and Kim et al. (2015)). Different from them, our study considers a novel type of workload; namely we measure the workload for individual surgeons on each day. To our best knowledge, we are the first to study the impact of surgeons' daily workload in the field of operations management.

In the hospital of our study, it is common for a cardiac surgeon to do multiple surgeries a day. In particular, the median of surgeons' daily workload is two surgeries, and the maximum is four surgeries. On average, each surgery takes more than seven hours to complete, and the surgeon has to be highly concentrated during the procedure. Although some parts of the surgery can be done by other members in the medical team, performing multiple surgeries a day is still a heavy burden for the surgeon. With long working hours, surgeons can

suffer from severe physical and mental fatigue (see, .e.g, Janhofer et al. (2019)), which may lead to worse medical outcomes. Thus, understanding the impact of surgeons' workload is important for hospitals to improve their surgical outcomes and system performance. Note that in our sample, the minimum of surgeon's workload level is one surgery a day, which is already burdensome and time-consuming for the surgeon. In this sense, our study focuses on the impact of workload in the region where the workload level is already high.

In this chapter, we examine the impact of surgeons' daily workload using a data set of cardiac surgeries from a large hospital. Our data contains detailed information of more than 5,600 cardiac surgeries that are performed over a horizon of four years. We measure the impact of surgeon's daily workload on multiple outcomes. First, we examine how surgeon's daily workload affects the surgery duration of each case, as measured by its incision time. This sheds light on the relationship between workload and service time in the context of cardiac surgery, i.e., whether the incision time increases or decreases when the surgeon performs more cases. Next, we analyze the effects of surgeon's daily workload on the patient's post-surgery length-of-stay (LOS) in the ICU and in the hospital. The post-surgery LOS is important for hospital as it affects the demand for downstream resources (e.g., ICU and ward beds) and overall throughput efficiency. Finally, we check the impact of surgeon's workload on the likelihood of adverse post-operation events for their patients, including reoperation, readmission, and mortality. The comprehensive scope of our analysis also differentiates our study with most of existing literature, which only considers one or two outcomes such as LOS and mortality. Besides, the effect of surgeon's workload may be heterogeneous for different types of patients. For example, the urgent and emergent patients are generally more severe than the elective patients, thus their surgical outcomes may be more sensitive to surgeon's fatigue. Such potential heterogeneity in the effects of workload is also accounted for in our study.

Our detailed data set allows us to control for a variety of demographic, risk, and operative factors that may also affect the surgical outcomes. However, we still face a major challenge in identifying the true effect of surgeon's daily workload. That is, the surgeon's daily workload is endogenous. This is because there exists risk factors that are considered by the surgeons when they schedule their cases, but these factors are not observable in the

data. These unobservable factors will affect both the surgeon's daily workload and the surgical outcomes, thus violating the exogeneity condition for identification. For example, a surgeon may schedule more cases if the unobservables imply less risk. This will generate a negative bias in the causal effect of surgeon's daily workload. We handle the endogeneity bias by proper instrument variables (IV). The IV method has been widely used in healthcare operations management for patient admission decisions (e.g., Kc and Terwiesch (2011), Kc and Terwiesch (2012), and Kim et al. (2015)). We now apply it in the context of cardiac surgeries to control the endogeneity in surgeon's workload.

A valid IV in our study should influence the surgical outcomes only via the surgeon's daily workload. We construct two instrument variables by leveraging the operational factors in cardiac surgeries. The first IV is the number of cases performed by other surgeons on the same day. As many resources are shared by surgeons in the cardiac department, more surgeries performed by other surgeons tend to limit the daily workload of the focal surgeon. We then construct another IV using novel operational data, which is the block schedule of surgeons. Specifically, we define the second IV as the number of days until the next scheduled block of the focal surgeon. This IV is based on the following surgeon's behavior: the surgeon may "squeeze in" more cases if his or her next scheduled block is far away. We validate the two IVs empirically with our data and show they are essential for correctly estimating the effect of surgeon's daily workload.

We find higher daily workload for surgeons is associated with longer incision time of the surgeries and worse outcomes for the patients. Specifically, adding one more case to a surgeon's daily workload increases the incision time by 26 minutes for each case performed by the surgeon in the day. This is a 9% relative increase. In addition, surgeon's daily workload leads to longer post-surgery LOS of patients in both the ICU and the hospital: when the surgeon needs to do one more surgery in a day, the affected patients are expected to stay in the ICU (resp. hospital) for 1.03 (resp. 1.41) more days after their surgeries. In addition, we find higher daily workload increases the patient's likelihood of reoperation and readmission. These consistent results highlight the negative impact of surgeon's fatigue due to daily workload. Recall our study focuses on the impact of workload when the workload level is already high. Thus, our results also reconcile the findings in Kuntz et al. (2015) and

Berry Jaeker and Tucker (2017) — they show very high levels of workload that exceeds a tipping point is associated with longer LOS and higher mortality respectively.

We further show there is substantial heterogeneity in the effect of daily workload for elective and non-elective patients. The non-elective patients refer to those in the urgent, emergent, and salvage surgeries. We find the effect of surgeon's daily workload on incision time is statistically significant for the elective patients, but not for the non-elective ones. On the contrary, the effects on post-surgery LOS (in both ICU and hospital), reoperation, and readmission are significant only for non-elective patients. A possible explanation for such heterogeneity in effect goes as follows: the surgeries for non-elective patients are more time-constrained, thus their incision time is less impacted by surgeon's fatigue. On the other hand, the non-elective patients are generally more severe, and their surgical outcomes (post-LOS, reoperation, and readmission) are more sensitive to surgeon's fatigue.

Based on the empirical results, we develop a surgery scheduling model that incorporates the effect of surgeon's daily workload. Operating rooms are expensive medical resources and generate up to a half of hospital's revenues (McDermott et al., 2017), and accordingly the literature on surgical scheduling is large (see, e.g., Keskinocak and Savva (2020)). In most of the existing literature, the surgery duration is assumed to be exogenous with deterministic or stochastic distributions. However, as shown by our study, the surgery duration endogenously depends on surgeon's daily workload. We thus propose a scheduling model that accounts for such effect. The motivation of our model is to smooth surgeons' daily workloads by switching the cases from different days. We consider the objective of minimizing the total expected incision time and show it can be formulated as a mixed-integer quadratic programming problem.

In summary, we make following key contributions in this chapter.

- **Impact of surgeon's daily workload:** We empirically estimate the impact of surgeon's daily workload on surgery duration and patient outcomes using a detailed data set of cardiac surgeries. We find surgeons' workloads increase surgery duration and lead to worse patient outcomes (post-surgery LOS, reoperation, and readmission). Besides, we show the effects are highly heterogeneous for different types of patients and

different outcomes. These results highlight the substantial impact of surgeon's daily workload. This provides a potential method for hospitals to improve their surgery performance.

- **Estimation methodology:** To address the endogeneity bias in surgeon's daily workload, we propose two novel IVs using the operational factors in the cardiac department. The first IV is based on the resource sharing across surgeons in the department. For the second IV, we leverage the surgeons' block schedule data to capture the "squeeze in" behavior that affects the surgeon's workload. We validate the two IVs empirically and show they are essential for identifying the true effect of surgeon's workload.

- **Surgery scheduling:** Our findings suggest surgeon's daily workload can substantially affect the surgery duration and patient outcomes. However, such impacts are largely ignored in the previous literature on surgery scheduling. Thus, we propose and formulate a surgery scheduling model that incorporates the effects from surgeon's daily workload. The model optimizes for the times of surgery to capture the benefit of smoothing surgeon's workload across days.

The rest of the chapter is organized as follows. The next section is a brief overview of related literature. Section 4.2 describes the data and clinical setting of our study. In Section 4.3, we develop the econometric framework and estimation methodology. Section 4.4 provides and assesses the main empirical findings. We propose our surgery scheduling model in Section 4.5. Section 4.6 concludes the chapter and discusses ongoing directions. The appendices include variable definitions and supplementary tables.

4.1.1 Literature Review

Our study is related to four streams of literature: (1) the effect of system workload on service rate and quality, (2) volume-outcome relationships, (3) the impact of surgeon's fatigue, and (4) operating room scheduling.

While traditional models usually assume a constant and exogenous service rate, there is rich literature, both analytical and empirical, that focuses on the relationship between

system workload and service rate. The dynamic queueing control literature has derived optimal service rates that balance the costs of acceleration and waiting time (e.g., George and Harrison (2001)). These analytical models suggest the system should increase its service rate in the face of high workload (i.e., queue length). In reality, such optimal policies can not always be achieved by human workers. To examine how human workers actually behave under varying workload, various empirical research has been conducted using observational data in real-world settings, and the results are mixed. Kc and Terwiesch (2009) show that workers for patient transport and cardiac surgery increase their service rate under high workload. Kc and Terwiesch (2012) and Chan et al. (2012) find hospitals are likely to discharge patients early when ICU occupancy is high, i.e., decreasing the service time. The opposite direction of the impact is also observed empirically. For example, Dietz (2011) finds a positive correlation between call volume and the average service time when call volume is high. Green and Nguyen (2001) show patient's LOS can increase when patient load becomes higher

The seemingly opposite effects of workload can be reconciled by an inverted-U shape pattern between service time and workload. That is, the service time first increases and then decreases with the workload level. Empirical evidence for this inverted-U shape pattern is found using restaurant chain data in Tan and Netessine (2014), and in the healthcare setting in Batt and Terwiesch (2016) and Berry Jaeker and Tucker (2017). In particular, Berry Jaeker and Tucker (2017) further show there exists a second tipping point at very high bed occupancies in the hospital, after which the service time (patient's LOS) increases again with workload. Different mechanisms have been proposed to explain the impact of workload on service time. For example, the decrease in the service time can be explained by server speedup (Staats and Gino, 2012; Kc and Terwiesch, 2009; Tan and Netessine, 2014), task reduction (Oliva and Sterman, 2001; Kuntz et al., 2015), or early task initiative between stages (Batt and Terwiesch, 2016). On the other hand, the slowdown in service time can be caused by multitasking (Tan and Netessine, 2014; Freeman et al., 2017), mental fatigue (Kuntz et al., 2015), and change in patient's average severity level (Berry Jaeker and Tucker, 2017). On the analytical side, Delasay et al. (2016) develop a state-dependent queueing model to capture the adaptive mechanisms leading to the nonlinear pattern between service

rate and workload. Delasay et al. (2019) provide a general framework that incorporates different effects of load on service time.

There is also a rich literature studying the effect of workload on servers' behavior and quality. Green et al. (2013) find the nurse absenteeism is positively correlated with the expected future workload. Hopp et al. (2007) use an analytical queueing model to show increasing servers may worsen congestion when servers have discretion over task completion time. Freeman et al. (2017) find that gatekeeper-providers would alter their service configuration and referral decisions in response to their workload. In multiple healthcare settings, the quality of care is found to suffer under high workload. Kc and Terwiesch (2009) and Kc and Terwiesch (2012) find that hospital's workload increases the patient's mortality and readmission rate. Kim et al. (2015) find the delay of ICU admission due to high occupancy leads to longer LOS and higher likelihood of transfer-up. This positive linkage between hospital workload and mortality is also observed in the medical literature (e.g., Schilling et al. (2010) and Neuraz et al. (2015)). The negative impact on quality is particularly significant when the workload is already high, producing a safety tipping point in occupancy level (Kuntz et al., 2015; Berry Jaeker and Tucker, 2017).

Our study contributes to this line of literature in the following aspects. We focus on a novel type of workload in the healthcare setting, which is the number of surgeries performed by a surgeon in a day. We find surgeon's high workload is associated with longer surgery duration and worse patient outcomes. This provides consistent evidence for the negative impact of very high workload level, see, e.g., Kuntz et al. (2015) and Berry Jaeker and Tucker (2017). Besides, we reveal the effect of surgeon's workload is highly heterogeneous across different patients and outcomes. Finally, to control for the endogeneity in surgeon's workload, we develop two IVs based on resource sharing and surgeon's block schedule. These types of IVs may be applied in other empirical settings to address the endogeneity bias in workload.

Next, our work is related to the literature on volume-outcome relationship in healthcare management. In the medical community, there is vast evidence supporting a positive relationship between a surgeon's (or a hospital's) volume and surgical outcome (see, e.g., Falcoz et al. (2014), Bashir et al. (2017), and Modrall et al. (2018)). The volume in these studies

usually refers to the number of surgeries performed by the surgeon in a relatively long period (e.g., the past one year). The volume-outcome relationship has also drawn attention in the field of operations management. Research in different empirical settings has been conducted to investigate the driver and mechanism behind the relationship, e.g., learning and specialization. For example, Kc and Terwiesch (2011) show that after controlling for selective patient admissions, the benefit of specialization disappears at the hospital level, but it still exists at the operating unit level in terms of a shorter patient's LOS. Kc and Staats (2012) disentangle the volume-outcome relationship by dividing experience into focal and related categories. They find a surgeon's focal experience has a greater impact on surgical outcome than related experience. Clark and Huckman (2012) identify the existence of complementarities resulting from cospecialization in focal and related segments, i.e., hospitals with greater specialization in related areas have a higher marginal benefit from specialization in the focal area. Using transaction data from a Japanese bank, Staats and Gino (2012) show specialization improves the performance in the short-term (single-day), while variety increases worker productivity in a longer-term (across days). In a recent work by Wang and Pourghannad (2020), they show the effects of surgical volume on surgery duration is heterogeneous across patients. Complementing to this line of literature, we investigate the impact of surgeon's short-term volume, i.e., number of cases performed in a day, on surgery duration and surgical outcomes.

Our work also relates and contributes to the literature on surgeon's fatigue. As surgeon's work is highly demanding both physically and mentally, the potential negative impact of surgeon's fatigue has long been a focus of the medical community (see a recent survey in Janhofer et al. (2019)). Long working hours and the consequent suboptimal sleep of surgeons can cause different types of fatigues, including muscular fatigue (Slack et al., 2008; Dorion and Darveau, 2013), mental fatigue (Gerdes et al., 2008; McCormick et al., 2012), and decision fatigue (Stewart et al., 2012). Under different medical settings, multiple studies have shown surgeon's fatigue is associated with worse surgical outcomes (see, e.g., Halldorson et al. (2009), Shanafelt et al. (2010), and Thomas et al. (2012)). However, other studies have found no significant impact from surgeon's fatigue on surgical outcomes (Ellman et al., 2005; Bagrodia et al., 2012; Govindarajan et al., 2015). In general, the medical literature

does not have a clear conclusion on the relation between surgeon's fatigue and worse patient outcomes. Our work sheds light on this problem using a detailed empirical data set of cardiac surgeries and rigorous econometric analysis. As an important difference from existing medical literature, we use proper IVs to control for the endogeneity in surgeon's daily workload. This accounts for the possibility that surgeons will schedule less severe cases when they know their workload is high. Ignoring such endogeneity may make it difficult, or even impossible, to identify the true effect of surgeon's fatigue .

We also contribute to the literature of operating room scheduling. Operating rooms are big cost centers and revenue generators of the hospital. However, efficient management of operating rooms often faces operational difficulties such as low utilization, late starts, overtime costs, and unexpected cancellations (Doebbeling et al., 2012). Thus, the literature on operating room scheduling is huge. Some review of the current research, challenges, and future directions of this field can be found in Cardoen et al. (2010), May et al. (2011), and Samudra et al. (2016) among many others. Different objectives are considered in operating room scheduling, including minimizing total costs (Denton et al., 2010; Batun et al., 2011), maximizing profit (Freeman et al., 2016), maximizing expected resource utilization (Gupta, 2007; Shylo et al., 2013), reducing patient wait times (Zenteno et al., 2015), and smoothing downstream census (Zenteno et al., 2016). Combinations of these objectives are also considered (e.g., Min and Yih (2010), Gul et al. (2011), and Li et al. (2017)). From a different aspect, Olivares et al. (2008) apply a structural estimation method on observational data to identify how the hospital actually balances the costs of reserving too much versus too little operating room capacity for cardiac surgeries. However, most of the existing literature assumes the surgery duration, either deterministic or stochastic, to be exogenous and independent of surgeon's workload. To the best of our knowledge, we are the first to develop a scheduling model that incorporates the effects of surgeon's daily workload. Another example of endogenous surgery duration is the recent work by Wang and Pourghannad (2020), in which the surgery duration is dynamically determined by the matching between patients and surgeons. The decision variables in their model is the assignment of patients to surgeons, while we consider the potential benefit of smoothing surgeon's daily workloads by changing the surgery time.

4.2 Data and Clinical Setting

In this study, we use a sample of cardiac surgeries from a large hospital over the period of July 2015 to July 2019. We combine two data sets. The first one is the cardiac surgery data collected from the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database.[1] The second one is the block schedule data for surgeons provided by the hospital.

The STS data contains detailed information of patient demographics, risk factors, preoperative status, operative procedures and timelines, as well as postoperative events. This comprehensive data set allows us to control for the severity of patients and complexity of surgeries when analyzing the impact of daily workload. The STS data contains basic patient demographics such as gender, age, and race. Besides, the risk factors section includes a patient's status for liver illness, lung disease, diabetes control, and renal failure; the preoperative section records whether the patient experienced heart fail, cardiogenic shock, or myocardial infarction (MI) before the surgery.

We obtain from the STS data each patient's hospital admission date, surgery date, and discharge date. Thus, we can compute the patient's LOS before and after the surgery. Besides, for each surgery, we can determine its OR time and incision time from the STS data using the timestamps of its OR entry and exit, as well as skin incision start and end. We also have the surgeon's identifier related to each case, which enables us to measure the daily workload for each surgeon and control for the difference in surgeons' experience. Finally, the STS data contains multiple medical outcomes of each surgery, such as the time spent in ICU, reoperation, readmission, and mortality.

The second data set is the block schedule of surgeons provided by the cardiac department. The hospital in our study employs a block booking framework to schedule their cardiac surgeries. Under the block booking, surgeons are assigned with fixed time slots (blocks) and dedicated resource (e.g., OR and staff) to perform their surgeries. The block schedule is decided in advance by the management board and adjusted infrequently, e.g., each quarter or twice a year. The block scheduling framework is widely accepted in the US as it is

[1] `https://www.sts.org/registries-research-center/sts-national-database/adult-cardiac-surgery-database/data-collection`

convenient for surgeons and hospitals to employ (Erdogan and Denton, 2010). Each block in our data specifies the date, OR number, and the surgeon assigned, e.g., OR 1 is assigned to Surgeon A on 10/1/2016. We note that in most cases of our study, each OR is assigned to only one surgeon for the entire day. That is, OR sharing by multiple surgeons is rare (only 3%). There are in total eight ORs for cardiac surgeries. However, some blocks of the ORs are assigned to other departments in the hospital (e.g., the pediatrics department). This is also documented in the block schedule data.

In principle, the block schedule data allows us to determine for each surgery whether it happens in or out of the block schedule. Here "in block schedule" means the surgery is performed in an OR that is assigned to its surgeon on the surgery date. However, we have two limitations in the available data. First, a significant proportion of the block schedule data is missing: out of the 48 months (resp. 5,604 cases) in our surgery data, we only have the block schedule data for 22 months (resp. 2,499 cases). Thus, we would need to impute the block schedule information for the missing periods. We note that the missing block data is due to the staff absence in the department administration, instead of any selection bias regarding the patients. Second, we do not have the location information (the OR number) for each surgery in our data. Because of this limitation, we can only determine the block status on the surgeon-day level, i.e., whether a surgeon is assigned a block on a specific day. We then use this block status for all the cases performed by the surgeon on that day.

4.2.1 Data Selection and Summary Statistics

In this section, we describe the data cleaning process and provide some summary statistics of the final sample in our study.

We start from 5,604 cases from the STS data. We first drop 20 cases that are eventually cancelled before or during the surgery. We then drop 232 cases from seven "infrequent" surgeons in our sample. These surgeons only performed a very small number of cases during the four year horizon. They are dropped for the following two reasons. First, these surgeons may be more likely to do unusual procedures that require special expertise. Second, the small sample size of these surgeons does not allow us to effectively control for the impact from their experience. Thus, we focus on the cases from the other eight surgeons, each of

which performed at least 200 cases in the sample. This leaves us with a sample of 5,352 cases in total, which consist of 95.5% of the initial sample. Among them, we have the block schedule information for 2,492 cases (46.5%). We refer to these cases as the block sample hereafter.

Table 4.1 reports the summary statistics of patients' gender, age, and critical status for both the full sample and the block sample. Specifically, a patient is classified as critical if he or she experiences a cardiogenic shock or syncope before the surgery, both of which are controlled in our estimation. In Appendix D.1, we provide a detailed description of other factors included in our econometric framework and their summary statistics.

Table 4.1: Summary Statistics of Patients for the Full Sample and Block Sample
(Full Sample: N = 5,352, Block Sample: N = 2,492)

	Full Sample			Block Sample		
	Mean	Median	Std	Mean	Median	Std
Gender: Male	0.675	-	-	0.671	-	-
Age	64.73	66.00	12.56	65.06	66.00	12.33
Critical	0.103	-	-	0.102	-	-

As a main risk indicator, the cardiac surgeries are divided to four status categories in the increasing order of patient severity: elective, urgent, emergent, and salvage. The elective cases are those surgeries that can be deferred without increased risk; the urgent cases are supposed to be performed during the same clinical stay to reduce further risk; the emergent and salvage cases refer to the situation that requires emergent operations with no delay upon the outbreak.[2] The surgery status has important implication on the surgical scheduling. While the hospital has relatively high flexibility in scheduling the elective cases, the schedules of urgent cases are more difficult to change, and the hospital has little control over the time of emergent and salvage cases. In Table 4.2 below, we provide the summary statistics for the four categories in both the full sample and the block sample. We have two findings from the statistics. First, a significant proportion of the surgeries are urgent or emergent cases. This consists of 53.5% of the full sample and 52.3% of the block sample. Indeed, the numbers of

[2]See page 154 in the training manual: `https://www.sts.org/sites/default/files/Training%20Manual%20V2-9%20June%202020.pdf`

elective and urgent cases are almost the same. Second, the distributions of the four status are very similar for the full and block samples.

Table 4.2: Statistics of Surgery Status in Full and Block Sample

	Full Sample		Block Sample	
Status	Number	Ratio	Number	Ratio
Elective	2479	46.3%	1184	47.5%
Urgent	2488	46.5%	1124	45.1%
Emergent	374	7.0%	180	7.2%
Salvage	11	0.2%	4	0.2%

Besides the surgery status, we can obtain the procedure information for each case from the STS data, i.e., which types of procedures are performed during the surgery. To control for the impact from different procedures, we classify the surgeries to different types as follows. First, there are eight standard types for those most commonly performed cardiac surgeries. For them, we directly use the classification provided by the STS data: coronary artery bypass graft (CABG), aortic valve replacement (AVR), mitral valve replacement (MVR), mitral valve repair (MVr), and their combinations CABG+AVR, CABG+MVR, CABG+MVr, and AVR+MVR. For other "non-standard" cases, we determine their surgery types based on the procedures actually performed. In total, we have 14 procedure types for the cases in our data. The heuristic rule for determining the surgery types of non-standard cases is described in Appendix D.1. There are 3,420 cases (63.9%) that fall into the eight standard types, and 1,932 cases (36.1%) for the non-standard types.

We compute the pre-surgery LOS (pre-LOS) for each patient as the number of days between the hospital admission and the surgery, and the post-surgery LOS (post-LOS) as that between the surgery and hospital discharge. Besides, the OR time of each case is calculated as the time elapsed between its OR entry and OR exit. As shown in Figure 4.1, the OR time can be decomposed to three stages: pre-incision time, incision time, and post-incision time. The incision stage corresponds to the time between skin incision start and end, and the pre-incision (resp. post-incision) stage refers to the time before (resp. after) it. Different tasks are performed in the three stages. The pre-incision stage includes pre-operative tests, positioning the patient in OR, and anesthetic. The post-incision stage

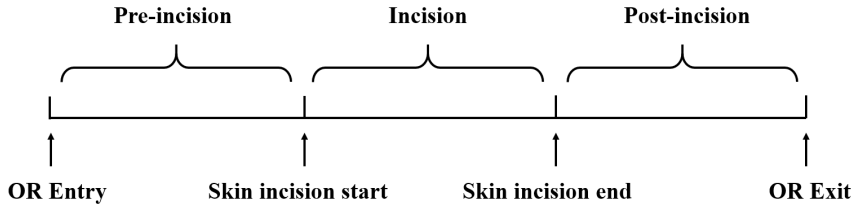


Figure 4.1: OR Timeline for a Cardiac Surgery

includes closing the incision and cleaning up the OR. In cardiac surgeries, these tasks can be largely performed by medical staff or surgical fellows without the presence of the focal surgeon. On the other hand, the incision stage requires relatively high participation of the surgeon. Thus, the incision time is a more accurate measure for a surgeon's working time than the total OR time.

We present the summary statistics for pre-LOS, post-LOS, and total LOS by the four surgery status in Table 4.3. We can see the elective cases have relatively short pre-LOS. This is because most of the elective patients are admitted one day before or on the same day of their surgeries. The elective cases also have the shortest post-LOS, while the emergent cases have the longest one. This reflects the fact that the patients of the elective cases are generally less severe than those of the urgent and emergent cases. Next, Table 4.4 provides the summary statistics for the pre-incision, incision, post-incision, and OR time by the four surgery status. We see that the average incision and OR time are longer for urgent and emergent cases than that for the elective cases. This is not surprising as the non-elective cases tend to be more complicated and thus take longer time to perform. On average, the incision stage consists of 67% of the total OR time.

Table 4.3: Summary Statistics of LOS by Surgery Status (in Days)

Status	Num Obs.	pre-LOS	post-LOS	total LOS
Elective	2479	1.16 (2.92)	8.77 (7.89)	9.93 (9.04)
Urgent	2488	4.90 (9.83)	13.28 (18.55)	18.17 (23.06)
Emergent	374	15.32 (30.11)	25.88 (22.28)	41.21 (37.86)
Salvage	11	8.09 (7.54)	20.55 (12.15)	28.64 (14.41)
All	5352	3.9 (11.19)	12.09 (15.58)	15.99 (21.19)

Table 4.4: Summary Statistics of OR and Incision Time by Surgery Status (in Hours)

Status	Num Obs.	Pre-incision	Incision	Post-incision	OR time
Elective	2479	1.48 (0.28)	4.52 (1.56)	0.73 (0.38)	6.79 (1.79)
Urgent	2488	1.51 (0.31)	4.88 (1.76)	0.78 (0.40)	7.24 (1.99)
Emergent	374	1.48 (0.45)	5.80 (2.31)	0.87 (0.47)	8.31 (2.59)
Salvage	11	1.18 (0.51)	5.96 (2.27)	0.95 (0.49)	8.28 (2.60)
All	5352	1.49 (0.31)	4.78 (1.75)	0.76 (0.40)	7.11 (1.99)

We then report the summary statistics for the following surgical outcomes. The first two are the total time in ICU after the surgery and post-LOS in the hospital (in days). The total time in ICU accounts for both the initial ICU visit and potential revisits. In addition, we consider three binary outcomes: reoperation, readmission to the hospital, and mortality. The reoperation accounts for all types of causes for returning to OR: bleeding, valve dysfunction, MI, aortic disease, other cardiac and non-cardiac reasons. However, we exclude the reoperations within 24 hours after the surgery due to acute bleeding. This is because these acute reoperations increase the surgeon's workload on the same day, but they are not documented in the STS data. The mortality includes death in 30 days after the surgery regardless of the location (e.g., in hospital or at home). The summary statistics of these surgical outcomes are reported in Tables 4.5 and 4.6. Not surprisingly, we see the urgent cases on average are associated with worse outcomes than the elective ones, and the emergent cases have the worst average outcomes among the three categories.

Table 4.5: Binary Surgical Outcomes by Status

Status	Num Obs.	Reoperation	Mortality	Readmission
Elective	2479	0.035	0.016	0.089
Urgent	2488	0.076	0.031	0.104
Emergent	374	0.251	0.061	0.118
Salvage	11	0.364	0.455	0.200
All	5352	0.070	0.027	0.098

Finally, we determine the block status on the surgeon-day level using the block schedule data. We find that out of the 1,744 surgeon-day pairs with block information, 1,343 (77%) of them are in block schedule, i.e., the surgeon has a block assignment on that day, while the remaining 401 (23%) pairs happen out of schedule. On average, a surgeon performs 1.4

Table 4.6: Total ICU Time and post-LOS by Status (in Days)

		Tot ICU			post-LOS		
Status	Num Obs.	Mean	Median	Std	Mean	Median	Std
Elective	2479	3.61	2.00	5.55	8.77	6.00	7.89
Urgent	2488	5.88	2.88	11.36	13.28	8.00	18.55
Emergent	374	13.29	7.69	17.35	25.88	21.00	22.28
Salvage	11	15.69	9.33	13.03	20.55	19.00	12.15
All	5352	5.37	2.67	10.09	12.09	7.00	15.58

(resp. 1.2) cases a day if he or she is in (resp. out of) block schedule. In total, we have 2,010 cases classified as in block schedule, 482 as out of block schedule, and 2,860 cases as unknown because the block information is missing for their surgery dates.

4.3 Econometric Framework

In this section, we develop the econometric framework for identifying the effect of daily workload on surgery duration and outcomes. For each case i, denote its surgeon and surgery date by s and t respectively. We consider two measures for the surgeon's daily workload. The first measure $NumCases_i$ is the total number of cases performed by surgeon s on day t. The second measure $SumInc_i$ represents the total incision time of other cases (excluding i) by surgeon s on day t. The summary statistics of the two daily workload measures is reported in Table 4.7 below. From the results, we see it is very common for a surgeon to perform multiple cases a day in our sample: the median of $NumCases_i$ is two for both the full sample and the block sample. That is, for half of cases in our sample, their surgeons perform at least two cases on the day of surgery. This is also reflected by the average of $SumInc_i$, which is 3.27 (resp 3.14) hours for the full (resp. block) sample. Note that if the surgeon performs only one case in a day, we would have $SumInc_i$ equal to zero by its definition. The two daily workload measures are highly correlated. The correlation is 0.91 and 0.92 for the full and block sample, respectively. The first measure $NumCases_i$ is easier to interpret, but it has less variation than $SumInc_i$ as it is forced to take integer values.

We control for a variety of demographic, medical, operative, and operational factors as independent variables in our estimation. The demographic variables include patient's gen-

Table 4.7: Summary Statistics of Daily Workload for Full and Block Sample
(Full Sample: N = 5,352, Block Sample: N = 2,492)

	Full Sample			Block Sample		
Workload	Mean	Median	Std	Mean	Median	Std
$NumCases_i$	1.69	2.00	0.69	1.66	2.00	0.65
$SumInc_i$	3.27	3.17	3.58	3.14	3.10	3.38

der, age, and race. The medical variables include 18 risk factors or preoperative conditions of the patient. The operative variables include the surgery status, the patient's admission type, the surgery type, the surgeon's identifier, an indicator for aorta procedure, and the number of arteries bypassed in the surgery. The admission type refers to the channels for the patient to be admitted to the hospital. The four admission types and their numbers of observations are: elective (3,777), emergency department (448), transfer-in (1,117), and other (10). The number of arteries bypassed is calculated if the coronary artery bypass is performed in the surgery, and set to zero otherwise. Besides, we include five operational variables: dummies for the weekday, month, and year of the surgery, the pre-LOS, and the block schedule status (in-schedule, out-of-schedule, or unknown). A detailed description of the independent variables used in our estimation can be found in Appendix D.1.

We represent the above independent variables (plus a constant) by X_i for case i. To estimate the effect of daily workload, we consider the following the econometric set-ups. For continuous dependent variable y_i, we employ the linear model:

$$y_i = X_i\beta + \gamma Workload_i + \varepsilon_i, \tag{4.1}$$

where $Workload_i$ is the daily workload of case i's surgeon on its surgery date, it is measured by the number of cases $NumCases_i$ or the total incision time of other cases $SumInc_i$; the error term ε_i is assumed to follow a normal distribution. On the other hand. for binary dependent variable y_i, we use the following probit model:

$$\begin{aligned} y_i^* &= X_i\beta + \gamma Workload_i + \varepsilon_i, \\ y_i &= \mathbf{1}\{y_i^* > 0\}, \end{aligned} \tag{4.2}$$

where y_i^* is a latent variable and the error term ε_i follows a standard normal distribution.

Note that by (4.1) and (4.2), we are estimating the "average" effect of daily workload for all cases performed by the surgeon on a same day.

The coefficient γ in models (4.1) and (4.2) measures the impact of daily workload on the dependent variable. As a naive approach, we can estimate the coefficients in (4.1) and (4.2) by simple ordinary least squares (OLS) or maximum likelihood estimation (MLE) and interpret γ as the effect of daily workload on the dependent variable y_i. However, this approach ignores the endogeneity in the daily workload of surgeons. That is, the surgeon's daily workload can be affected by patients' severity factors that are unobservable in the data but are considered by the surgeons (e.g., a patient's cognitive state). For example, the surgeon may schedule more cases a day if the unobservable severity factors imply shorter incision times. Consequently, both the dependent variable (incision time) and the daily workload (number of cases) are affected by regressor X_i as well as the unobserved severity factors. If we ignore this, the unobserved severity factors will be absorbed as a part of the error term ε_i and thus violate the strict exogeneity constraint required for consistent estimation. In the case described above, it could introduce a negative bias to the estimates of γ, as the unobservable severity factors are negatively correlated with the daily workload (and we expect γ to be positive). Thus, simple OLS on (4.1) may yield a negative γ even the true effect is positive.

Figure 4.2 shows the endogeneity issue of surgeon's daily workload in more details. To address the endogeneity bias, we employ the IV method that yields consistent estimates of the coefficients. We construct two IVs using the operational data from the cardiac department. We describe and validate the two IVs in next section.

4.3.1 Instrument Variables

To address the endogeneity bias, a valid IV must satisfy the following two conditions. The first is the instrument *relevance*, i.e., the instrument variable has to affect the endogenous variable. The second condition is the instrument *exogeneity*: the instrument variable should impact the outcome only via the endogenous variable (after other exogenous variables are controlled). That is, the instrument variable must be uncorrelated with the unobservable factors that also affect the patient outcomes (Angrist and Krueger, 2001). While the

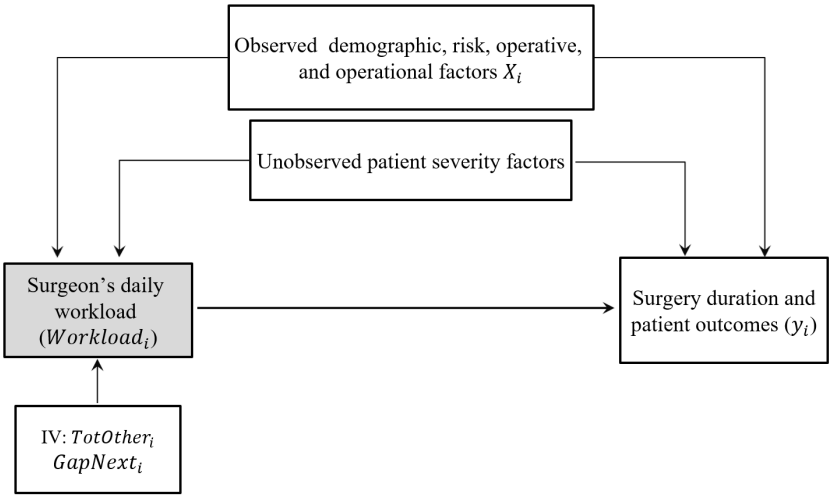


Figure 4.2: Relationship between Surgeon Daily Workload, Observed/Unobserved factors, and Surgery Duration/Patient Outcomes

relevance condition can be tested statistically, the exogeneity condition is generally difficult to test as the error term is unobserved. Thus, the construction and validation of instrument variable depend on the specific economic context of the problem and data generation process. In the following, we propose two IVs using the operational data and demonstrate their validity. The two IVs are both computed on the surgeon-day level, i.e., they are the same across the cases performed by a surgeon on the same day.

The first IV, $\boldsymbol{TotOther_i}$, is the total number of cases performed by other surgeons on the same day of case i. It is supposed to affect the daily workload of the focal surgeon via the channel of *resource sharing* across surgeons. When surgeons perform surgeries on the same day, many resources in the cardiac department are shared by them, e.g., medical staff, medicine, and equipment. Thus, more cases performed by other surgeons on the same day tend to limit the daily workload of the focal surgeon. As such, we expect $\boldsymbol{TotOther_i}$ to be negatively correlated with the focal surgeon's daily workload. This supports the relevance condition. We then argue $\boldsymbol{TotOther_i}$ also satisfies the exogeneity condition. Based on our discussion with the doctors, the surgeons in the cardiac department have high ownership of their patients, and they rarely coordinate when scheduling their own cases. That is, an individual surgeon has little control over others' patients and workload. This suggests the workload of other surgeons should be uncorrelated with the unobservable severity factors of

the focal surgeon's patients.

Besides the total workload by other surgeons, we further construct a novel IV using the block schedule data. Our argument is based on the following behavioral analysis: if a surgeon has to wait for a long period until the next scheduled block, the surgeon may be motivated to "squeeze" more cases into the current day. Thus, we expect the gap to next block (in days) of the focal surgeon, $GapNext_i$, to be positively correlated with the surgeon's daily workload. Thus, $GapNext_i$ satisfies the relevance condition. On the other hand, the block schedule of each surgeon is fixed in advance and adjusted very infrequently (e.g., twice a year), thus it is unlikely for $GapNext_i$ to be correlated with the unobservable severity factors of the surgeon's patients. This supports the exogeneity requirement for $GapNext_i$ as an IV.

For the periods with the block schedule, we can directly compute $GapNext_i$ for each case. However, the block schedule data is missing for a significant proportion of the sample horizon. For the periods without the block schedule, we construct $GapNext_i$ as follows. First, we impute the block schedule on the surgeon-day level using a logistic model. Then, we calculate the *expected* $GapNext_i$ based on our imputation. This enables us to maintain the entire sample for estimation. Indeed, the block schedule data is missing for more than half (53%) of the cases in our sample. So simply dropping the missing periods would largely reduce our sample size.

We impute the block schedule information on the surgeon-day level. Let $Y_{s,t}$ be a binary variable which takes value one if surgeon s has an assigned block on day t (i.e., in block schedule), and zero otherwise. We estimate $Y_{s,t}$ using a logistic model, i.e.,

$$\ln\left[\frac{\mathsf{Pr}(Y_{s,t}=1|X'_{s,t})}{\mathsf{Pr}(Y_{s,t}=0|X'_{s,t})}\right] = X'_{s,t}\beta + e_{s,t}, \tag{4.3}$$

where $X'_{s,t}$ is a set of independent variables and $e_{s,t}$ denotes the error term. Note that we only estimate the model for the surgeon-day combinations that appear in our sample. That is, we implicitly set $\mathsf{Pr}(Y_{s,t}=1)=0$ if surgeon s does not perform any case on day t. The regressor $X'_{s,t}$ contains 23 independent variables (plus a constant term) for surgeon s on day t. For example, it includes the numbers of elective, urgent, and emergent cases by the focal and other surgeons on day t. As we are imputing instead of predicting $Y_{s,t}$, the regressor

$X'_{s,t}$ also contains variables that depend on the "future" information after day t, e.g., the number of days worked by surgeon s in the current calendar week. A complete description of the independent variables in $X'_{s,t}$ can be found in Appendix D.2.

We estimate the logit model (4.3) using the surgeon-day pairs in the block sample, for which the values of $Y_{s,t}$ are known. We then use the estimated model to impute for the periods where the block schedule is missing. With the fitted probability $\mathsf{Pr}(Y_{s,t} = 1|X'_{s,t})$, we calculate the expected gap to next block as:

$$GapNext_i = \sum_{l=1}^{T-1} \Big[\prod_{j=1}^{l-1} (1 - p_{t+j}^{(blk)}) p_{t+l}^{(blk)} \Big] \times l + \prod_{j=1}^{T-1} (1 - p_{t+j}^{(blk)}) \times T, \tag{4.4}$$

where

$$p_{t+l}^{(blk)} := \mathsf{Pr}(Y_{s,t+l} = 1|X'_{s,t+l}).$$

Here t denotes the surgery date of case i; $p_{t+l}^{(blk)}$ is the probability that case i's surgeon has a block on day $t + l$; T is the truncation level for the maximum expected gap. We set it as 14 days in our computation, as we find from the block schedule data that it is rare for a surgeon to stay idle without any block assignment in consecutive two weeks. The calculation in (4.4) is based on an implicit assumption that whether a surgeon has a block assignment is independent across days. Thus, the term $\prod_{j=1}^{l-1}(1-p_{t+j}^{(blk)})p_{t+l}^{(blk)}$ represents the probability that the first block assignment after day t occurs on day $t + l$. Note that for the periods with block schedule data, the corresponding $GapNext_i$ can also be computed by (4.4) with $p_{t+l}^{(blk)}$ set to zero or one according to the block schedule.

4.3.2 Estimation Methods

With the two IVs introduced above, we can estimate the effect of daily workload in models (4.1) and (4.2). We describe the estimation methods below.

For continuous dependent variable y_i, we estimate the linear model (4.1) using the two-stage least squares (TSLS) regression (see Woodridge (2010)). The TSLS estimation is conducted as follows. In the first stage, we regress the daily workload on the exogenous variables X_i and the two IVs using OLS:

$$Workload_i = X_i\beta + \eta_1 TotOther_i + \eta_2 GapNext_i + \xi_i. \tag{4.5}$$

The first stage regression measures the impact of the two IVs on a surgeon's daily workload. For the two IVs to affect the daily workload (i.e., the relevance condition), at least one of η_1 and η_2 should be statistically different from zero. Then in the second stage, we replace $Workload_i$ in (4.1) with its fitted values from (4.5) and estimate γ by OLS. Note that the standard errors in the second stage need to be adjusted as we are plugging in estimates of $Workload_i$.

For binary dependent variable y_i, we use the full maximum likelihood estimation method to estimate the effect γ in the probit model (4.2) (Woodridge, 2010; Cameron and Trivedi, 2013). Specifically, the models for the daily workload in (4.5) and the outcome in (4.2) are estimated jointly under the assumption that the error terms (ε_i, ξ_i) follow a bivariate normal distribution. To capture the endogeneity in daily workload, we allow ε_i and ξ_i to be correlated. Thus, there can be unobservable severity factors that affect the surgical outcomes and daily workload simultaneously.

We find that the distributions of incision time, post-LOS, and total ICU time have long tails on the right end, thus we winsorize them by their 97.5th percentiles to mitigate the impact from extreme values. Our estimation results are robust to other choices of winsorization levels. In addition, for both the linear and probit models in (4.1) and (4.2), we cluster the standard errors by the surgeon's identifier to account for the heteroskedasticity across the cases by different surgeons

4.4 Main Empirical Results

This section provides the main empirical results regarding the impact of daily workload on surgery duration and outcomes. Section 4.4.1 includes the results for the schedule imputation model (4.3) and regression (4.5), which measures the effect of IVs on daily workload. Then, the results for the main models (4.1) and (4.2) are assessed in Section 4.4.2.

4.4.1 Schedule Imputation and Impact of IVs on Daily Workload

In this section, we discuss the empirical results related to schedule imputation and regression (4.5), i.e., the impact of IVs on surgeon's daily workload. For continuous dependent

variables, it is the first stage regression when we estimate the effect of daily workload in (4.1). For binary dependent variables, it is estimated jointly with (4.2) jointly by full MLE method.

4.4.1.1 Schedule Imputation Model

We first provide the results for the schedule imputation model (4.3), which is used when we construct the second IV $GapNext_i$ by (4.4) for the periods without block information. A detailed description of the independent variables in $X'_{s,t}$ is given in Appendix D.2. We estimate the model using the periods with block schedule information. The weekends are dropped as all blocks are assigned on weekdays. This leaves us with 1680 surgeon-day pairs in the block sample for estimation. To measure the model performance, we compute the McFadden's R-squared as

$$R^2 = 1 - \frac{\ln(l^{mod})}{\ln(l^{null})},$$

where l^{mod} is the likelihood from the estimated model, l^{null} is the likelihood from the null model with only intercept. The McFadden's R-squared of the estimated model is 0.31. Besides, the AUC from the model classification is 0.86. Both measures show the imputation model fits the block schedule data well.

In Table 4.8 below, we report the estimated coefficients and average marginal effects (AME) for select variables in model (4.3). Due to space limitation, we only include the variables that have a p-value smaller than 0.05 besides the surgeon and weekday dummies. We have the following findings from the estimated results. First, more elective and urgent cases of other surgeons on day t ($ElecOth_{s,t}$ and $UrgOth_{s,t}$) decrease the probability that surgeon s is assigned a block schedule (i.e., $\mathsf{Pr}(Y_{s,t} = 1)$), while more elective cases ($ElecCur_{s,t}$) and patients admitted ($AdmCur_{s,t}$) by the focal surgeon increase $\mathsf{Pr}(Y_{s,t} = 1)$. This reflects the resource sharing among surgeons in the department on the same day. Next, a late start after 8AM ($StartLate_{s,t}$) of the cases by the focal surgeon s decreases the probability $\mathsf{Pr}(Y_{s,t} = 1)$.

Besides, we find the surgeon's workload around current day t also has explanatory power for $\mathsf{Pr}(Y_{s,t} = 1)$. For example, the number of days worked in the current calendar week ($NumCurWeek_{s,t}$) and the distance to next work day ($DistNext_{s,t}$) are negatively associated

with $\mathsf{Pr}(Y_{s,t} = 1)$. Finally, the variable $WDElecRatio_{s,t}$ denotes the proportion of elective cases by surgeon s in $[t-180, t+180]$ that fall on the same weekday as t. We see its coefficient and AME are significantly positive. This is because blocks tend to be assigned on specific weekdays for each surgeon in adjacent periods. Moreover, more than 90% of elective cases are performed in their surgeons' block schedule. Thus, the ratio $WDElecRatio_{s,t}$ has strong explanatory power for the surgeons' block assignment.

Table 4.8: Select Coefficients in the Logistic Model (4.3)
N = 1,680, R-squared=0.31

Variable	Coefficient	AME
$ElecOth_{s,t}$	-0.206^{***}	-0.020^{***}
	(0.059)	(0.006)
$UrgOth_{s,t}$	-0.140^{*}	-0.014^{*}
	(0.062)	(0.006)
$ElecCur_{s,t}$	1.297^{***}	0.126^{***}
	(0.300)	(0.028)
$AdmCur_{s,t}$	0.332^{***}	0.032^{**}
	(0.099)	(0.010)
$NumCurWeek_{s,t}$	-0.293^{*}	-0.029^{*}
	(0.141)	(0.014)
$DistNext_{s,t}$	-0.125^{**}	-0.012^{**}
	(0.045)	(0.004)
$StartLate_{s,t}$	-1.457^{***}	-0.205^{**}
	(0.351)	(0.065)
$WDElecRatio_{s,t}$	6.598^{***}	0.643^{***}
	(0.790)	(0.081)

Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Select coefficients for schedule imputation model (4.3).

4.4.1.2 Impact of IVs on Daily Workload

With the estimated logistic model, we compute the expected gap to next block following (4.4). The summary statistics of the two IVs are shown in Table 4.9 for both the full sample and the block sample. The statistics of the first IV $TotOther_i$ suggests that other surgeons on average perform 4.1 cases on the surgery day of the focal surgeon. For the second IV $GapNext_i$, we find the average (resp. median) gap to next block schedule is 3.46 days (resp. 2.21 days) for the full sample. The standard deviation of $GapNext_i$ is 3.04 days, reflecting a large variation in the gap to next block. This is because the blocks of each surgeon are

distributed unevenly across the days. We also notice that the statistics of $GapNext_i$ is very similar for the full sample and the block sample. Recall that $GapNext_i$ can be calculated accurately for the block sample. Thus, the comparison here supports the effectiveness of our schedule imputation model (4.3) as the distribution of $GapNext_i$ imputed by the model is close to that calculated directly from the block data.

Table 4.9: Summary Statistics of the IVs for Full and Block Sample
(Full Sample: N = 5,352, Block Sample: N = 2,492)

	Full Sample			Block Sample		
IV	Mean	Median	Std	Mean	Median	Std
$TotOther_i$	4.11	4.00	1.78	4.18	4.00	1.77
$GapNext_i$	3.46	2.21	3.04	3.40	2.00	3.22

In Table 4.10, we show the estimated coefficients and standard errors for the two IVs in the OLS regression (4.5). The daily workload in (4.5) is specified as the number of cases ($NumCases_i$) or the total incision time of other cases ($SumInc_i$) by the focal surgeon on the given day. The estimation results are reported for the full sample and block sample respectively. We see the IVs are all statistically significant with expected signs. For the full sample, we see one more case by other surgeons on the same day leads to a reduction of 0.074 (resp. 0.37 hours) in the number of cases (resp. total incision time of other cases) by the focal surgeon. This reflects the resource sharing among surgeons, i.e., more cases by other surgeons limit the workload by the focal surgeon. On the other hand, the gap to next block $GapNext_i$ is associated with higher daily workload of the focal surgeon. This shows the effect of "squeezing in" more cases by the focal surgeon if the next scheduled block is far away.

The regression results in Table 4.10 show that the two IVs indeed significantly affect the daily workload of the focal surgeon, thus validating the relevance condition required for IV. We also notice that the results are similar when we estimate (4.5) using the block sample, for which the second IV $GapNext_i$ can be accurately computed. This further supports the validity and robustness of the two IVs used in our estimation. Besides, we estimate regression (4.5) using the samples consisting of elective and non-elective cases respectively. The results are summarized in Table 4.11. It shows the two IVs still have expected signs for

the elective and non-elective samples, although the impact of $GapNext_i$ on the number of cases becomes insignificant. This is potentially due to the smaller sizes of the elective and non-elective samples.

Table 4.10: Impact of IVs on Daily Workload (Full and Block Sample)

	Full Sample		Block Sample	
IV	$NumCases_i$	$SumInc_i$	$NumCases_i$	$SumInc_i$
$TotOther_i$	-0.074^{***}	-0.369^{***}	-0.070^{***}	-0.357^{***}
	(0.006)	(0.028)	(0.008)	(0.040)
$GapNext_i$	$0.006^{\dagger}$	0.045^{**}	0.009^{*}	0.064^{**}
	(0.003)	(0.016)	(0.004)	(0.022)
Num Obs.	5345	5344	2490	2489
Adj R^2	0.147	0.161	0.176	0.179

Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Coefficients and standard errors of two IVs in (4.5) for full sample and block sample.

Table 4.11: Impact of IVs on Daily Workload (Elective and Non-elective Sample)

	Elective Sample		Non-elective Sample	
IV	$NumCases_i$	$SumInc_i$	$NumCases_i$	$SumInc_i$
$TotOther_i$	-0.065^{***}	-0.338^{***}	-0.081^{***}	-0.392^{***}
	(0.008)	(0.039)	(0.008)	(0.040)
$GapNext_i$	0.006	0.053^{*}	0.005	0.042^{*}
	(0.005)	(0.026)	(0.004)	(0.021)
Num Obs.	2474	2474	2871	2871
Adj R^2	0.122	0.142	0.160	0.171

Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Coefficients and standard errors of two IVs in (4.5) for elective and non-elective samples.

Finally, Table 4.12 reports the coefficients η_1 and η_2 for the two IVs when we estimate (4.5) and (4.2) jointly by full MLE for binary dependent variables. Due to space limitation, we only show the estimation results for the full sample. Unlike the OLS estimates in Tables 4.10 and 4.11, the estimated η_1 and η_2 from the full MLE may vary when we choose different dependent variable y_i in (4.2). Thus in Table 4.12, we show the estimation results of (4.5) for three binary outcomes separately[3], i.e., reoperation, readmission, and mortality. We

[3]The sample size varies for the three outcomes. This is because some levels of categorical variables (e.g., specific procedure types) lead to perfect predictions of the binary outcome, thus the corresponding observations are dropped from the estimation.

find that the two IVs have expected signs for all the three outcomes and both workload measures. Furthermore, the estimated coefficients are statistically significant in most cases. This supports the validity of the two IVs for estimating the effect of daily workload on binary outcomes.

Table 4.12: Estimated Coefficients of IVs by Full MLE of (4.5) and (4.2) (Full Sample)

	Reoperation		Readmission		Mortality	
IV	$NumCases_i$	$SumInc_i$	$NumCases_i$	$SumInc_i$	$NumCases_i$	$SumInc_i$
$TotOhter_i$	-0.074^{***}	-0.369^{***}	-0.072^{***}	-0.359^{***}	-0.075^{***}	-0.373^{***}
	(0.008)	(0.051)	(0.008)	(0.052)	(0.008)	(0.051)
$GapNext_i$	0.005	$0.043^{\dagger}$	$0.007^{\dagger}$	0.053^{*}	$0.006^{\dagger}$	0.047^{*}
	(0.004)	(0.022)	(0.004)	(0.026)	(0.004)	(0.022)
Num Obs.	5345	5345	5116	5116	5081	5081

Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Coefficients and standard errors of two IVs in (4.5) when we estimate (4.2) and (4.5) jointly by full MLE for reoperation, readmission, and mortality.

4.4.2 Effect of Daily Workload on Surgery Duration and Patient Outcomes

In this section, we provide the estimation results for our main models (4.1) and (4.2) to analyze how surgeon's daily workload impacts surgery duration and patient outcomes. To save space, we only show the results when we measure surgeon's daily workload by the number of cases performed, i.e., with $Workload_i$ specified as $NumCases_i$ in (4.1) and (4.2). The results for $SumInc_i$ (the total incision time of other cases) as the daily workload measure are qualitatively similar, and are reported in Tables D.5 and D.6 in Appendix D.3. Note that the estimation results here capture the average effect for all the cases performed by the surgeon in a day.

In Table 4.13, we show the estimation results for the full sample. For the three continuous dependent variables ("Incision time", "Post-LOS", and "Total ICU time" columns), we show the estimated γ in (4.1) and its standard errors. For the three binary dependent variables ("Reoperation", "Readmission", and "Mortality" columns), we report the estimated average marginal effects (AME) of daily workload as they are easier to interpret. The AME calculates the marginal effect of daily workload for each case given the other independent variables are

unchanged, and then averages across the resulting effect estimates. In Panel A, we show the estimation results from the TSLS and full MLE with the two IVs, as described in Section 4.3.2. For comparison, we also show in Panel B the results when we ignore the endogeneity bias, i.e., from simple OLS on (4.1) or MLE on (4.2). The estimated coefficient γ in (4.2) for the three binary dependent variables are given in Tables D.7 and D.8 in Appendix D.3.

To account for the heterogeneity in the impact of daily workload, we further report in Table 4.14 the estimated effects when we estimate the models using elective patients (Panel A) and non-elective patients (Panel B) respectively. The non-elective patients include those in the urgent, emergent, and salvage surgeries. For the two subsamples, we still estimate models (4.1) and (4.2) following the methods in Section 4.3.2. The effectiveness of the two IVs for the full and sub-samples is validated in Section 4.4.1.

Table 4.13: Estimated Effects of Daily Workload (Number of Cases) on Surgery Duration and Patient Outcomes: Full Sample

	Continuous y_i: Coefficients			Binary y_i: AME		
	Incision time	Post-LOS	Total ICU time	Reoperation	Readmission	Mortality
Panel A: Full	0.430^{*}	1.408^{*}	1.031^{*}	0.030^{**}	$0.065^{\dagger}$	0.010
	(0.217)	(0.565)	(0.482)	(0.011)	(0.039)	(0.011)
Num Obs.	5345	5344	5319	5345	5116	5081
Panel B: Full (w/o IV)	-0.102^{**}	−0.020	−0.002	−0.002	−0.000	0.010^{*}
	(0.039)	(0.147)	(0.086)	(0.004)	(0.008)	(0.004)
Num Obs.	5345	5344	5319	5345	5116	5081

The estimated effects of surgeon's daily workload (number of cases) on surgery duration and patient outcomes for the full sample. We report the estimated coefficients in (4.1) for the three continuous dependent variables, and the AME from (4.2) for the three binary dependent variables. Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

4.4.2.1 Impact of workload on incision time

We first discuss the impact of workload on the surgery incision time. First, we see from the "Incision time" in Panel A of Table 4.13 that higher daily workload tends to increase the incision time of the cases performed by the focal surgeon. In particular, adding one more case increases the incision time of each case performed by the surgeon by 0.43 hour (26 minutes). This translates to a 9% relative increase of the average incision time. The effect is statistically significant at the 5% level. On the other hand, if we ignore the endogeneity in

Table 4.14: Estimated Effects of Daily Workload (Number of Cases) on Surgery Duration and Patient Outcomes: Elective and Non-elective Sample

	Continuous y_i: Coefficients			Binary y_i: AME		
	Incision time	Post-LOS	Total ICU time	Reoperation	Readmission	Mortality
Panel A: Elec	0.378^{**}	−0.288	0.121	0.017	0.029	0.060
	(0.140)	(1.015)	(0.729)	(0.039)	(0.066)	(0.050)
Num Obs.	2474	2474	2454	2394	2398	1897
Panel B: Non-elec	0.486	3.004^{*}	$1.906^{\dagger}$	0.049^{*}	0.082^{*}	−0.004
	(0.341)	(1.501)	(1.054)	(0.021)	(0.039)	(0.012)
Num Obs.	2871	2870	2865	2871	2697	2769

The estimated effects of surgeon's daily workload (number of cases) on surgery duration and patient outcomes for the elective and non-elective sample. We report the estimated coefficients in (4.1) for the three continuous dependent variables, and the AME from (4.2) for the three binary dependent variables. Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

daily workload and estimate the model by OLS, the results become completely opposite. By the 'Incision time" column in Panel B, we see the coefficient of $\boldsymbol{NumCases_i}$ is significantly negative. The negative coefficient implies surgeons may schedule more cases a day if the unobservable factor implies shorter incision times. This shows it is essential to control the endogeneity in the daily workload by proper IVs.

A priori, the surgeon's daily workload may impact the incision time in both directions. First, surgeons may "speed up" the surgeries when they have more cases to perform in a day, leading to a shorter incision time. This type of speedup effect is found in, e.g., Kc and Terwiesch (2009). On the other hand, surgeons may take more time to complete their tasks due to fatigue associated with high daily workloads. For example, Berry Jaeker and Tucker (2017) finds patient's LOS increases in the occupancy after the occupancy level exceeds a tipping point. Our empirical results here support the second direction, i.e., higher daily workloads of surgeons leads to longer incision times. This can be explained as follows. First, as cardiac surgeries are complex and delicate procedures, it is very difficult for surgeons to speed up in their operations. Second, as a demanding task, performing multiple cases a day can lead to severe fatigue of surgeons, both physically and mentally. For these reasons, the impact of surgeon's fatigue outweighs other potential channels for speedup, and causes longer incision time for each case performed.

We notice that by model (4.1), the expected total incision time of all cases performed by

a surgeon in a day increases not linearly, but quadratically in the number of cases. This is because the coefficient γ measures the average effect of daily workload for all cases performed by the surgeon. To see this, index the cases by a surgeon on a given day by $i = 1, 2, ..., n$. By (4.1), their total expected incision time is given by

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} X_i\beta + \sum_{i=1}^{n} \gamma n = \sum_{i=1}^{n} X_i\beta + \gamma n^2. \tag{4.6}$$

This quadratic structure highlights the negative effect of surgeon's high daily workload. On the other hand, it suggests the hospital can improve its performance by 'smoothing" surgeon's workloads across days. For example, if we assign two cases by a surgeon in a day to two separate days, their expected total incision time would decrease by $0.43\times2^2-0.43\times2 = 0.86$ hour. Following this spirit, we propose in Section 4.5 a scheduling model for minimizing the total expected incision time.

We also estimate the impact on incision time for elective and non-elective cases separately. This captures the potential heterogeneity in the effect for different types of patients. By the "Incision time" column in Table 4.14, we see that the effect of daily workload on incision time is statistically significant (at 1% level) for elective cases (Panel A), but it turns out to be insignificant for non-elective cases (Panel B). The magnitude of the impact for elective cases is similar to that for the full sample. In particular, performing one more case increases the incision time of each elective case by 23 minutes, which is equivalent to an 8% increase on average. One possible explanation for the difference in effects is the non-elective cases (urgent and emergent cases) are generally more urgent and time sensitive, thus their incision time is less impacted by surgeon's daily workload.

4.4.2.2 Impact of workload on patient outcomes

Next, we examine the effect of daily workload on patient outcomes, including two continuous outcomes (post-LOS and total ICU time) and three binary outcomes (reoperation, readmission, and mortality). The estimation results are reported in the corresponding columns in Table 4.13 for the full sample, and in Table 4.14 for elective and non-elective samples separately.

By the "post-LOS" and "total ICU time" columns in Panel A of Table 4.13, we find that higher daily workload increases the post-LOS and total ICU time. Specifically, adding one more case increases the total ICU time and post-LOS by 1.05 and 1.45 days, respectively, for the cases performed by the surgeon on the same day. This is equivalent to a 12% increase for post-LOS and a 19% increase for total ICU time. However, the effects becomes insignificant for both outcomes if we do not control the endogeneity by IVs, as shown in Panel B of Table 4.13.

We then compare the effects on post-LOS and total ICU time for elective and non-elective cases. For the post-LOS, we see the coefficients of daily workload are statistically significant for the non-elective cases, but insignificant for the elective cases ("Post-LOS" column in Table 4.14). Furthermore, the magnitude of the effect is larger for the non-elective cases than that for the full sample. For example, adding one more case leads to 3.13 more days in the post-LOS of non-elective patients, which is equivalent to a 21% relative increase. This effect is more than twice as that for the full sample (1.45 days). We also find a similar heterogeneous effect for total ICU time ("Total ICU time" column in Table 4.14). The estimated coefficient of $NumCases_i$ is 1.967 days for the total ICU time of non-elective cases, which is almost twice as large as that for the full sample. On the other hand, the daily workload does not significantly impact the total ICU time of elective cases.

The heterogeneous effect for the post-LOS and total ICU time can be explained as follows. The non-elective cases are generally more urgent and complicated than the elective ones as the non-elective patients are usually more severe. Thus, the outcomes of non-elective cases tend to be more sensitive to surgeon's fatigue due to daily workload. On the other hand, the elective patients are on average less severe, and they recover more quickly after the surgery. This is shown by the summary statistics of post-LOS and total ICU time in Table 4.6, as we see the average post-LOS and total ICU time are much longer for the non-elective cases than that for the elective ones. Moreover, the standard deviations of total ICU time and post-LOS are also much larger for the non-elective cases, implying there is more variation in the surgical outcome of non-elective patients.

The effect on total ICU time and post-LOS, as identified above, is important for the hospital to manage its patient flow. With longer post-surgery LOS, the patients would

occupy the beds in ICU or ward for more time, thus increasing the demand for downstream resources and reducing the system throughput efficiency. This can lead to overcrowding in the perioperative environment and delay in surgeries (Zenteno et al., 2016). Besides, the ICU is often congested and extremely expensive to operate (e.g., Halpern (2011)). Given almost all the patients (99%) in our sample are sent to the ICU after surgery, understanding the factors that impact their ICU recovery time provides a potential method for managing ICU congestion. Similar to (4.6) for incision time, the total expected ICU time and post-LOS also grow quadratically in the number of cases of the surgeon, so smoothing the surgeon's workload across days can reduce the total time needed for recovery of the patients. According to the estimation results for the full sample, moving two cases performed by a surgeon in a day to two different days leads to a reduction of 2.06 (resp. 2.82) days in total expected ICU time (resp. post-LOS).

We then check the impact of workload on three binary patient outcomes (reoperation, readmission, and mortality), which are estimated by full MLE on (4.5) and (4.2) jointly. The estimated AME and their standard errors are reported in the corresponding columns in Table 4.13 for the full sample, and in Table 4.14 for the elective and non-elective subsamples. The estimated coefficient γ in (4.2) for the three binary outcomes are provided in Tables D.7 and D.8 in Appendix D.3.

By the "Reoperation" column in Panel A of Table 4.13, we see a higher daily workload increases the probability of reoperation when estimated from the full sample. Specifically, adding one more case leads to a three percentage points increase in the reoperation probability for each case performed by the surgeon on the same day. The magnitude of such an increase seems large at first, as the original reoperation probability is only 7% (see Table 4.5). However, we note that the median of surgeons' daily workload is two cases, thus adding one more case is equivalent to a 50% increase in daily workload. The large impact of workload on medical outcome is also observed in the literature. For example, Kc and Terwiesch (2009) finds that 10% increase in overwork increases the morality rate by 2.2 percentage points, which is a 32% relative change in their setting.

When we estimate the impact on reoperation probability for elective and non-elective patients separately, we find that daily workload significantly increases the reoperation prob-

ability for non-elective patients ("Reoperation" column in Table 4.14). In particular, adding one more case increases the reoperation probability by 4.9 percentage points for the non-elective cases. This AME is larger than that for the full sample (three percentage points). However, no statistically significant impact is observed for elective patients. Such heterogeneity reconciles the discussion for post-LOS and total ICU time in the previous section: the non-elective cases are generally more severe and urgent than the elective ones, thus their outcomes are more sensitive to surgeon's fatigue due to high daily workload.

A similar effect of daily workload is also observed for hospital readmission probability. The "Readmission" column in Panel A of Table 4.13 shows that higher daily workload increases the probability of readmission. Specifically, adding one more case increases the readmission probability by 6.5 percentage points when estimated from the full sample. Moreover, as shown by the "Readmission" column in Table 4.14, the effect on readmission probability is statistically significant for the non-elective cases, but not so for the elective cases. The magnitude of the AME is also larger for the non-elective cases (0.082) than that for the full sample (0.065). These results are consistent with other outcomes, i.e., post-LOS, total ICU time, and reoperation.

For the 30-day mortality, however, the effect of daily workload vanishes. By the "Mortality" column in Tables 4.13 and 4.14, we see the AME of daily workload are insignificant for both the full sample and the two subsamples once we control the endogeneity bias using IVs. The insignificant effect for mortality seems surprising, especially given the negative impact of daily workload on other outcomes. One possible reason for such difference is surgeons tend to pay greater attention to the patients with high risk of death, thus the mortality rate is less impacted by surgeon's daily workload.

The consistent results in Tables 4.13 and 4.14 demonstrate the negative impact of surgeon's high daily workload on multiple patient outcomes. Such effect is especially significant for non-elective patients, who are generally more severe. Our results provide new evidence for the link between high workload level and worse patient outcomes (see, e.g., Kc and Terwiesch (2009) and Kuntz et al. (2015)). From the managerial perspective, it suggests that when hospitals design their surgery schedules, they should take into account the effect of surgeon's fatigue due to high daily workload. We shed light on this direction in our next

section.

4.5 A Surgery Scheduling Model with Impact from Daily Workload

In this section, we propose a shuffling model for surgery scheduling based on the effect of daily workload estimated in the previous sections. While there is a rich literature on surgery scheduling, most of it assumes exogenous distributions for the surgery duration and patient outcomes (e.g., post-LOS). That is, the behavioral impact from surgeon's workload is ignored. However, our econometric analysis suggests that higher daily workloads for surgeons are associated with longer incision times and worse surgical outcomes. Thus, our model in this section sheds lights on the potential benefit we can obtain by incorporating these effects in the surgical scheduling.

There are different layers of decisions to be made in OR scheduling, such as determining the number of ORs to open, assigning surgeries to ORs and surgeons, choosing surgery date and starting time, and sequencing of surgeries within each OR in a day. Different studies usually focus on one or more of these aspects while assume the others to be fixed. For example, Shylo et al. (2013) consider the allocation of surgeries to blocks under a block booking framework. Denton et al. (2010) determine the number of ORs to open and the assignment of surgeries to ORs. The sequencing of surgeries is considered in Gupta (2007) and Batun et al. (2011). The model in Wang and Pourghannad (2020) solves for the optimal matching between patients and surgeons.

In our model, we consider the decision to switch two cases that are performed within the same calendar week in our sample. In the switching, we keep the surgeon of each surgery unchanged and only switch the surgery dates. In addition, we require the number of days worked by each surgeon in the week does not increase after switching. These constraints lead to three properties of the shuffling model that facilitate its practical implementation. First, the total number of cases by all surgeons performed on each day is unchanged. Thus, the cardiac department does not need to adjust its resource allocation across days or change the schedule of its ORs shared with other departments. Second, the patients assigned to each

surgeon remain unchanged. This reflects the reality that surgeons in the cardiac department have high ownership of their patients. Finally, the number of working days in the week does not increase for each surgeon. This is imposed because it would be difficult to ask the surgeons to work for more days than they originally do, although this can smooth their daily workload.

Our objective is to minimize the total expected incision time of all cases in a given week. By the empirical results in Section 4.4.2.1, the total incision time will endogenously depend on surgeon's daily workload. We show this objective can be formulated as a mixed-integer quadratic programming (MIQP) problem. That is, the objective is quadratic in the decision variables, the constraints are linear, and some of the decision variables are forced to take integer values. Note that as we can only switch the dates of surgeries with their surgeons unchanged in our model, the corresponding results provide a conservative estimate for the benefit of incorporating the effect of surgeons' workloads in surgery scheduling. If we introduce more flexibility to the model, e.g., changing the number of ORs or the surgeons assigned to each patient, we may be able to achieve larger improvements.

We develop and solve the shuffling model for each calendar week separately. We impose the following constraints on the switching of different types of cases. First, the elective cases can be moved to any weekday of the week; second, the urgent cases can only be moved to one day before or after the original surgery date; finally, the emergent and salvage cases cannot be moved. These constraints reflect the fact that the hospital has less flexibility in changing the schedule of non-elective cases. Thus, we can switch two elective cases regardless of their original surgery dates. However, for an urgent case, we can only switch it with the cases that are performed one day before or after its surgery date. Besides, we assume the cases performed on weekends cannot be moved.

Let A be the set of movable cases described above. For cases $i, j \in A$, we introduce the binary variable $x_{i,j}$, which takes value one if cases i and j are switched and zero otherwise. By symmetry, we have $x_{i,j} = x_{j,i}$. In our model, we only consider one-round shuffling, i.e., the pairs (i, j) that are switched do not overlap. This constraint facilitates the formulation of the model as an MIQP problem. In addition, the model solution from the one-round shuffling

is easier to interpret and implement. This constraint can be mathematically expressed as:

$$\sum_j x_{i,j} \leq 1, \quad \forall i \in A. \tag{4.7}$$

Denote the surgeons and weekdays by set S and T, respectively. For each case $i \in A$, we represent its surgeon and original surgery date by $\tilde{s}(i)$ and $\tilde{t}(i)$. As we do not change the surgeon assigned to each case in our shuffling model, it will be meaningless to switch the cases that are performed on the same day. Besides, as we measure the surgeon's daily workload by the number of cases performed, switching two cases by the same surgeon on different days will not change the surgeon's daily workload. Thus we only consider the switching between cases that are performed by different surgeons on different days. Thus we have

$$x_{i,j} = 0, \quad \text{if } \tilde{t}(i) = \tilde{t}(j) \text{ or } \tilde{s}(i) = \tilde{s}(j). \tag{4.8}$$

As the urgent cases can only be moved to one day before or after the original date, we impose

$$x_{i,j} = 0, \quad \text{if } i \text{ or } j \in A_{urg} \text{ and } |\tilde{t}(i) - \tilde{t}(j)| > 1, \tag{4.9}$$

where A_{urg} denotes the set of urgent cases.

We then formulate the constraint that the number of days worked by each surgeon does not increase after switching. Let $C_{s,t}^{(1)}$ and $C_{s,t}^{(2)}$ denote the sets of elective and non-elective cases by surgeon s on day t before switching. To compute the number of cases by surgeon s on day t after switching, we need to account for both the cases that are moved out and the cases that are moved in. To characterize the cases that are moved out, we introduce the set:

$$O_{s,t}^{(\iota)} = \left\{(i,j) | i \in C_{s,t}^{(\iota)}, \tilde{s}(j) \neq s\right\}$$

for $\iota \in \{1,2\}$. It is easy to verify the set $O_{s,t}^{(\iota)}$ has the following interpretations: suppose $(i,j) \in O_{s,t}^{(1)}$ (resp. $(i,j) \in O_{s,t}^{(2)}$) and $x_{i,j} = 1$, then the number of elective (resp. non-elective) cases performed by surgeon s on day t is reduced by one. This is because the case i by surgeon s on day t is switched with another case j, which must be performed by another surgeon on a different day by constraint (4.8). Similarly, to account for the cases that are moved in, we define the set:

$$I_{s,t}^{(\iota)} = \left\{(i,j) | \tilde{t}(i) = t, j \in C_{s,t'}^{(\iota)} \text{ for some } t'\right\}$$

for $\iota \in \{1, 2\}$. The set $I_{s,t}^{(\iota)}$ can be interpreted as follows: if $(i, j) \in I_{s,t}^{(1)}$ (resp. $(i, j) \in I_{s,t}^{(2)}$) and $x_{i,j} = 1$, then the number of elective (resp. non-elective) cases performed by surgeon s on day t is increased by one. This is because case j performed by surgeon s is now moved to day t, replacing the original case i performed by another surgeon.

With the above preparation, we can express the number of elective and non-elective cases by surgeon s on day t after switching as

$$\tilde{n}_{s,t}^{(1)} = n_{s,t}^{(1)} - \sum_{(i,j) \in O_{s,t}^{(1)}} x_{i,j} + \sum_{(i,j) \in I_{s,t}^{(1)}} x_{i,j}, \tag{4.10}$$

and

$$\tilde{n}_{s,t}^{(2)} = n_{s,t}^{(2)} - \sum_{(i,j) \in O_{s,t}^{(2)}} x_{i,j} + \sum_{(i,j) \in I_{s,t}^{(2)}} x_{i,j}, \tag{4.11}$$

where $n_{s,t}^{(1)}$ and $n_{s,t}^{(2)}$ denote the numbers of elective and non-elective cases by surgeon s on day t before the switching, i.e., $n_{s,t}^{(1)} = |C_{s,t}^{(1)}|$ and $n_{s,t}^{(2)} = |C_{s,t}^{(2)}|$. As we discussed, the second and third terms in equations (4.10) and (4.11) represent the cases that are moved out of or into day t for surgeon s.

Then, the constraint on the number of days worked by each surgeon can be formulated as

$$\sum_{t \in T} \mathbf{1}\{\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)} > 0\} \leq N_s, \, \forall s \in S.$$

where N_s is the number of days worked by surgeon s before switching. However, this constraint is non-linear as it involves the indicator function. To facilitate implementation, we employ the following equivalent linear formulation:

$$\sum_{t \in T} z_{s,t} \leq N_s, \, \forall s \in S, \tag{4.12}$$

with

$$z_{s,t} \leq M \cdot (\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)}) \text{ and } z_{s,t} \geq m \cdot (\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)}). \tag{4.13}$$

Here $z_{s,t} \in \{0, 1\}$ are binary variables; M (resp. m) is a big (resp. small) enough constant number. It is easy to verify $z_{s,t}$ satisfies

$$z_{s,t} = \mathbf{1}\{\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)} > 0\}.$$

Next, we formulate the objective function to minimize the total expected incision time of all cases in the week. For each case i, we decompose its expected incision time l_i to two parts

$$y_i = l_i + d_i,$$

with

$$l_i = X_i\beta \text{ and } d_i = \gamma \cdot NumCases_i. \tag{4.14}$$

where $NumCases_i$ denotes the number of cases performed by the surgeon of case i on its surgery date; the coefficients β and γ are estimated by TSLS on model (4.1) as reported in Section **??**. Thus, the part d_i captures the effect of $NumCases_i$ on the incision time. To focus on the impact of daily workload, we assume the part l_i remains unchanged after switching. However, the part d_i will be affected by the new daily workload of the surgeon. Combining the two parts, the total expected incision time of surgeon s on day t can be expressed as

$$Y_{s,t} = \sum_{i \in C_{s,t}} l_i - \sum_{i \in C_{s,t}} l_i \cdot x_{i,j} + \sum_{\tilde{t}(i)=t, \tilde{s}(j)=s} l_j \cdot x_{i,j} + D_{s,t}. \tag{4.15}$$

The first term is the sum of l_i before switching; the second and third terms represent the changes in the sum of l_i due to switching. The final term capture the impact from the new daily workload (number of cases) of surgeon s on the total incision time. By (4.6), it can be expressed as

$$D_{s,t} = \gamma\left(\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)}\right)^2, \tag{4.16}$$

which is quadratic in the new daily workload $\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)}$. We can also incorporate the heterogeneous effect for elective and non-elective cases as shown in Section 4.4.2.1. In this case, the term $D_{s,t}$ can be expressed as

$$D_{s,t} = \left(\tilde{n}_{s,t}^{(1)} + \tilde{n}_{s,t}^{(2)}\right)\left(\gamma^{(1)}\tilde{n}_{s,t}^{(1)} + \gamma^{(2)}\tilde{n}_{s,t}^{(2)}\right), \tag{4.17}$$

where $\gamma^{(1)}$ and $\gamma^{(2)}$ denote the effects for elective and non-elective cases respectively.

Our objective is to minimize the total expected incision time in the week, i.e.,

$$\min \sum_{s,t} Y_{s.t}$$

where $Y_{s,t}$ is given by (4.15). Apparently, the objective is quadratic in the decision variable $x_{i,j}$. Then, the MIQP is formulated by combining (4.15) and the linear constraints in (4.7) – (4.13).

4.6 Conclusion and Discussion

In many human-involved service systems, the service time and quality are found to be endogenously affected by the level of workload. To shed light on this topic, we empirically investigate the relationship between workload and performance in the context of cardiac surgery. Specifically, we study how surgery duration and patient outcomes are impacted by surgeon's daily workload, i.e., number of cases performed in a day. Using a detailed data set of cardiac surgeries, we find that higher surgeons' daily workloads lead to longer surgery durations and worse patient outcomes, including longer post-surgery LOS in the ICU and hospital, as well as higher likelihoods of reoperation and readmission. Our study provides new evidence for the negative impact of surgeon's fatigue due to high daily workload. It suggests hospitals may improve their surgery performance if they could smooth surgeons' workloads across days. Based on our findings, we develop a surgery scheduling model that incorporates the effect of surgeon's workload.

When identifying the true effect of surgeon's workload, it is crucial to address the endogeneity bias that arises from unobservable risk factors. To handle this challenge, we develop two IVs using the operational factors in the cardiac department. In particular, we leverage a novel data set, which is the surgeons' block schedule assigned by the department. We find surgeons tend to schedule more surgeries if their next scheduled block is far away. This introduces exogenous variation in surgeon's daily workload, which is essential for constructing proper IVs. We also find there is substantial heterogeneity in the effects of daily workload for different types of patients: the impact on incision time is more significant for elective patients, while the surgery outcomes of non-elective patients are more affected by surgeon's daily workload.

We are currently exploring following directions in our study. First, we are implementing the surgery scheduling model proposed in Section 4.5 on our sample. We aim to quantify

the benefit of incorporating the effect of surgeon's daily workload in surgery scheduling, i.e., the reduction in total expected incision time. We also plan to investigate other types of objectives, such as total expected overtime or patient's post-LOS. Besides, we are conducting more robustness checks of our empirical results, as well as communicating with doctors to discuss other potential mechanisms for the effects of surgeon's workload. Finally, as a potential future research, we hope to obtain the surgery data in the period of the COVID-19 pandemic. During the first several months of the pandemic, the hospital cancelled most of its elective surgeries. This provides an exogenous shock to the surgeon's workload and patient's waiting time. If we were able to get this data (i.e., extending our sample by one year), we may be able to find other interesting results related to the effects of surgeon's workload and delays in patient treatment.

Bibliography

Abbring, Jaap H, Øystein Daljord. 2019. Identifying the discount factor in dynamic discrete choice models. *Becker Friedman Institute for Research in Economics Working Paper* (2017-17).

Admati, Anat R. 1985. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica: Journal of the Econometric Society* 629–657.

Aït-Sahalia, Yacine, Jianqing Fan, Dacheng Xiu. 2010. High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association* **105**(492) 1504–1517.

Ait-Sahalia, Yacine, Per A Mykland, Lan Zhang. 2005. How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* **18**(2) 351–416.

Aït-Sahalia, Yacine, Per A Mykland, Lan Zhang. 2011. Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics* **160**(1) 160–175.

Akşin, Zeynep, Barış Ata, Seyed Morteza Emadi, Che-Lin Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Science* **59**(12) 2727–2746.

Akşin, Zeynep, Baris Ata, Seyed Morteza Emadi, Che-Lin Su. 2016. Impact of delay announcements in call centers: An empirical approach. *Operations Research* **65**(1) 242–265.

Alfonsi, Aurélien, Alexander Schied, Alla Slynko. 2012. Order book resilience, price manipulation, and the positive portfolio problem. *SIAM Journal on Financial Mathematics* **3**(1) 511–533.

Allen, Franklin, Stephen Morris, Hyun Song Shin. 2006. Beauty contests and iterated expectations in asset markets. *The Review of Financial Studies* **19**(3) 719–752.

Allon, Gad, Sarang Deo, Wuqin Lin. 2013. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61**(3) 544–562.

Andersen, Torben G, Tim Bollerslev. 1997. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* **4**(2-3) 115–158.

Andersen, Torben G, Tim Bollerslev, Ashish Das. 2001. Variance-ratio statistics and high-frequency data: Testing for changes in intraday volatility patterns. *Journal of Finance* **56**(1) 305–327.

Andersen, Torben G, Tim Bollerslev, Francis X Diebold, Paul Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* **71**(2) 579–625.

Ang, Erjie, Sara Kwasnick, Mohsen Bayati, Erica L Plambeck, Michael Aratow. 2016. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* **18**(1) 141–156.

Angrist, Joshua D, Alan B Krueger. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives* **15**(4) 69–85.

Appel, Ian R, Todd A Gormley, Donald B Keim. 2016. Passive investors, not passive owners. *Journal of Financial Economics* **121**(1) 111–141.

Avdis, Efstathios. 2016. Information tradeoffs in dynamic financial markets. *Journal of Financial Economics* **122**(3) 568–584.

Bagrodia, Aditya, Varun Rachakonda, Karen Delafuente, Suzette Toombs, Owen Yeh, Joseph Scales, Claus G Roehrborn, Yair Lotan. 2012. Surgeon fatigue: impact of case order on perioperative parameters and patient outcomes. *The Journal of Urology* **188**(4) 1291–1296.

Bajari, Patrick, C Lanier Benkard, Jonathan Levin. 2007. Estimating dynamic models of imperfect competition. *Econometrica* **75**(5) 1331–1370.

Barndorff-Nielsen, Ole E, Peter Reinhard Hansen, Asger Lunde, Neil Shephard. 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* **76**(6) 1481–1536.

Barndorff-Nielsen, Ole E, Peter Reinhard Hansen, Asger Lunde, Neil Shephard. 2011. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* **162**(2) 149–169.

Barndorff-Nielsen, Ole E, Neil Shephard. 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(2) 253–280.

Barro, Robert J. 2009. Rare disasters, asset prices, and welfare costs. *American Economic Review* **99**(1) 243–64.

Bashir, Mohamad, Amer Harky, Matthew Fok, Matthew Shaw, Graeme L Hickey, Stuart W Grant, Rakesh Uppal, Aung Oo. 2017. Acute type a aortic dissection in the united kingdom: surgeon volume-outcome relation. *The Journal of Thoracic and Cardiovascular Surgery* **154**(2) 398–406.

Batt, Robert J, Diwas S Kc, Bradley R Staats, Brian W Patterson. 2019. The effects of discrete work shifts on a nonterminating service system. *Production and Operations Management* **28**(6) 1528–1544.

Batt, Robert J, Christian Terwiesch. 2016. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* **63**(11) 3531–3551.

Batun, Sakine, Brian T Denton, Todd R Huschka, Andrew J Schaefer. 2011. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing* **23**(2) 220–237.

Berry Jaeker, Jillian A, Anita L Tucker. 2017. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* **63**(4) 1042–1062.

Biais, Bruno, Peter Bossaerts, Chester Spatt. 2010. Equilibrium asset pricing and portfolio choice under asymmetric information. *The Review of Financial Studies* **23**(4) 1503–1543.

Bibinger, Markus, Nikolaus Hautsch, Peter Malec, Markus Reiss. 2019. Estimating the spot covariation of asset prices – statistical theory and empirical evidence. *Journal of Business and Economic Statistics* **37**(3) 419–435.

Bogousslavsky, Vincent. 2016. Infrequent rebalancing, return autocorrelation, and seasonality. *Journal of Finance* **71**(6) 2967–3006.

Bollerslev, Tim, Nour Meddahi, Serge Nyawa. 2019. High-dimensional multivariate realized volatility estimation. *Journal of Econometrics* **212**(1) 116–136.

Boudt, Kris, Jin Zhang. 2015. Jump robust two time scale covariance estimation and realized volatility budgets. *Quantitative Finance* **15**(6) 1041–1054.

Brancati, Emanuele, Marco Macchiavelli. 2019. The information sensitivity of debt in good and bad times. *Journal of Financial Economics* **133**(1) 99–112.

Bray, Robert L, Yuliang Yao, Yongrui Duan, Jiazhen Huo. 2019. Ration gaming and the bullwhip effect. *Operations Research* **67**(2) 453–467.

Brennan, Michael J, H Henry Cao. 1997. International portfolio investment flows. *The Journal of Finance* **52**(5) 1851–1880.

Buccheri, Giuseppe, Giacomo Bormetti, Fulvio Corsi, Fabrizio Lillo. 2020. A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics. *Journal of Business and Economic Statistics* 1–17.

Cameron, A Colin, Pravin K Trivedi. 2013. *Regression analysis of count data*, vol. 53. Cambridge university press.

Campbell, John Y. 1990. A variance decomposition for stock returns. Tech. rep., National Bureau of Economic Research.

Campbell, John Y, Tarun Ramadorai, Allie Schwartz. 2009. Caught on tape: Institutional trading, stock returns, and earnings announcements. *Journal of Financial Economics* **92**(1) 66–91.

Cardoen, Brecht, Erik Demeulemeester, Jeroen Beliën. 2010. Operating room planning and scheduling: A literature review. *European Journal of Operational Research* **201**(3) 921–932.

CESA-BIANCHI, AMBROGIO, EMILIO FERNANDEZ-CORUGEDO. 2018. Uncertainty, financial frictions, and nominal rigidities: a quantitative investigation. *Journal of Money, Credit and Banking* **50**(4) 603–636.

Chalfin, Donald B, Stephen Trzeciak, Antonios Likourezos, Brigitte M Baumann, R Phillip Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical care medicine* **35**(6) 1477–1483.

Chamley, Christophe Paul. 2007. Complementarities in information acquisition with short-term trades. *Theoretical Economics* **2**(4) 441–467.

Chan, Carri W, Vivek F Farias, Nicholas Bambos, Gabriel J Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations research* **60**(6) 1323–1341.

Chan, Carri W, Vivek F Farias, Gabriel J Escobar. 2016. The impact of delays on service times in the intensive care unit. *Management Science* **63**(7) 2049–2072.

Chen, Lena M, Marta Render, Anne Sales, Edward H Kennedy, Wyndy Wiitala, Timothy P Hofer. 2012. Intensive care unit admitting patterns in the veterans affairs health care system. *Archives of Internal Medicine* **172**(16) 1220–1226.

Ching, Andrew T, Matthew Osborne. 2017. Identification and estimation of forward-looking behavior: The case of consumer stockpiling. *Rotman School of Management Working Paper* (2594032).

Ching, Andrew T, Matthew Osborne. 2020. Identification and estimation of forward-looking behavior: The case of consumer stockpiling. *Marketing Science* .

Clark, Jonathan R, Robert S Huckman. 2012. Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Science* **58**(4) 708–722.

Coopersmith, Craig M, Hannah Wunsch, Mitchell P Fink, Walter T Linde-Zwirble, Keith M Olsen, Marilyn S Sommers, Kanwaljeet JS Anand, Kathryn M Tchorz, Derek C Angus, Clifford S Deutschman. 2012. A comparison of critical care research funding and the financial burden of critical illness in the united states. *Critical Care Medicine* **40**(4) 1072–1079.

Corallo, Ashley N, Ruth Croxford, David C Goodman, Elisabeth L Bryan, Divya Srivastava, Therese A Stukel. 2014. A systematic review of medical practice variation in oecd countries. *Health Policy* **114**(1) 5–14.

Cushing, David, Ananth Madhavan. 2000. Stock returns and trading at the close. *Journal of Financial Markets* **3**(1) 45–67.

Dallery, Yves, Stanley B Gershwin. 1992. Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems* **12**(1) 3–94.

Dang, Tri Vi, Gary Gorton, Bengt Holmström. 2012. Ignorance, debt and financial crises. *Yale University and Massachusetts Institute of technology, working paper* **17**.

Dang, Tri Vi, Gary Gorton, Bengt Holmström. 2020. The information view of financial crises. *Annual Review of Financial Economics* **12** 39–65.

De Groote, Olivier, Frank Verboven. 2019. Subsidies and time discounting in new technology adoption: Evidence from solar photovoltaic systems. *American Economic Review* **109**(6) 2137–72.

Delasay, Mohammad, Armann Ingolfsson, Bora Kolfal. 2016. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research* **64**(4) 867–885.

Delasay, Mohammad, Armann Ingolfsson, Bora Kolfal, Kenneth Schultz. 2019. Load effect on service times. *European Journal of Operational Research* **279**(3) 673–686.

Denton, Brian T, Andrew J Miller, Hari J Balasubramanian, Todd R Huschka. 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research* **58**(4) 802–816.

Dietz, Dennis C. 2011. Practical scheduling for call center operations. *Omega* **39**(5) 550–557.

Doebbeling, Bradley N, Matthew M Burton, Eric A Wiebke, Spencer Miller, Laurence Baxter, Donald Miller, Jorge Alvarez, Joseph Pekny. 2012. Optimizing perioperative decision making: improved information for clinical workflow planning. *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 154.

Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2018. The impact of delay announcements on hospital network coordination and waiting times. *Management Science* **65**(5) 1969–1994.

Dorion, Dominique, Simon Darveau. 2013. Do micropauses prevent surgeon's fatigue and loss of accuracy associated with prolonged surgery? an experimental prospective study. *Journal of Vascular Surgery* **57**(4) 1173.

Dow, James, Itay Goldstein, Alexander Guembel. 2017. Incentives for information production in markets where prices affect real investment. *Journal of the European Economic Association* **15**(4) 877–909.

Driebusch, Corrie, Alexander Osipovich, Gregory Zuckerman. 2018. What's the biggest trade on the New York stock exchange? The last one. `https://www.wsj.com/articles/at-closing-time-the-stock-market-heats-up-like-a-bar-at-last-call-1521038300`.

Dubé, Jean-Pierre, Günter J Hitsch, Pranav Jindal. 2014. The joint identification of utility and discount functions from stated choice data: An application to durable goods adoption. *Quantitative Marketing and Economics* **12**(4) 331–377.

Dutta, Sunil, Alexander Nezlobin. 2017. Information disclosure, firm growth, and the cost of capital. *Journal of Financial Economics* **123**(2) 415–431.

Edbrooke, David L, Cosetta Minelli, Gary H Mills, Gaetano Iapichino, Angelo Pezzi, Davide Corbella, Philip Jacobs, Anne Lippert, Joergen Wiis, Antonio Pesenti, et al. 2011. Implications of icu triage decisions on patient mortality: a cost-effectiveness analysis. *Critical Care* **15**(1) R56.

Ellman, Peter I, Irving L Kron, Jeffrey S Alvis, Carlos Tache-Leon, Thomas S Maxey, T Brett Reece, Benjamin B Peeler, John A Kern, Curtis G Tribble. 2005. Acute sleep deprivation in the thoracic surgical resident does not affect operative outcomes. *The Annals of Thoracic Surgery* **80**(1) 60–65.

Emadi, Seyed Morteza, Bradley R Staats. 2019. A structural estimation approach to agent attrition. *Management Science, to appear* .

Epps, Thomas W. 1979. Comovements in stock prices in the very short run. *Journal of the American Statistical Association* **74**(366a) 291–298.

Erdogan, S Ayca, Brian T Denton. 2010. Surgery planning and scheduling. *Wiley encyclopedia of operations research and management science* .

Escobar, Gabriel J, Marla N Gardner, John D Greene, David Draper, Patricia Kipnis. 2013. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Medical Care* 446–453.

Escobar, Gabriel J, Juan Carlos LaGuardia, Benjamin J Turk, Arona Ragins, Patricia Kipnis, David Draper. 2012. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine* **7**(5) 388–395.

Falcoz, Pierre-Emmanuel, Marc Puyraveau, Caroline Rivera, Alain Bernard, Gilbert Massard, Frederic Mauny, Marcel Dahan, Pascal-Alexandre Thomas, Epithor Group. 2014. The impact of hospital and surgeon volume on the 30-day mortality of lung cancer surgery: a nation-based reappraisal. *The Journal of Thoracic and Cardiovascular Surgery* **148**(3) 841–848.

Farboodi, Maryam, Laura Veldkamp. 2020. Long-run growth of financial data technology. *American Economic Review* **110**(8) 2485–2523.

Fisher, Elliott S, David E Wennberg, Thérèse A Stukel, Daniel J Gottlieb. 2004. Variations in the longitudinal efficiency of academic medical centers: Increased intensity of care does not appear to be associated with higher quality or to result in better survival at amcs. *Health Affairs* **23**(Suppl2) VAR–19.

Foucault, Thierry, Ohad Kadan, Eugene Kandel. 2005. Limit order book as a market for liquidity. *Review of Financial Studies* **18**(4) 1171–1217.

Franzini, Luisa, Kavita R Sail, Eric J Thomas, Laura Wueste. 2011. Costs and cost-effectiveness of a telemedicine intensive care unit program in 6 intensive care units in a large health care system. *Journal of Critical Care* **26**(3) 329–e1.

Frederick, Shane, George Loewenstein, Ted O'donoghue. 2002. Time discounting and time preference: A critical review. *Journal of economic literature* **40**(2) 351–401.

Freeman, Michael, Nicos Savva, Stefan Scholtes. 2017. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* **63**(10) 3147–3167.

Freeman, Nickolas K, Sharif H Melouk, John Mittenthal. 2016. A scenario-based approach for operating theater scheduling under uncertainty. *Manufacturing & Service Operations Management* **18**(2) 245–261.

Gabler, Nicole B, Sarah J Ratcliffe, Jason Wagner, David A Asch, Gordon D Rubenfeld, Derek C Angus, Scott D Halpern. 2013. Mortality among patients admitted to strained intensive care units. *American Journal of Respiratory and Critical Care Medicine* **188**(7) 800–806.

Ganguli, Jayant Vivek, Liyan Yang. 2009. Complementarities, multiplicity, and supply information. *Journal of the European Economic Association* **7**(1) 90–115.

Gao, Lei, Yufeng Han, Sophia Zhengzi Li, Guofu Zhou. 2018. Market intraday momentum. *Journal of Financial Economics* **129**(2) 394–414.

Gattinoni, Luciano, Danilo Radrizzani, Bruno Simini, Guido Bertolini, Luca Ferla, Giovanni Mistraletti, Francesca Porta, Dinis R Miranda, et al. 2004. Volume of activity and occupancy rate in intensive care units. association with mortality. *Intensive Care Medicine* **30**(2) 290–297.

Gençay, Ramazan, Giuseppe Ballocchi, Michel Dacorogna, Richard Olsen, Olivier Pictet. 2002. Real-time trading models and the statistical properties of foreign exchange rates. *International Economic Review* 463–491.

Gençay, Ramazan, Michel Dacorogna, Ulrich A Muller, Olivier Pictet, Richard Olsen. 2001. *An introduction to high-frequency finance*. Elsevier.

George, Jennifer M, J Michael Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Operations Research* **49**(5) 720–731.

Gerdes, Jodi, Kanav Kahol, Marshall Smith, Mario J Leyba, John J Ferrara. 2008. Jack barney award: the effect of fatigue on cognitive and psychomotor skills of trauma residents and attending surgeons. *The American Journal of Surgery* **196**(6) 813–820.

Glasserman, Paul, Fulin Li, Harry Mamaysky. 2019. Time variation in the news-returns relationship. *Available at SSRN 3420981* .

Goldstein, Itay, Yaron Leitner. 2018. Stress tests and information disclosure. *Journal of Economic Theory* **177** 34–69.

Goldstein, Itay, Liyan Yang. 2015. Information diversity and complementarities in trading and information acquisition. *The Journal of Finance* **70**(4) 1723–1765.

Gorton, Gary, Guillermo Ordonez. 2014. Collateral crises. *American Economic Review* **104**(2) 343–78.

Govindarajan, Anand, David R Urbach, Matthew Kumar, Qi Li, Brian J Murray, David Juurlink, Erin Kennedy, Anna Gagliardi, Rinku Sutradhar, Nancy N Baxter. 2015. Outcomes of daytime procedures performed by attending surgeons after night work. *New England Journal of Medicine* **373**(9) 845–853.

Green, Linda V, Vien Nguyen. 2001. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* **36**(2) 421.

Green, Linda V, Sergei Savin, Nicos Savva. 2013. "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10) 2237–2256.

Grossman, Sanford J, Joseph E Stiglitz. 1980. On the impossibility of informationally efficient markets. *The American economic review* **70**(3) 393–408.

Grynkiv, Iaryna, Kimberly Russell. 2015. A look inside the shifting volume smile for us equities. *Journal of Trading* **11**(1) 26–37.

Gul, Serhat, Brian T Denton, John W Fowler, Todd Huschka. 2011. Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management* **20**(3) 406–417.

Gupta, Diwakar. 2007. Surgical suites' operations management. *Production and Operations Management* **16**(6) 689–700.

Haas, Nathan L, Sage P Whitmore, James A Cranford, Ryan E Tsuchida, Adam Nicholson, Caryn Boyd, Kyle J Gunnerson, Roma Y Gianchandani, Benjamin S Bassin. 2020. An emergency department–based intensive care unit is associated with decreased hospital and intensive care unit utilization for diabetic ketoacidosis. *The Journal of emergency medicine* **58**(4) 620–626.

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

Halldorson, Jeffrey B, Ramasamy Bakthavatsalam, Jorge D Reyes, James D Perkins. 2009. The impact of consecutive operations on survival after liver transplantation. *Liver Transplantation* **15**(8) 907–914.

Halpern, Neil A, Stephen M Pastores. 2010. Critical care medicine in the united states 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical Care Medicine* **38**(1) 65–71.

Halpern, Neil A, Stephen M Pastores. 2015. Critical care medicine beds, use, occupancy and costs in the united states: a methodological review. *Critical Care Medicine* **43**(11) 2452.

Halpern, Scott D. 2011. Icu capacity strain and the quality and allocation of critical care. *Current opinion in critical care* **17**(6) 648–657.

Hansen, Peter Reinhard, Zhuo Huang, Howard Howan Shek. 2012. Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics* **27**(6) 877–906.

Hellwig, Martin F. 1980. On the aggregation of information in competitive markets. *Journal of economic theory* **22**(3) 477–498.

Heston, Steven L, Robert A Korajczyk, Ronnie Sadka. 2010. Intraday patterns in the cross-section of stock returns. *Journal of Finance* **65**(4) 1369–1407.

Hopp, Wallace J, Seyed MR Iravani, Gigi Y Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77.

Hotz, V Joseph, Robert A Miller. 1993. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies* **60**(3) 497–529.

Huang, David T. 2004. Clinical review: impact of emergency department care on intensive care unit costs. *Critical Care* **8**(6) 498.

Huberman, Gur, Werner Stanzl. 2004. Price manipulation and quasi-arbitrage. *Econometrica* **72**(4) 1247–1275.

Hugonnet, Stéphane, Jean-Claude Chevrolet, Didier Pittet. 2007. The effect of workload on infection risk in critically ill patients. *Critical care medicine* **35**(1) 76–81.

Ibanez, Maria R, Jonathan R Clark, Robert S Huckman, Bradley R Staats. 2017. Discretionary task ordering: Queue management in radiological services. *Management Science* **64**(9) 4389–4407.

Jacod, Jean, Yingying Li, Per A Mykland, Mark Podolskij, Mathias Vetter. 2009. Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and Their Applications* **119**(7) 2249–2276.

Janhofer, David E, Chrisovalantis Lakhiani, David H Song. 2019. Addressing surgeon fatigue: current understanding and strategies for mitigation. *Plastic and Reconstructive Surgery* **144**(4) 693e–699e.

Kahn, Jeremy M, Gordon D Rubenfeld, Jeffery Rohrbach, Barry D Fuchs. 2008. Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients. *Medical Care* 1226–1233.

Kappou, Konstantina, Chris Brooks, Charles Ward. 2010. The s&p 500 index effect reconsidered: Evidence from overnight and intraday stock price performance and volume. *Journal of Banking and Finance* **34**(1) 116–126.

Karolyi, G Andrew, Kuan-Hui Lee, Mathijs A Van Dijk. 2012. Understanding commonality in liquidity around the world. *Journal of Financial Economics* **105**(1) 82–112.

Kc, Diwas S, Christian Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

Kc, Diwas Singh. 2014. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.

Kc, Diwas Singh, Bradley R Staats. 2012. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* **14**(4) 618–633.

Kc, Diwas Singh, Christian Terwiesch. 2011. The effects of focus on performance: Evidence from california hospitals. *Management Science* **57**(11) 1897–1912.

Kc, Diwas Singh, Christian Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65.

Keskinocak, Pinar, Nicos Savva. 2020. A review of the healthcare-management (modeling) literature published in manufacturing & service operations management. *Manufacturing & Service Operations Management* **22**(1) 59–72.

Kim, Song-Hee, Carri W Chan, Marcelo Olivares, Gabriel Escobar. 2015. Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* **61**(1) 19–38.

Kim, Song-Hee, Jordan Tong, Carol Peden. 2019. Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Available at SSRN 3219451* .

Koch, Andrew, Stefan Ruenzi, Laura Starks. 2016. Commonality in liquidity: a demand-side explanation. *Review of Financial Studies* **29**(8) 1943–1974.

Komarova, Tatiana, Fabio Sanches, Daniel Silva Junior, Sorawoot Srisuma. 2018. Joint analysis of the discount factor and payoff parameters in dynamic discrete choice models. *Quantitative Economics* **9**(3) 1153–1194.

Kreps, David M, Evan L Porteus. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica: journal of the Econometric Society* 185–200.

Kuntz, Ludwig, Roman Mennicken, Stefan Scholtes. 2015. Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* **61**(4) 754–771.

Kyle, Albert S. 1985. Continuous auctions and insider trading. *Econometrica* 1315–1335.

Li, Jun, Nelson Granados, Serguei Netessine. 2014. Are consumers strategic? structural estimation from the air-travel industry. *Management Science* **60**(9) 2114–2137.

Li, Xiangyong, N Rafaliya, M Fazle Baki, Ben A Chaouch. 2017. Scheduling elective surgeries: the tradeoff among bed capacity, waiting patients and operating room utilization using goal programming. *Healthcare Management Science* **20**(1) 33–54.

Liu, Fei, Athanasios A Pantelous, Hans-Jörg von Mettenheim. 2018. Forecasting and trading high frequency volatility on large indices. *Quantitative Finance* **18**(5) 737–748.

Lu, Yina, Andrés Musalem, Marcelo Olivares, Ariel Schilkrut. 2013. Measuring the effect of queues on customer purchases. *Management Science* **59**(8) 1743–1763.

Magnac, Thierry, David Thesmar. 2002. Identifying dynamic discrete decision processes. *Econometrica* **70**(2) 801–816.

Mamaysky, Harry. 2020. Financial markets and news about the coronavirus. *Available at SSRN 3565597* .

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* **60**(3) 531–542.

Manzano, Carolina, Xavier Vives. 2011. Public and private learning from prices, strategic substitutability and complementarity, and equilibrium multiplicity. *Journal of Mathematical Economics* **47**(3) 346–369.

May, Jerrold H, William E Spangler, David P Strum, Luis G Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management* **20**(3) 392–405.

McCormick, Frank, John Kadzielski, Christopher P Landrigan, Brady Evans, James H Herndon, Harry E Rubash. 2012. Surgeon fatigue: a prospective analysis of the incidence, risk, and intervals of predicted fatigue-related impairment in residents. *Archives of Surgery* **147**(5) 430–435.

McDermott, Kimberly W, William J Freeman, Anne Elixhauser. 2017. Overview of operating room procedures during inpatient stays in us hospitals, 2014: statistical brief# 233. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville: Agency for Healthcare Research and Quality* .

Mehta, Nitin, Jian Ni, Kannan Srinivasan, Baohong Sun. 2017. A dynamic model of health insurance choices and healthcare consumption decisions. *Marketing Science* **36**(3) 338–360.

Miessi Sanches, Fabio A, Daniel Junior Silva, Sorawoot Srisuma. 2016. Ordinary least squares estimation of a dynamic game model. *International Economic Review* **57**(2) 623–634.

Min, Daiki, Yuehwern Yih. 2010. Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research* **206**(3) 642–652.

Min, Seungki, Costis Maglaras, Ciamac C Moallemi. 2018. Cross-sectional variation of intraday liquidity, cross-impact, and their effect on portfolio execution. *Columbia Business School Research Paper* (19-4).

Modrall, J Gregory, Rebecca M Minter, Abu Minhajuddin, Javier Eslava-Schmalbach, Girish P Joshi, Shivani Patel, Eric B Rosero. 2018. The surgeon volume-outcome relationship: not yet ready for policy. *Annals of Surgery* **267**(5) 863–867.

Mondria, Jordi. 2010. Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory* **145**(5) 1837–1864.

Mullan, Fitzhugh. 2004. Wrestling with variation: An interview with jack wennberg: The creator of modern-day evaluative clinical sciences discusses what motivated him to define and pursue this area of study. *Health Affairs* **23**(Suppl2) VAR–73.

Needleman, Jack, Peter Buerhaus, V Shane Pankratz, Cynthia L Leibson, Susanna R Stevens, Marcelline Harris. 2011. Nurse staffing and inpatient hospital mortality. *New England Journal of Medicine* **364**(11) 1037–1045.

Neuraz, Antoine, Claude Guérin, Cécile Payet, Stéphanie Polazzi, Frédéric Aubrun, Frédéric Dailler, Jean-Jacques Lehot, Vincent Piriou, Jean Neidecker, Thomas Rimmelé, et al. 2015. Patient mortality is associated with staff resources and workload in the icu: a multicenter observational study. *Critical Care Medicine* **43**(8) 1587–1594.

O'Connor, Annette M, Hilary A Llewellyn-Thomas, Ann Barry Flood. 2004. Modifying unwarranted variations in health care: Shared decision making using patient decision aids: A review of the evidence base for shared decision making. *Health Affairs* **23**(Suppl2) VAR–63.

Oliva, Rogelio, John D Sterman. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* **47**(7) 894–914.

Olivares, Marcelo, Christian Terwiesch, Lydia Cassorla. 2008. Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science* **54**(1) 41–55.

Pagano, Michael S, Lin Peng, Robert A Schwartz. 2008. The quality of price formation at market openings and closings: Evidence from the nasdaq stock market. Tech. rep., CFS Working Paper.

Peress, Joel. 2010. The tradeoff between risk sharing and information production in financial markets. *Journal of Economic Theory* **145**(1) 124–155.

Pesendorfer, Martin, Philipp Schmidt-Dengler. 2008. Asymptotic least squares estimators for dynamic games. *The Review of Economic Studies* **75**(3) 901–928.

Rath, Sandeep, Kumar Rajaram. 2018. Staff planning for hospitals with cost estimation and optimization. *Kenan Institute of Private Enterprise Research Paper* (18-28).

Roberts, Rebecca R, Paul W Frutos, Ginevra G Ciavarella, Leon M Gussow, Edward K Mensah, Linda M Kampe, Helen E Straus, Gnanaraj Joseph, Robert J Rydman. 1999. Distribution of variable vs fixed costs of hospital care. *Jama* **281**(7) 644–649.

Rust, John. 1987. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society* 999–1033.

Rust, John. 1994. Structural estimation of markov decision processes. *Handbook of Econometrics* **4** 3081–3143.

Samudra, Michael, Carla Van Riet, Erik Demeulemeester, Brecht Cardoen, Nancy Vansteenkiste, Frank E Rademakers. 2016. Scheduling operating rooms: achievements, challenges and pitfalls. *Journal of Scheduling* **19**(5) 493–525.

Schilling, Peter L, Darrell A Campbell, Michael J Englesbe, Matthew M Davis. 2010. A comparison of in-hospital mortality risk conferred by high hospital occupancy, differences in nurse staffing levels, weekend admission, and seasonal influenza. *Medical Care* 224–232.

Schultz, Kenneth L, David C Juran, John W Boudreau, John O McClain, L Joseph Thomas. 1998. Modeling and worker motivation in jit production systems. *Management Science* **44**(12-part-1) 1595–1607.

Shanafelt, Tait D, Charles M Balch, Gerald Bechamps, Tom Russell, Lotte Dyrbye, Daniel Satele, Paul Collicott, Paul J Novotny, Jeff Sloan, Julie Freischlag. 2010. Burnout and medical errors among american surgeons. *Annals of Surgery* **251**(6) 995–1000.

Shmueli, Amir, Charles L Sprung, Edward H Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6**(3) 131–136.

Shylo, Oleg V, Oleg A Prokopyev, Andrew J Schaefer. 2013. Stochastic operating room scheduling for high-volume specialties under block booking. *INFORMS Journal on Computing* **25**(4) 682–692.

Singer, Adam J, Henry C Thode Jr, Peter Viccellio, Jesse M Pines. 2011. The association between length of emergency department boarding and mortality. *Academic Emergency Medicine* **18**(12) 1324–1329.

Slack, PS, CJ Coulson, X Ma, K Webster, DW Proops. 2008. The effect of operating time on surgeons' muscular fatigue. *The Annals of The Royal College of Surgeons of England* **90**(8) 651–657.

Song, Hummy, Anita Tucker, Ryan Graue, Sarah Moravick, Julius Yang. 2019. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science, to appear* .

Song, Hummy, Anita L Tucker, Karen L Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

Spiegel, Matthew. 1998. Stock price volatility in a multiple security overlapping generations model. *The Review of Financial Studies* **11**(2) 419–447.

Staats, Bradley R, Francesca Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management science* **58**(6) 1141–1159.

Stewart, Adam F., Donna M. Ferriero, S. Andrew Josephson, Daniel H. Lowenstein, Robert O. Messing, Jorge R. Oksenberg, S. Claiborne Johnston, Stephen L. Hauser. 2012. Fighting decision fatigue. *Annals of Neurology* **71**(1) A5–A15.

Stroud, Jonathan R, Michael S Johannes. 2014. Bayesian modeling and forecasting of 24-hour high-frequency volatility. *Journal of the American Statistical Association* **109**(508) 1368–1384.

Strumpf, Dan. 2015. Stock-market traders pile in at the close. `https://www.wsj.com/articles/traders-pile-in-at-the-close-1432768080`.

Tan, Tom Fangyun, Serguei Netessine. 2014. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.

Thomas, Mathew, Mark S Allen, Dennis A Wigle, K Robert Shen, Stephen D Cassivi, Francis C Nichols III, Claude Deschamps. 2012. Does surgeon workload per day affect outcomes after pulmonary lobectomies? *The Annals of Thoracic Surgery* **94**(3) 966–972.

Van Nieuwerburgh, Stijn, Laura Veldkamp. 2010. Information acquisition and under-diversification. *The Review of Economic Studies* **77**(2) 779–805.

Veldkamp, Laura L. 2006. Media frenzies in markets for financial information. *American Economic Review* **96**(3) 577–601.

Vuolteenaho, Tuomo. 2002. What drives firm-level stock returns? *The Journal of Finance* **57**(1) 233–264.

Wang, Guihua, Behrooz Pourghannad. 2020. Matching patients with surgeons: Heterogeneous effects of surgical volume on surgery duration. *Available at SSRN* .

Wang, Jiang. 1993. A model of intertemporal asset prices under asymmetric information. *The Review of Economic Studies* **60**(2) 249–282.

Wang, Jiang. 1994. A model of competitive stock trading volume. *Journal of political Economy* **102**(1) 127–168.

Wang, Yang. 2014. Dynamic implications of subjective expectations: Evidence from adult smokers. *American Economic Journal: Applied Economics* **6**(1) 1–37.

Watanabe, Masahiro. 2008. Price volatility and investor behavior in an overlapping generations model with information asymmetry. *The Journal of Finance* **63**(1) 229–272.

Weingart, Scott D, Robert L Sherwin, Lillian L Emlet, Isaac Tawil, Julie Mayglothling, Jon C Rittenberger. 2013. Ed intensivists and ed intensive care units. *The American journal of emergency medicine* **31**(3) 617–620.

Westert, Gert P, Stef Groenewoud, John E Wennberg, Catherine Gerard, Phil DaSilva, Femke Atsma, David C Goodman. 2018. Medical practice variation: public reporting a first necessary step to spark change. *International Journal for Quality in Health Care* **30**(9) 731–735.

Wolff, Ronald W. 1989. *Stochastic modeling and the theory of queues*. Pearson College Division.

Wood, Robert A, Thomas H McInish, J Keith Ord. 1985. An investigation of transactions data for nyse stocks. *Journal of Finance* **40**(3) 723–739.

Woodridge, JM. 2010. *Econometric Analysis of Cross Section and Panel Data, 2nd Edtion*. The MIT Press.

Wu, Yanbin. 2019. Closing auction, passive investing, and stock prices. *Available at SSRN 3440239* .

Xu, Kuang, Carri W Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* **18**(3) 314–331.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2016. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63**(1) 1–20.

Zenteno, Ana C, Tim Carnes, Retsef Levi, Bethany J Daily, Devon Price, Susan C Moss, Peter F Dunn. 2015. Pooled open blocks shorten wait times for nonelective surgical cases. *Annals of Surgery* **262**(1) 60–67.

Zenteno, Ana Cecilia, Tim Carnes, Retsef Levi, Bethany J Daily, Peter F Dunn. 2016. Systematic or block allocation at a large academic medical center. *Annals of Surgery* **264**(6) 973–981.

Zhang, Lan. 2011. Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics* **160**(1) 33–47.

Zhang, Lan, Per A Mykland, Yacine Aït-Sahalia. 2005. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* **100**(472) 1394–1411.

Appendix A

Chapter 1 Supplemental Information

A.1 Information Production

We decompose the time $t+1$ dividend M_{t+1} into a total of $\bar{N}$ units of information η_i, so that $M_{t+1} = \eta_1 + \cdots + \eta_{\bar{N}}$. These units are i.i.d. normally distributed with variance $\mathsf{var}(M)/\bar{N}$. To capture the idea that each piece of information is small, we assume $\bar{N}$ is large. Let f_t be the fraction of the $\bar{N}$ units that are observable at time t, and $1 - f_t$ be the fraction of unobservable units. Then the components of M_{t+1} can be written as:

$$\tilde{m}_t = \sum_{i \in \text{observable}} \eta_i, \qquad \epsilon_{t+1} = \sum_{j \in \text{unobservable}} \eta_j.$$

And we'll have $\mathsf{var}(\tilde{m}_t) = f_t \mathsf{var}(M)$ and $\mathsf{var}(\epsilon_{t+1}) = (1 - f_t)\mathsf{var}(M)$.

We assume that in every period: (a) with probability $\pi^{o:u}$ any previously observable piece of information may become unobservable next period; (b) with probability $\pi^{u:o}$ any previously unobservable piece of information becomes observable next period; and (c) a certain number $\epsilon_{f,t+1}$ of information units transitions at random from unobservable at time t to observable at time $t+1$, or vice versa. It is possible for a given t that $\epsilon_{f,t} = 0$. This shock proxies for large aggregate changes in observability, as opposed to the micro changes that are captured by $\pi^{u:o}$ and $\pi^{o:u}$. Note that $\epsilon_{f,t+1}$ should be specified to ensure that $f_{t+1} \in [0, 1]$.

When $b_f = 0$, we can derive the f_t dynamics in (1.5) and (1.6) in terms of these two transition probabilities. Given f_t, and in the absence of any information production, next period's fraction of information units that are knowable will be

$$f_{t+1} \approx \underbrace{(1 - f_t) \times \pi^{u:o}}_{\text{unobservable} \to \text{observable}} + \underbrace{f_t \times (1 - \pi^{o:u})}_{\text{observable} \to \text{observable}} + \epsilon_{f,t+1}.$$

The approximation becomes exact for large $\bar{N}$, where the fraction of units that change states approaches the probabilities $\pi^{u:o}$ and $\pi^{o:u}$. Matching this expression to the corresponding parts in our f_t dynamics in (1.5) requires

$$(1 - f_t) \times \pi^{u:o} + f_t \times (1 - \pi^{o:u}) = a_f + \kappa_f \times (f_t - a_f). \tag{A.1}$$

Equating the coefficient of f_t on the two sides yields

$$\pi^{u:o} = (1 - \kappa_f) \times a_f, \quad \pi^{o:u} = (1 - \kappa_f) \times (1 - a_f),$$

which we solve to get

$$\kappa_f = 1 - \pi^{u:o} - \pi^{o:u}, \quad a_f = \frac{\pi^{u:o}}{\pi^{u:o} + \pi^{o:u}}.$$

The persistence parameter κ_f is therefore greater when the transition probabilities for individual units are smaller. The ratio defining a_f is the stationary probability that an individual unit is observable, so (A.1) says that f_t mean-reverts to the average fraction of observable units.

Given the model calibration from Table A.1, the relationships above imply

$$\pi^{u:o} = 0.01575 \qquad \pi^{o:u} = 0.07425.$$

It is about five times more likely that a currently observable piece of information becomes unobservable, than a currently unobservable piece of information becomes observable. This result is a direct outcome of the steady state level of information in the absence of feedback, i.e. a_f, being low and is an inevitable feature of a two-regime model (since the no-feedback regime has to be at a low value of f_t).

Information Production Sector

Each news outlet j can discover I_j units of information, i.e. η_i's from above, at a fixed cost per unit of c_P. We assume each outlet discovers a unique set of information. Once a unit of information is discovered, the marginal cost of revealing it to investors is zero, and that unit becomes a generic observable unit with a $1 - \pi^{o:u}$ probability of remaining observable in the next period.

Recall from Section 1.2.2 that informed investors pay an amount c_M (out of the total cost c_I of becoming informed) to acquire information from the news outlets. We show in Section A.5.3 that in our calibration for a low enough c_M all informed investors prefer spending c_M on acquiring information from the news producers, over consuming c_M. [1] In deciding how much information to produce, news producers forecast next period's demand for news. A market price of a unit of information is determined to clear the news production market.

Each news outlet believes that the number of informed investors at time $t+1$ will be $\mathsf{E}_t[\lambda_{t+1}]$, and that each of the $t+1$ informed will choose to buy all of the outlet's production I_j at a price of p per unit of information. Since each outlet is small and ex-ante identical, it assumes the price p is fixed. Each outlet's profit is given by

$$\mathsf{E}_t[\lambda_{t+1}]I_j p - c_P I_j.$$

The λ_{t+1} term is present due to the zero marginal cost of transmitting information, once it's been discovered. Since the market is competitive, each outlet must operate at zero profit, which implies

$$p = \frac{c_P}{\mathsf{E}_t[\lambda_{t+1}]}.$$

Because of the zero marginal cost of sharing information with additional investors, the per unit price of information is decreasing in the number of informed investors.[2] This is similar to Veldkamp (2006), except in our model investors buy more information as the price falls, whereas in Veldkamp (2006) the cost of becoming informed varies but investors cannot choose the quantity of information they acquire.

Each investor chooses to spend c_M of the total cost c_I on buying information from all the producers. The choice to spend c_M will be discussed in Section A.5.3. Therefore, the budget constraint becomes

$$c_M = I \times p,$$

[1] We also show that uninformed investors would not choose to opt into an information set where they become informed but choose to consume c_M rather than pay it to the information production sector.

[2] A similar result obtains if we assume monopolistic competition, i.e. each media outlet produces a differentiated piece of information, while taking as given the prices of all other news outlets. This case is analyzed in Perloff and Salop (1985) and Veldkamp (2006), though it does not add to the intuition here. Monopolistic competition leads to a higher constant price, and lower I_j, than the competitive case.

where $I = \sum_j I_j$. Combining this with the zero profit condition we get

$$\frac{c_M}{I} = \frac{c_P}{\mathsf{E}_t[\lambda_{t+1}]}.$$

So the aggregate news production will be given by

$$I = \frac{c_M}{c_P}\mathsf{E}_t[\lambda_{t+1}]. \tag{A.2}$$

The information producers are assumed to be myopic and believe that tomorrow's number of informed is equal to today's number of informed.

$$\mathsf{E}_t[\lambda_{t+1}] \approx \lambda_t \tag{A.3}$$

Figure (A.1) shows that in our equilibrium, this is a very accurate approximation because, in the steady state, the economy spends very little time in the $\lambda_t \in [0, 0.05]$ region in which the approximation is less precise (see equilibrium λ in Figure 1.4 and the equilibrium steady-state distribution in Figure 1.5). With this we augment the f_{t+1} dynamics in the prior section with an additional $c_M/c_P\lambda_t$ units of information. Setting

$$b_f = \frac{c_M}{c_P} \tag{A.4}$$

yields the reconciliation of our process with that of (1.6). Section A.5.3 discusses the magnitude of feasible c_Ms in our calibration; and since c_P is a free parameter in the model, from (A.4) we see that given a c_M any b_f is attainable for some c_P.

The foregoing discussion describes the interior behavior of f_t, given in (1.5). At the left boundary, we disallow realizations of $\epsilon_{f,t+1}$ that will push f_{t+1} below zero. At the right boundary, we similarly disallow realizations of $\epsilon_{f,t+1}$ that will push f_{t+1} above one. When the $b_f\lambda_t$ term would push f_{t+1} above one, we assume this effect is exactly offset by $\epsilon_{f,t+1}$.

A.2 Model Calibration

In calibrating the model to the aggregate market, we take one period in the model to represent one month. We estimate a monthly dividend process of the form (1.1) using daily dividend data for the S&P 500 index from 1998–2018, then aggregating this up to the quarterly level (to mitigate seasonality effects), and estimating an ARMA(1,1) process

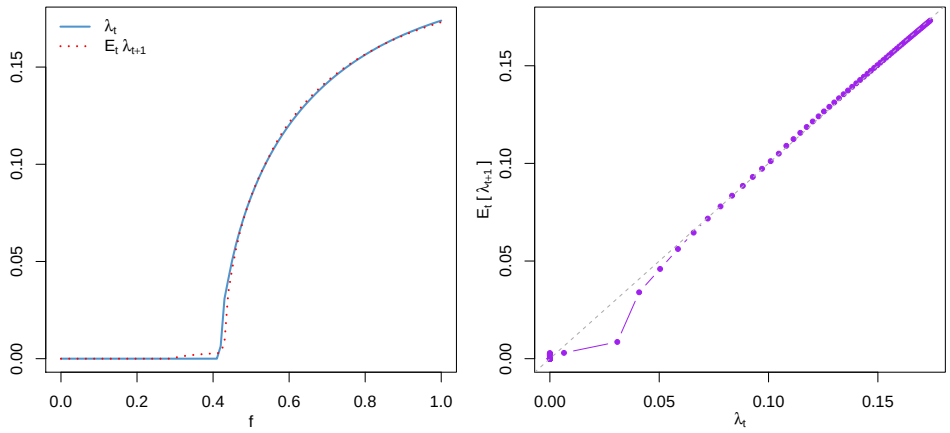


Figure A.1: Comparison of $\mathsf{E}_t[\lambda_{t+1}]$ against λ_t.

Note: The figure uses parameter values from Table A.1.

for the quarterly dividend. From this we back out the monthly parameters $\rho = .967$ and $\sigma_M = 0.0471$. See the Supplementary Appendix for details.

We adopt the normalization $\bar{D} = 1$ and $\bar{X} = 1$, so dividends and share supplies are measured in units of their monthly averages. We calibrate σ_X^2 to match monthly turnover, meaning the number of shares traded per month divided by the shares outstanding. Recall from Section 1.2.4 that in each period t, investors buy the new supply X_t originating from liquidity demanders, and investors from the previous period sell back X_{t-1} shares to the previous period's liquidity demanders unwinding their trades. The total trading volume in period t is therefore $|X_t| + |X_{t-1}|$. Using the normality of the supply shocks, the expected volume per period becomes

$$\mathsf{E}[|X_t| + |X_{t-1}|] = 2\sigma_X\sqrt{\frac{2}{\pi}} \approx 1.596\sigma_X. \tag{A.5}$$

In the Supplementary Appendix, we find that the average weekly turnover of the Dow Industrials index is 0.065.[3] To model a period of stress, we assume that turnover, or the turnover expectation of market participants, is four times higher than normal, so $\sigma_X = 4 \times 0.065 \times 1/2 \times \sqrt{\pi/2} = 0.1629$.[4]

[3]Lo and Wang (2000, Table 3) show that from 1987-1996, weekly turnover on a value-weighted index of NYSE and AMEX common shares was 1.25%. Therefore the monthly turnover on this index was $52/12 \times 1.25\% = 5.42\%$.

[4]The trading volume of SPY, which tracks the S&P 500 index and is one of the most liquid exchange-traded funds, has spiked by a factor of four during stress periods. For example, in 1/18/2011 the trailing

We use a monthly gross risk-free return of $R = 1.0015$ and set the risk-aversion parameter at $\gamma = 0.46$. This yields an annualized excess return of roughly 15%, which is not unreasonable for periods of stress. We choose a per month cost of being informed of $c_I = 0.2627$, which should be compared with a monthly aggregate average dividend of 1 (since $\bar{X} = \bar{D} = 1$). This high cost of information is comparable to the 2/20 fee structure of many hedge funds, and leads to an equilibrium number of informed of under 18% of the overall population.

For the dynamics of the f_t process in (1.6), we set $a_f = 0.175$, $\kappa_f = 0.91$ and $b_f = 0.384$. In the context of our information production microfoundation (see Section A.1), these choices of a_f and κ_f imply the probability of a unit of information transitioning from observable to unobservable is roughly five times larger than the probability of transitioning from unobservable to observable. A value of κ_f close to 1 makes the information state persistent by making both of the above probabilities small, and a positive b_f produces positive feedback from the fraction informed to the level of accessible information. When f_t is low, $\lambda(f_t) = 0$, and f_t is pulled toward a level of $a_f = 0.175$.[5] We fix $\phi = 0.35$, implying that 65% of the knowable information is publicly known. This introduces a high degree of information asymmetry between informed and uninformed investors, and leads to a large drop in price in the high-information regime; see the discussion in Section 1.5.1.

Beyond these qualitative considerations, these specific parameters were chosen to produce plausible model dynamics. Finally, for the shocks $\epsilon_{f,t+1}$, we use a three-point distribution taking values $\{-0.135, 0, 0.135\}$ with probabilities $\{0.03, 0.94, 0.03\}$, so shocks are rare. Model parameters are summarized in Table A.1.

Our results are robust to changes in model parameters. For a wide range of values in our non-dividend parameters (since ρ and σ_M are estimated from actual data) in Table A.1, there exists a ϕ close to its base value of 0.35 which generates the bimodal f_t distribution and the large price drops that we discuss below. In fact, in many cases the resultant price

month's average daily trading volume was 106.2 million and in 8/26/2011 the trailing month's daily trading volume was 407.0 million.

[5] Fama and French (2000) show R^2's of year-ahead firm-level earnings forecasts are between 5% and 20%.

$\bar{D}$	$\bar{X}$	R	σ_X	ρ	σ_M	γ	c_I	ϕ	$\mathsf{P}[\epsilon_f \neq 0]$	ϵ_f	a_f	b_f	κ_f
1	1	1.0015	0.1629	0.967	0.0471	0.46	0.2627	0.35	0.03×2	0.135	.175	.384	.91

Table A.1: Calibrated parameters for model (1.6)

drops are larger than the one that occurs under our base parameterization.[6]

A.3 Model Solution: Statement of Main Results

This appendix provides precise statements of the results described in Section 1.3. For some of the results in this section (Propositions 4–6), and for numerical calculations, we discretize the state space by restricting $\mathcal{D}$ — the set of values that f_t can take — to be a finite subset of $[0, 1]$. This discretization allows us to represent any function of f as an n-dimensional vector. Our numerical procedure is discussed in the Supplementary Appendix. The proofs are given in Sections A.4 and A.5.

A.3.1 Market Equilibrium

Recall that a market equilibrium with conjectured variance V_B consists of a price process and investor demand functions that clear the market and solve (1.9), with the true conditional variance replaced by V_B, as in (1.16).

Proposition 3. *For any conjectured variance function $V_B(\cdot)$ bounded above and bounded away from zero, and any $\lambda(\cdot)$, there exists a market equilibrium with a price process of the form (1.11) in which a_t, b_t, and c_t are functions of f_t and do not otherwise depend on t, and the constants d and g are given by $d = \rho/(R - \rho)$ and $g = 1/(R - \rho)$.*

The proof of this proposition is in Appendix A.4.1. Given price coefficient functions b and c, the conjectured-variance updating equation (1.17) defines a new V_B, and given V_B, Proposition 3 defines new coefficients b and c. (The coefficient a depends on V_B but does not enter in the update of V_B.) Combining the two steps yields a mapping from an initial pair of functions (b, c) to an updated pair (b, c). We have a rational expectations market equilibrium, at a fixed point of this mapping. We need some modest parameter restrictions to ensure existence of a fixed point. The following condition is particularly simple to state:

[6] This analysis is available from the authors.

Proposition 4. *Suppose that* $R \in [1, 1.2]$ *and the model parameters satisfy*

$$(1+d)\gamma\sigma_M\sigma_X \leq 0.28. \tag{A.6}$$

Then for any fixed $\lambda(\cdot)$*, there exists a fixed point of the conjectured-variance updating mapping. This fixed point defines a self-consistent variance conjecture and thus a rational expectations market equilibrium with prices of the form in (1.11).*

The point of condition (A.6) is that we need $(1+d)\gamma\sigma_M\sigma_X$ to be small, and with the mild bounds on R we can show that 0.28 is small enough. If we fix $f_t \equiv 0$ and $\lambda \equiv 0$ we have an OLG model without asymmetric information, similar to the one in Spiegel (1998). As in Spiegel's (1998) model, the coefficients in the price function can be expressed through solutions of quadratic equations,[7] and a condition like (A.6) arises naturally.

A.3.2 Information Equilibrium

Given a market equilibrium, we need to ensure the existence of an endogenous fraction informed. For every f, we need to find a $\lambda(f)$ that makes investors exactly indifferent between paying the cost c_I of becoming informed or staying uninformed; if no such λ exists, we set λ equal to zero or one according to Definition 1.

For technical reasons, in the following result we assume that the shocks $\epsilon_{f,t}$ have a density. The dynamics in (1.6) project f_t to a finite grid, even when the shocks are continuous. In our numerical examples, it is easier to work with a discrete distribution for $\epsilon_{f,t}$. The details of our numerical procedure are included in Appendix A.5.1 and the Supplementary Appendix.

---

[7] When $f_t \equiv 0$, the equation for a self-consistent variance conjecture reduces to solving a quadratic equation with two real roots, which is given by

$$V_B^2 + \left[2(1+d)^2\sigma_M^2 - \frac{R^2}{\gamma^2\sigma_X^2}\right] V_B + (1+d)^4\sigma_M^4 = 0.$$

The two roots describe two market equilibria, one with high price variance and one with low price variance, and we need an upper bound on the left side of (A.6) to ensure that both roots are positive. However, the high variance equilibrium is unstable under arbitrarily small parameter perturbations; only the low variance equilibrium is robust to such changes. In our numerical experiments, we find that if we start from a low value of $V_B(\cdot)$ we converge to the low variance equilibrium.

Proposition 5. *Suppose the shocks $\epsilon_{f,t}$ have a density. Then for any strictly positive V_B there exists an endogenous λ in the sense of Definition 1.*

Given an exogenous $\lambda(\cdot)$ and the associated self-consistent V_B, we can solve for a new endogenous $\lambda(\cdot)$. However, once we change λ, the conjectured variance may no longer be self-consistent. When we solve the model numerically, we start with an arbitrary[8] variance conjecture V_B, we then solve for the endogenous fraction informed (as provided by Proposition 5), we then calculate the realized variance (1.17) using the endogenous λ, update the conjectured variance and repeat. We can formulate this process as starting with a pair of coefficient functions (b, c), from which we calculate λ and then a new (b, c). Combining the two steps yields a mapping from an initial (b, c, λ) to a new (b, c, λ). A fixed point of this mapping defines an information equilibrium, in the sense that it yields a rational expectations market equilibrium in which investors do not want to deviate from their information choices. We now address the existence of such a fixed point.

For any coefficient functions (b, c), let $\Lambda_o(b, c)$ denote the set of λ satisfying Definition 1, which we know from Proposition 5 is nonempty. Let $\Lambda(b, c)$ denote the set of all convex combinations of elements of $\Lambda_o(b, c)$. If there is just one λ in $\Lambda_o(b, c)$, then $\Lambda(b, c) = \Lambda_o(b, c) = \{\lambda\}$. In our numerical experiments, instances of multiple λ satisfying Definition 1 occur rarely, and we have never encountered multiple solutions λ when using self-consistent V_Bs. However, because we have not proved the uniqueness of λ, we need to work with the potentially larger set $\Lambda(b, c)$ in establishing the existence of an information equilibrium (b, c, λ).

A convex combination $\lambda \in \Lambda(b, c)$ represents a mixed equilibrium in the following heuristic sense. For each $f \in \mathcal{D}$, we can write

$$\lambda(f) = w_f \lambda_1(f) + (1 - w_f)\lambda_2(f), \tag{A.7}$$

with $w_f \in [0, 1]$ and $\lambda_1, \lambda_2 \in \Lambda_o(b, c)$, $\lambda_2(f) > \lambda_1(f)$. Interpret this to mean that a fraction w_f of investors thought equilibrium $\lambda_1(f)$ would be selected, and a fraction $1 - w_f$ thought $\lambda_2(f)$ would be selected. At the outcome $\lambda(f)$, the marginal investor is not indifferent

[8] More precisely, we start with a V_B within the region where Proposition 4 ensures the existence of a fixed point.

between becoming informed or not. A fraction w_f of investors, expecting an outcome of $\lambda_1(f)$, will see $\lambda(f)$ as too high, and in response a fraction $w_f(\lambda(f)-\lambda_1(f))$ of investors will switch from informed to uninformed. Similarly, a fraction $1-w_f$, expecting $\lambda_2(f)$, will see $\lambda(f)$ as too low, resulting in a fraction $(1-w_f)(\lambda_2(f)-\lambda(f))$ switching from uninformed to informed. But then (A.7) implies that these effects offset each other, leaving the fraction informed at $\lambda(f)$.

We establish existence of an information equilibrium — a joint equilibrium in (b, c, λ) — within the broader class of information choices in $\Lambda(b, c)$. For the following, let $M(\lambda)$ be the set of market equilibrium parameters (b, c) consistent with the fraction informed function $\lambda = \{\lambda(f), f \in \mathcal{D}\}$; these are the fixed points in Proposition 4. The following proposition is proved in the Supplementary Appendix.

Proposition 6. *Suppose the conditions of Propositions 4 and 5 hold. Then there exists an information equilibrium (b, c, λ), meaning that $(b, c) \in M(\lambda)$ and $\lambda \in \Lambda(b, c)$. In other words, (b, c) defines a market equilibrium given λ, and λ defines a (possibly mixed) endogenous fraction informed given (b, c).*

A.4 Market Equilibrium

A.4.1 Proof of Proposition 3 (Existence of a Market Equilibrium)

Investor Demands for the Risky Asset

We prove Proposition 3 by solving explicitly for the coefficients of the price in (1.11). To allow for arbitrary V_B, we write the investor optimization problem (1.9) as

$$\hat{J}_t^\iota \equiv \max_q \mathsf{E}\left[\mathsf{E}[W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}] - \frac{\gamma}{2}\hat{\mathsf{var}}(W_{t+1}|\mathcal{I}_t^\iota, f_{t+1})\middle|\mathcal{I}_t^\iota\right], \quad \iota \in \{I, U\}, \tag{A.8}$$

where, using (1.16),

$$\hat{\mathsf{var}}(W_{t+1}|\mathcal{I}_t, f_{t+1}) = q^2(1+d)^2\left[\mathsf{var}(m_t|\mathcal{I}_t) + (1-f_t)\sigma_M^2\right] + q^2 V_B(f_{t+1}). \tag{A.9}$$

If the conjectured variance V_B is self-consistent, then (A.9) yields the conditional variance, but (A.8) makes explicit investors' objectives with arbitrary V_B.

We can write the terminal wealth in (1.7) as $W_{t+1} = RW_t + q\pi_{t+1}$. Recalling that W_t is known to time-t investors, we set $\hat{\mathsf{var}}(\pi_{t+1}|\mathcal{I}_t, f_{t+1}) = \hat{\mathsf{var}}(W_{t+1}|\mathcal{I}_t, f_{t+1})/q^2$ to get

$$\hat{\mathsf{var}}(\pi_{t+1}|\mathcal{I}_t, f_{t+1}) = (1+d)^2\left[\mathsf{var}(m_t|\mathcal{I}_t) + (1-f_t)\sigma_M^2\right] + V_B(f_{t+1}). \tag{A.10}$$

The first-order condition for optimality in (A.8) becomes

$$q_t^\iota = \frac{1}{\gamma}\frac{\mathsf{E}\left\{\mathsf{E}[\pi_{t+1}|\mathcal{I}_t^\iota, f_{t+1}]\middle|\mathcal{I}_t^\iota\right\}}{\mathsf{E}\left\{\hat{\mathsf{var}}(\pi_{t+1}|\mathcal{I}_t^\iota, f_{t+1})\middle|\mathcal{I}_t^\iota\right\}} \equiv \frac{1}{\gamma}\frac{q_N^\iota}{q_D^\iota}, \tag{A.11}$$

where q_N^ι is the conditional expectation of the net profit, and q_D^ι is the expectation of its conditional variance, given a price variance of V_B. Through (A.10), the conditional variances, reflecting the cash-flow component and the time-t price variance, take the form

$$\begin{aligned} q_D^I &= (1+d)^2(1-f_t)\sigma_M^2 + \mathsf{E}_t V_B(f_{t+1}), \\ q_D^U &= q_D^I + (1+d)^2\mathsf{var}(m_t|P_t, \theta_t). \end{aligned} \tag{A.12}$$

Evaluating the conditional mean in the numerator of (A.11) as in (1.14), the demands for time-t informed and uninformed agents become

$$\begin{aligned} q^I &= \frac{q_N^I}{\gamma q_D^I} = \frac{1}{\gamma q_D^I}\Big[(1+d)(\mu_D + \rho D_t + \theta_t + m_t) + \mathsf{E}_t a(f_{t+1}) - RP_t\Big], \\ q^U &= \frac{q_N^U}{\gamma q_D^U} = \frac{1}{\gamma q_D^U}\Big[(1+d)(\mu_D + \rho D_t + \theta_t + \mathsf{E}[m_t|P_t, D_t, \theta_t]) + \mathsf{E}_t a(f_{t+1}) - RP_t\Big]. \end{aligned} \tag{A.13}$$

For the informed, we have used the fact that $\mathsf{E}[m_t|\mathcal{I}_t^I] = m_t$ and $\mathsf{var}(m_t|\mathcal{I}_t^I) = 0$. For the uninformed, we evaluate (A.13) using[9]

$$\begin{aligned} \mathsf{E}[m_t|P_t, \theta_t] &= K_t(b_t m_t - c_t X_t), \\ \mathsf{var}(m_t|P_t, \theta_t) &= \phi f_t \sigma_M^2(1 - K_t b_t) = \phi f_t \sigma_M^2(1 - \mathcal{R}_t^2), \end{aligned} \tag{A.14}$$

with

$$K_t = \frac{\mathsf{cov}(m_t, P_t|\theta_t, D_t)}{\mathsf{var}(P_t|\theta_t, D_t)} = \frac{b_t\phi f_t\sigma_M^2}{b_t^2\phi f_t\sigma_M^2 + c_t^2\sigma_X^2} \qquad \text{and} \qquad \mathcal{R}_t^2 \equiv K_t b_t. \tag{A.15}$$

[9]Say $E[m|P] = K(bm + cX)$ for $K = \mathsf{cov}(m, P)/\mathsf{var}(P)$ and $\mathsf{var}(m|P) \equiv \mathsf{var}(m - E[m|P])$. Since $m - E[m|P] = (1-Kb)m - KcX$ then $\mathsf{var}(m - E[m|P]) = (1-Kb)^2\mathsf{var}(m) + K^2c^2\mathsf{var}(X)$. This equals $(1 - 2Kb + K^2b^2)\mathsf{var}(m) + K^2c^2\mathsf{var}(X) = (1-2Kb)\mathsf{var}(m) + K^2(b^2\mathsf{var}(m) + c^2\mathsf{var}(X))$. Note that $K = b\mathsf{var}(m)/(b^2\mathsf{var}(m) + c^2\mathsf{var}(X))$ and therefore $K^2(b^2\mathsf{var}(m) + c^2\mathsf{var}(X)) = b^2\mathsf{var}^2(m)/(b^2\mathsf{var}(m) + c^2\mathsf{var}(X)) = Kb\mathsf{var}(m)$. And therefore $\mathsf{var}(m|P) = (1-Kb)\mathsf{var}(m)$.

Market Clearing and Price Coefficients

We now impose market clearing (1.10), taking λ as given. We substitute investor demands q^ι in (1.10), use the price function from (1.11), and collect terms. We do not have to expand q_D^I or q_D^U in the following because these depend on f_t but not on D_t, m_t, θ_t, or X_t. Equation (1.10) becomes

$$\begin{aligned}\lambda q_D^U\Big[(1+d)(\mu_D+\rho D_t+\theta_t+m_t)+\mathsf{E}_t a(f_{t+1})-RP_t\Big]\\ +(1-\lambda)q_D^I\Big[(1+d)(\mu_D+\rho D_t+\theta_t+\mathsf{E}[m_t|P_t,\theta_t])+\mathsf{E}_t a(f_{t+1})-RP_t\Big]\\ =\gamma q_D^I q_D^U X_t+\gamma q_D^I q_D^U \bar{X}.\end{aligned} \tag{A.16}$$

Collecting the D_t terms and then the θ_t terms yields the constants

$$d=\frac{\rho}{R-\rho}\qquad\text{and}\qquad g=\frac{1}{R-\rho}. \tag{A.17}$$

Collecting the constant terms in (A.16) yields

$$a_t=\frac{1}{R}\left[(1+d)\mu_D-\frac{\gamma q_D^I q_D^U}{\lambda q_D^U+(1-\lambda)q_D^I}\bar{X}+\mathsf{E}_t a(f_{t+1})\right]. \tag{A.18}$$

The function a appears on both sides. Assuming for a moment that a solution $a_t=a(f_t)$ exists, we can proceed to solve for b and c because a plays no role in the inference the uninformed make from the price in (A.15). We return to solve (A.18) after solving for b and c.

Collecting the m_t terms in (A.16) we get

$$b_t=\frac{1+d}{R}\times\frac{\lambda q_D^U+(1-\lambda)q_D^I K_t b_t}{\lambda q_D^U+(1-\lambda)q_D^I}. \tag{A.19}$$

Collecting the X_t terms and simplifying — mainly dividing the resulting equation by (A.19) — we find

$$b_t/c_t=\frac{\lambda(1+d)}{\gamma q_D^I},\quad\text{with } c_t=\frac{\gamma q_D^U}{R}\text{ if }\lambda=0. \tag{A.20}$$

We can now combine these equations to solve for b and c through the following steps, each

of which involves only known quantities on the right side:

$$
\begin{aligned}
q_D^I(f) &= (1+d)^2(1-f)\sigma_M^2 + \mathsf{E}[V_B(f_{t+1})|f_t = f], && \text{(A.21)}\\
r(f) &= \lambda(f)(1+d)/(\gamma q_D^I(f)), && \text{(A.22)}\\
\mathcal{R}^2(f) &= \frac{r^2(f) f\phi\sigma_M^2}{r^2(f) f\phi\sigma_M^2 + \sigma_X^2}, && \text{(A.23)}\\
q_D^U(f) &= q_D^I(f) + (1+d)^2 f\phi\sigma_M^2(1-\mathcal{R}^2(f)), && \text{(A.24)}\\
b(f) &= \frac{1+d}{R}\frac{\lambda(f)q_D^U(f) + (1-\lambda(f))q_D^I(f)\mathcal{R}^2(f)}{\lambda(f)q_D^U(f) + (1-\lambda(f))q_D^I(f)}, && \text{(A.25)}\\
c(f) &= \begin{cases} b(f)/r(f), & \lambda(f) > 0; \\ \gamma q_D^U(f)/R, & \lambda(f) = 0. \end{cases} && \text{(A.26)}
\end{aligned}
$$

Equation (A.21) restates the first line of (A.12); (A.22) is the ratio in (A.20); (A.23) rewrites the expression for $\mathcal{R}^2$ in (A.15); (A.24) follows from the second line of (A.12); (A.25) and (A.26) come from (A.19) and (A.20).

A.4.2 Solving for the $a()$ curve

We now return to (A.18). Using the price function from (1.11) and the net profit from (1.8), we see that

$$
\begin{aligned}
\mathsf{E}_t[\pi_{t+1}] &= \mathsf{E}_t[D_{t+1} + P_{t+1} - RP_t] \\
&= \mathsf{E}_t[\mu_D + \rho D_t + a_{t+1} + d(\mu_D + \rho D_t) - Ra_t - RdD_t] \\
&= (1+d)\mu_D + \mathsf{E}_t[a_{t+1}] - Ra_t \\
&= \gamma\bar{X}\frac{q_D^I q_D^U}{\lambda q_D^U + (1-\lambda)q_D^I}, && \text{(A.27)}
\end{aligned}
$$

where the third step follows from the definition of d in (A.17) and the fourth step follows from a_t in (A.18). Using (A.27) we can rewrite a_t in (A.18) as

$$
\begin{aligned}
a_t &= \frac{1}{R}\left[(1+d)\mu_D - \mathsf{E}_t[\pi_{t+1}] + \mathsf{E}_t a(f_{t+1})\right] \\
&= \frac{1}{R}\left[(1+d)\mu_D - \mathsf{E}_t[\pi_{t+1}] + \mathsf{E}_t\left\{\frac{1}{R}\left[(1+d)\mu_D - \mathsf{E}_{t+1}[\pi_{t+2}] + \mathsf{E}_{t+1}a(f_{t+2})\right]\right\}\right] \\
&= \cdots = \frac{(1+d)\mu_D}{R-1} - \sum_{i=1}^{\infty}\frac{1}{R^i}\mathsf{E}_t[\pi_{t+i}] && \text{(A.28)}
\end{aligned}
$$

If the conjectured variance V_B is bounded above and bounded away from zero, then $|\mathsf{E}_t[\pi_{t+i}]|$ is bounded and the expression in (A.28) is well-defined and finite. The quantities in (A.21)–(A.26) and (A.27) are all functions solely of the information state f, so the conditional expectations in (A.18) and (A.28) are taken with respect to the evolution of the information state in (1.6), for given λ. Equation (A.28) shows a_t is equal to the present value of all future expected dividend payments minus a discount reflecting the expected present value of all future net profits. We show how to calculate the expectation in (A.28) in the Supplementary Appendix.

A.4.3 Statement and Proof of Proposition 4 (Existence of a Rational Expectations Equilibrium)

In this section, we prove the existence of a rational expectations equilibrium by showing that the conjectured-variance updating mapping has a fixed point. This demonstrates the existence of self-consistent V_B given an exogenously specified $\lambda()$ curve.[10]

We now precisely state Proposition 4 for model (1.6). With f_t restricted to a finite set $\mathcal{D}$, we represent any function of f_t as a vector of dimension $n = |\mathcal{D}|$. We suppose $\lambda()$ is fixed (not necessarily constant) with $0 \leq \lambda(f) \leq 1$ for all f. Let $F(b, c)$ be the mapping that sends the initial coefficients (b, c) to updated coefficients (b', c') through (1.17) and (A.21)–(A.26). A fixed point refers to the coefficients b and c such that $(b, c) = F(b, c)$.

Assume there exists a scalar $\bar{c} > 0$ satisfying the following four polynomial conditions:

$$\gamma\sigma_X\left(2\bar{q} + (1+d)^2\sigma_M^2\phi\right) - \bar{c}R\sigma_X\left(4 - (1+d)^2\gamma^2\sigma_M^2\sigma_X^2\phi\right) \leq 0, \tag{A.29}$$

$$4\gamma R\sigma_X^2\bar{c}\bar{q} - 4R^2\sigma_X^2\bar{c}^2 + (1+d)^2\sigma_M^2\phi\left(1 + \gamma R\sigma_X^2\bar{c}\right)^2 \leq 0, \tag{A.30}$$

$$\gamma^2\sigma_X^2\left(\bar{q} + (1+d)^2\sigma_M^2\phi\right) \leq 1, \tag{A.31}$$

$$\gamma\left((1+d)^2\sigma_M^2\eta + \bar{c}^2\sigma_X^2\right) - R\bar{c} \leq 0, \tag{A.32}$$

where $\bar{q}$ in (A.29), (A.30), and (A.31) is a quadratic function in $\bar{c}$,

$$\bar{q} = (1+d)^2\sigma_M^2\delta + \bar{c}^2\sigma_X^2, \tag{A.33}$$

[10]Note that this exogenous $\lambda()$ curve need not result in equivalent utilities for informed and uninformed investors. We endogenize the $\lambda()$ curve in Section A.5.

and the constants η and δ in (A.32) and (A.33) are defined by

$$\begin{aligned}\eta &:= \max_{f \in \mathcal{D}} \left\{ \frac{1}{R^2} \mathsf{E}\left[f_{t+1} | f_t = f \right] + 1 - (1-\phi) f \right\}, \\ \delta &:= \max_{f \in \mathcal{D}} \left\{ \frac{1}{R^2} \mathsf{E}\left[f_{t+1} | f_t = f \right] + 1 - f \right\}.\end{aligned} \tag{A.34}$$

Our simplest conditions would require $\delta \leq \eta \leq 2$, which is natural for $f \in [0,1]$, $\phi \in [0,1]$ and $R > 1$. The precise statement of Proposition 4 is that the mapping F has a fixed point in $[0, \bar{b}]^n \times [0, \bar{c}]^n$ with $n = |\mathcal{D}|$, $\bar{c}$ satisfying (A.29)–(A.32), and

$$\bar{b} = \frac{1+d}{R}.$$

The existence of a $\bar{c}$ satisfying (A.29)–(A.32), and thus the existence of fixed point for mapping F, only depends on model parameters.

As a shortcut for checking that these conditions hold, we show in Appendix A.4.3.1 that if $R \in [1, 1.2]$ and condition (A.6) holds, then F has a fixed point in $[0, \bar{b}]^n \times [0, \bar{c}]^n$ with $\bar{b} = (1+d)/R$ and $\bar{c} = R/(2\gamma\sigma_X^2)$.

Proof of Proposition 4

To prove the result, we use Brouwer's fixed point theorem, which states that if F is a continuous function mapping a compact convex set S to itself, then F has a fixed point in this set, meaning there exists $x \in S$ for which $x = F(x)$; see, for example, p.29 of Border (1989).

We show conditions for the Brouwer's fixed point theorem are satisfied with the V_B updating mapping F and the compact convex set $[0, \bar{b}]^n \times [0, \bar{c}]^n$, $n = |\mathcal{D}|$. First, it is evident that F is continuous as each mapping in (A.21)–(A.26) is continuous in its input. It is also evident that $F(b,c) \geq 0$ since each step in (A.21)–(A.26) returns a nonnegative value. Next, for any input (b,c), we have $b'(f) \leq (1+d)/R = \bar{b}$ for all f: it is easy to see $q_D^I(f) \geq 0$, $\mathcal{R}^2(f) \in [0,1]$, and $q_D^U(f) \geq 0$, so the second factor in (A.25) is in $[0,1]$ and the bound on $b'(f)$ follows.

It only remains to show that if $(b,c) \in [0, \bar{b}]^n \times [0, \bar{c}]^n$ and $(b', c') = F(b,c)$, then $c'(f) \leq \bar{c}$ for all f. For this we first establish two useful bounds. To lighten notation, in the following we abbreviate conditional expectations of the form $\mathsf{E}[V_B(f_{t+1}, \phi) | f_t = f]$ as

$\mathsf{E}_t[V_B(f)]$. Recalling (1.17), we get the bound

$$\begin{aligned}\mathsf{E}_t[V_B(f)] &= \mathsf{E}_t\left[b(f)^2 f\right]\phi\sigma_M^2 + g^2(1-\phi)\mathsf{E}_t[f]\sigma_M^2 + \mathsf{E}_t\left[c(f)^2\right]\sigma_X^2 \\ &\leq \bar{b}^2\mathsf{E}_t[f]\phi\sigma_M^2 + g^2(1-\phi)\mathsf{E}_t[f]\sigma_M^2 + \bar{c}^2\sigma_X^2 \\ &= (1+d)^2\frac{1}{R^2}\mathsf{E}_t[f]\sigma_M^2 + \bar{c}^2\sigma_X^2. \end{aligned} \tag{A.35}$$

The inequality uses the assumption $(b,c) \in [0,\bar{b}]^n \times [0,\bar{c}]^n$, and the last equality uses the relationship $\bar{b} = (1+d)/R = g$, which follows (A.17). Combining this with (A.21), we can bound $q_D^I(f)$ by

$$\begin{aligned} q_D^I(f) &= (1+d)^2(1-f)\sigma_M^2 + \mathsf{E}_t[V_B(f)] \\ &\leq (1+d)^2\sigma_M^2\delta + \bar{c}^2\sigma_X^2 = \bar{q}, \end{aligned} \tag{A.36}$$

where $\bar{q}$ is given in (A.33). The inequality follows from the definition of δ in (A.34). We now proceed to prove the desired bound $c'(f) \leq \bar{c}$ for all f by the following two cases.

I. The Case Without Informed Investor: $\lambda(f) = 0$

We first prove the bound for $c'(f)$ for the case $\lambda(f) = 0$. By the second case in (A.26), $c'(f)$ is now given by $c'(f) = \gamma q_D^U(f)/R$. With $\lambda(f) = 0$, (A.22), (A.23), and (A.25) lead to $r(f) = \mathcal{R}^2(f) = b'(f) = 0$. Plugging these into (A.21) and (A.24) we have

$$\begin{aligned} q_D^U(f) &= q_D^I(f) + (1+d)^2 f\phi\sigma_M^2 \\ &= (1+d)^2(1-f)\sigma_M^2 + \mathsf{E}_t[V_B(f)] + (1+d)^2 f\phi\sigma_M^2. \end{aligned}$$

With the bound for $\mathsf{E}_t[V_B(f)]$ in (A.35), we can derive

$$\begin{aligned} q_D^U(f) &\leq (1+d)^2\sigma_M^2\left(1 - f + \frac{1}{R^2}\mathsf{E}_t[f] + f\phi\right) + \bar{c}^2\sigma_X^2 \\ &\leq (1+d)^2\sigma_M^2\eta + \bar{c}^2\sigma_X^2, \end{aligned}$$

where the last inequality follows from the definition of η in (A.34). Then the updated coefficient $c'(f)$ satisfies

$$c'(f) = \frac{\gamma}{R}q_D^U(f) \leq \frac{\gamma}{R}\left((1+d)^2\sigma_M^2\eta + \bar{c}^2\sigma_X^2\right) \leq \bar{c},$$

which follows from condition (A.32).

II. The Case With Informed Investors: $\lambda(f) > 0$

Next, we prove the bound $c(f) \leq \bar{c}$ under the case $\lambda(f) > 0$. Applying the first case in (A.26), we get

$$c'(f) = \frac{b(f)}{r(f)} = \frac{\gamma}{R}\left(\frac{q_D^I(f)}{\lambda(f)} \cdot \frac{\lambda(f)q_D^U(f) + (1-\lambda(f))q_D^I(f)\mathcal{R}^2(f)}{\lambda(f)q_D^U(f) + (1-\lambda(f))q_D^I(f)}\right). \tag{A.37}$$

We need to derive a bound for the quantity in the parenthesis above.

We substitute $q_D^U(f)$ and $\mathcal{R}^2(f)$ in (A.37) using (A.24), (A.22), and (A.23). Then the right-hand side of (A.37) can be expressed as

$$\frac{\gamma}{R}\left(\frac{\lambda q_D^I(1+d)^2\sigma_M^2 f\phi + \left(q_D^I\right)^2\gamma^2\sigma_X^2(1+d)^2\sigma_M^2 f\phi + (q_D^I)^3\gamma^2\sigma_X^2}{\lambda^2(1+d)^2\sigma_M^2 f\phi + \lambda q_D^I\gamma^2\sigma_X^2(1+d)^2\sigma_M^2 f\phi + (q_D^I)^2\gamma^2\sigma_X^2}\right), \tag{A.38}$$

where we have dropped the dependence on f to simplify notation. In the following, we establish the needed bound for $c'(f)$ by combining expression (A.38), bound for $q_D^I(f)$ in (A.36), and conditions (A.29)–(A.31). In particular, we need a bound of (A.38) that is valid for all $\lambda \in (0,1]$, as we do not make λ endogenous in this proof.

We first consider the case $f = 0$ or $\phi = 0$. If this holds, (A.38) directly reduces to $\gamma q_D^I(f)/R$ itself. By (A.36), $c'(f)$ satisfies

$$c'(f) = \frac{\gamma}{R}q_D^I(f) \leq \frac{\gamma}{R}\bar{q} = \frac{\gamma}{R}\left((1+d)^2\sigma_M^2\delta + \bar{c}^2\sigma_X^2\right).$$

As we have $\delta \leq \eta$ by (A.34), the desired bound for $c'(f)$ can be established as

$$c'(f) \leq \frac{\gamma}{R}\left((1+d)^2\sigma_M^2\delta + \bar{c}^2\sigma_X^2\right) \leq \frac{\gamma}{R}\left((1+d)^2\sigma_M^2\eta + \bar{c}^2\sigma_X^2\right) \leq \bar{c},$$

where the last inequality directly follows from condition (A.32).

Next, we consider the case (A.38) for $f > 0$ and $\phi > 0$. Here we need to consider the maximum of (A.38) over all $\lambda \in (0,1]$. We compute the derivative of (A.38) with respect to λ. As the numerator and denominator of (A.38) are linear and quadratic functions in λ respectively, the numerator of its derivative is a quadratic function in λ. Furthermore, it is easy to check this quadratic function is concave and the denominator of the derivative is always positive. Thus the maximum of (A.38) is attained at the larger root of its derivative, as long as that root falls in $(0,1]$. Through algebraic manipulations, the larger root of the

derivative is given by

$$\tilde{\lambda} = \frac{-\tau + q_D^I \gamma \sigma_X \sqrt{\tau + (1+d)^2 \sigma_M^2 f\phi}}{(1+d)^2 \sigma_M^2 f\phi}, \tag{A.39}$$

where $\tau = q_D^I \gamma^2 \sigma_X^2 \left(q_D^I + (1+d)^2 \sigma_M^2 f\phi\right)$. Also, $\tilde{\lambda} \in (0,1]$ is guaranteed by condition (A.31) and the bounds $q_D^I \leq \bar{q}$ and $f \leq 1$. Thus the maximum of (A.38) is indeed attained at $\lambda = \tilde{\lambda}$.

Letting $\lambda = \tilde{\lambda}$ in (A.38), we can derive

$$c'(f) \leq \frac{\gamma \sigma_X \left(2q_D^I + (1+d)^2 \sigma_M^2 f\phi\right) + 2\sqrt{(1+d)^2 \sigma_M^2 f\phi + q_D^I \gamma^2 \sigma_X^2 \left(q_D^I + (1+d)^2 \sigma_M^2 f\phi\right)}}{R\sigma_X (4 - (1+d)^2 \gamma^2 \sigma_M^2 \sigma_X^2 f\phi)}. \tag{A.40}$$

Denote the right side by $\kappa(f, q_D^I)$. It is positive as condition (A.29) directly implies the denominator is greater than zero. As $\kappa(f, q_D^I)$ clearly increases in both q_D^I and f, it can be further bounded by setting q_D^I and f at their upper bounds $\bar{q}$ and one, respectively. Thus to establish the needed bound $c'(f) \leq \bar{c}$, it suffices to show

$$c'(f) \leq \kappa(f, q_D^I) \leq \kappa(1, \bar{q}) \leq \bar{c}. \tag{A.41}$$

Setting $q_D^I = \bar{q}$ and $f = 1$ on the right side of (A.40), the needed inequality $\kappa(1, \bar{q}) \leq \bar{c}$ becomes

$$\begin{aligned} 2\sqrt{(1+d)^2 \sigma_M^2 \phi + \bar{q}\gamma^2 \sigma_X^2 \left(\bar{q} + (1+d)^2 \sigma_M^2 \phi\right)} \quad &\leq \quad \bar{c} R \sigma_X \left(4 - (1+d)^2 \gamma^2 \sigma_M^2 \sigma_X^2 \phi\right) \\ &\quad -\gamma \sigma_X \left(2\bar{q} + (1+d)^2 \sigma_M^2 \phi\right). \end{aligned} \tag{A.42}$$

By condition (A.29), the right side of (A.42) is positive, so this inequality is equivalent to the one obtained by squaring both sides. Taking squares and simplifying, we can show the quadratic terms $\bar{q}^2$ cancel out, and inequality (A.42) is equivalent to

$$4\gamma R \sigma_X^2 \bar{c} \bar{q} \leq 4R^2 \sigma_X^2 \bar{c}^2 - (1+d)^2 \sigma_M^2 \phi \left(1 + \gamma R \sigma_X^2 \bar{c}\right)^2,$$

which holds as long as $\bar{c}$ and $\bar{q}$ satisfy condition (A.30). By (A.41), this proves the needed bound $c'(f) \leq \bar{c}$ for $\lambda(f) > 0$. Combining the two cases with $\lambda(f) = 0$ and $\lambda(f) > 0$, we have proved $c' \leq \bar{c}$ holds when conditions (A.29)–(A.32) are satisfied. The existence of a fixed point for the V_B updating mapping now follows by Brouwer's theorem.

A.4.3.1 The Simplified Condition in (A.6)

Finally, we prove that (A.6) is a simple sufficient condition for the existence of fixed point. We follow the general conclusions established above and show that as long as $R \in [1, 1.2]$ and condition (A.6) hold, then the value $\bar{c} = R/(2\gamma\sigma_X^2)$ satisfies conditions (A.29)–(A.32). Thus by the proposition, a fixed point of V_B updating exists in $[0, (1+d)/R]^n \times [0, R/(2\gamma\sigma_X^2)]^n$.

Plugging $\bar{c} = R/(2\gamma\sigma_X^2)$ into (A.29)–(A.32), algebraic computation shows that they simplify to following equivalent conditions:

$$(1+d)^2\gamma^2\sigma_M^2\sigma_X^2 \leq \frac{3R^2}{3R^2+4\delta+6}, \quad (1+d)^2\gamma^2\sigma_M^2\sigma_X^2 \leq \frac{2R^4}{(4+4R^2+R^4)\phi+8R^2\delta}, \tag{A.43}$$

$$(1+d)^2\gamma^2\sigma_M^2\sigma_X^2 \leq \frac{4-R^2}{4(\delta+\phi)}, \text{ and } (1+d)^2\gamma^2\sigma_M^2\sigma_X^2 \leq \frac{R^2}{4\eta}, \tag{A.44}$$

respectively. A nice property of these new conditions is that they are all upper bounds for the product $(1+d)\gamma\sigma_M\sigma_X$. Thus it suffices to propose a bound for $(1+d)\gamma\sigma_M\sigma_X$ that is small enough such that all the four conditions are satisfied. Since all the bounds in the right-hand sides of (A.43) and (A.44) are decreasing in ϕ, δ, and η, it suffices to consider their values at $\phi = 1$ and $\delta = \eta = 2$, as by (A.34) we clearly have $\delta \leq \eta \leq 2$. On the other hand, as the bounds do not monotonically depend on the risk-free return R, we impose a relatively mild condition $R \in [1, 1.2]$. Plugging these values into the right-hand sides of (A.43) and (A.44), the minimum of the four upper bounds approximately equals to 0.283, thus (A.6) is sufficient for the four conditions. Consequently, the corresponding $\bar{c} = R/(2\gamma\sigma_X^2)$ is a valid upper bound for the existence of fixed point.

A.5 Information Equilibrium

A.5.1 Equate Expected Utility of Informed to Uninformed given V_B

In this section we discuss the procedure for solving for an endogenous λ given a conjectured variance V_B. Given the demands in (A.12) and V_B, we have, for $\iota \in \{I, U\}$,

$$\begin{aligned} \mathsf{E}\big[\mathsf{E}[W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}]\big|\mathcal{I}_t^\iota\big] &= q^\iota \times q_N^\iota + RW_t = \frac{(q_N^\iota)^2}{\gamma q_D^\iota} + RW_t, \\ \mathsf{E}\big[\mathsf{var}(W_{t+1}|\mathcal{I}_t^\iota, f_{t+1})\big|\mathcal{I}_t^\iota\big] &= (q^\iota)^2 \times q_D^\iota = \frac{(q_N^\iota)^2}{\gamma^2 q_D^\iota}. \end{aligned} \tag{A.45}$$

We can therefore write the agent's value function in (1.9), conditional on $\mathcal{I}_t^\iota$ as

$$J_t^\iota = RW_t + \frac{1}{2\gamma}\frac{(q_N^\iota)^2}{q_D^\iota}, \quad \iota \in \{I, U\}. \tag{A.46}$$

To find an endogenous λ in the sense of Definition 1, we need to evaluate the conditional expectation of $J_t^I - J_t^U$ given the information state f_t. We can pull the denominators q_D^ι out of the conditional expectation because we know from (A.21) and (A.24) that they are purely functions of the information state. For the numerator terms, using the demands from (A.13), the price process from (1.11) and the condition on a_t in (A.18), it is straightforward to show that

$$q_N^\iota = \mathsf{E}_t[\pi_{t+1}] + (1+d)\mathsf{E}[m_t|\mathcal{I}_t^\iota] - Rb_t m_t + Rc_t X_t$$

where $\mathsf{E}_t[\pi_{t+1}]$ — which is a function of λ — is given by (A.27). The θ_t and D_t terms drop out, as do the terms involving a_t.[11] Note that q_N^ι equals the expected net profit $\mathsf{E}_t[\pi_{t+1}]$, which only conditions on f_t, adjusted for the information set of agent ι.

Since $\mathsf{E}[m_t|\mathcal{I}_t^I] = m_t$ we have

$$\mathsf{E}[(q_N^I)^2|f_t] = (\mathsf{E}_t[\pi_{t+1}])^2 + (1+d-Rb_t)^2\phi f_t\sigma_M^2 + R^2c_t^2\sigma_X^2. \tag{A.47}$$

And from (A.14) we have $\mathsf{E}[m_t|\mathcal{I}_t^U] = K_t b_t m_t - K_t c_t X_t$. From this we have that

$$\begin{aligned}\mathsf{E}[(q_N^U)^2|f_t] &= (\mathsf{E}_t[\pi_{t+1}])^2 + [(1+d)K_t b_t - Rb_t]^2\phi f_t\sigma_M^2 + [Rc_t - (1+d)K_t c_t]^2\sigma_X^2 \\ &= (\mathsf{E}_t[\pi_{t+1}])^2 + [(1+d)K_t - R]^2(b_t^2\phi f_t\sigma_M^2 + c_t^2\sigma_X^2).\end{aligned} \tag{A.48}$$

Combining these expressions with J_t^ι in (A.46), we get an expression for the difference in conditional expectations

$$\Delta_f = \mathsf{E}[J_t^I - Rc_I|f_t = f] - \mathsf{E}[J_t^U|f_t = f]. \tag{A.49}$$

When this difference is positive, the marginal investor has an incentive to become informed. For a given f we numerically solve for the $\lambda \in [0,1]$ which sets $\Delta_f = 0$. If this difference is always strictly positive we set $\lambda = 1$, and if it is always strictly negative we set $\lambda = 0$.

[11] In particular, we do not need to evaluate $a(\cdot)$ to find the endogenous $\lambda(\cdot)$, which is useful in solving the model numerically.

A.5.2 Proof of Proposition 5 (Existence of $\lambda()$ given V_B)

Recall that we have restricted f_t to a finite set $\mathcal{D}$. We need to be more explicit about the mapping to $\mathcal{D}$ in (1.6). Suppose $\mathcal{D} = \{s_1, \ldots, s_n\} \subset [0,1]$. Partition the extended real line using

$$-\infty = c_0 < c_1 < \cdots < c_n < c_{n+1} = \infty,$$

and let $\Pi_{\mathcal{D}} : (c_j, c_{j+1}] \mapsto s_{j+1}$, $j = 0, 1, \ldots, n$. We prove the proposition for this choice of $\Pi_{\mathcal{D}}$.

We can write the difference in expected utilities (A.49) as

$$\Delta_f = \frac{1}{2\gamma}\mathsf{E}\left[\frac{q_N^{I2}}{q_D^I} - \frac{q_N^{U2}}{q_D^U}\middle| f_t = f\right] - Rc_I = \frac{1}{2\gamma}\left(\frac{\mathsf{E}[q_N^{I2}|f_t = f]}{q_D^I(f)} - \frac{\mathsf{E}[q_N^{U2}|f_t = f]}{q_D^U(f)}\right) - Rc_I. \quad \text{(A.50)}$$

The terms on the right depend on the mapping $\lambda : \mathcal{D} \mapsto [0,1]$. However, for each f, Δ_f depends on λ only through $\lambda(f)$. This follows from the expressions in (A.21)–(A.26). We may therefore write Δ_f as $\Delta_f(\ell)$, with the interpretation that ℓ is the value of $\lambda(f)$. The proposition will follow once we show that $\Delta_f(\cdot)$ is continuous: given continuity, either $\Delta_f(\ell^*) = 0$ at some $\ell^* \in [0,1]$ (in which case we set $\lambda(f) = \ell^*$), or $\Delta_f(\ell) < 0$ for all $\ell \in [0,1]$ (in which case we set $\lambda(f) = 0$), or $\Delta_f(\ell) > 0$ for all $\ell \in [0,1]$ (in which case we set $\lambda(f) = 1$). This specification satisfies the conditions in Definition 1.

To establish continuity of Δ, we use the representation in (A.50). It is evident that, holding f fixed, each of the operations in (A.22)–(A.25) is continuous in $\ell = \lambda(f)$. But λ is also implicit in (A.21) through the conditional expectation of the conjectured variance, which takes the form

$$\mathsf{E}[V_B(f_{t+1})|f_t = s_i] = \sum_{s_j \in \mathcal{D}} \mathsf{P}(f_{t+1} = s_j|f_t = s_i)V_B(s_j).$$

With $\lambda(s_i) = \ell$, the transition probabilities take the form

$$\begin{aligned}\mathsf{P}(f_{t+1} = s_j|f_t = s_i) &= \mathsf{P}(a_f + b_f\ell + \kappa_f(f_t - a_f) + \epsilon_{f,t+1} \in (c_j, c_{j+1}]|f_t = s_i) \\ &= \mathsf{P}(\epsilon_{f,t+1} \in (c_j - [a_f + b_f\ell + \kappa_f(s_i - a_f)], c_{j+1} - [a_f + b_f\ell + \kappa_f(s_i - a_f)]),\end{aligned}$$

which is the integral of the density of $\epsilon_{f,t+1}$ over the indicated interval and is therefore continuous in the endpoints and in ℓ. It follows that $\mathsf{E}[V_B(f_{t+1})|f_t = f]$ is continuous in

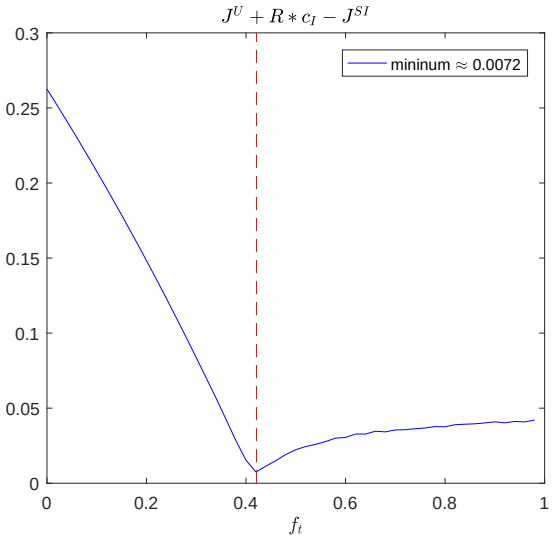


Figure A.2: Uninformed utility J^U and semi-informed utility J^{SI} as functions of f

Note: Comparison of the uninformed utility J^U to that of the semi-informed utility J^{SI} averaged over the steady-state conditional probabilities as shown in (A.52). The red, vertical line shows the threshold f_t at which $\lambda > 0$. The figure uses parameter values from Table A.1.

$\lambda(f)$, and therefore that the mapping (A.21)–(A.25) is continuous in $\lambda(f)$, including, in particular, $q_D^I(f)$ and $q_D^U(f)$.

Next we turn to (A.26) and verify that $c(f)$ is continuous at $\lambda(f) = 0$. Using (A.37), we can write, for $\lambda(f) > 0$,

$$c(f) = \frac{\gamma}{R} q_D^I(f) \left(\frac{\lambda(1+d)^2\sigma_M^2 f\phi + q_D^I(f)\gamma^2\sigma_X^2(1+d)^2\sigma_M^2 f\phi + (q_D^I(f))^2\gamma^2\sigma_X^2}{\lambda^2(1+d)^2\sigma_M^2 f\phi + \lambda q_D^I(f)\gamma^2\sigma_X^2(1+d)^2\sigma_M^2 f\phi + (q_D^I(f))^2\gamma^2\sigma_X^2} \right).$$

As $\lambda(f) \to 0$, we have $\mathcal{R}^2(f) \to 0$ and

$$c(f) \to \frac{\gamma}{R} \left((1+d)^2\sigma_M^2 f\phi + q_D^I(f) \right) = \frac{\gamma}{R} q_D^U(f),$$

which coincides with the value specified for $c(f)$ in (A.26) at $\lambda(f) = 0$.

A.5.3 Out of Equilibrium Utility and the Value of c_M

Given our model equilibrium, we analyze the incentive of an atomic agent to become informed but to forego the output of the information production sector. Because the agent

is atomic, the agent's deviation will not affect the equilibrium. By foregoing the information production output, the agent only spends $c_I - c_M$ on acquiring information; we refer to such agents as semi-informed. The semi-informed learn only a portion m_o (suppressing the t subscript) of m_t

$$m_t = m_o + m_u \qquad \text{with } m_o \perp m_u, \tag{A.51}$$

where m_u is the output of the information production sector. An agent evaluating this semi-informed information set $\mathcal{I}^{SI}$ solves the same problem as in (1.9). The solution is similar to the derivation in Section A.4 and is in the Supplementary Appendix. We now verify the conditions under which uninformed and informed agents prefer their information sets to being semi-informed.

In particular, we verify in our model calibration that for a low enough c_M, the expected utility of the semi-informed $J^{SI} + Rc_M$ and that of the uninformed $J^U + Rc_I$ satisfy

$$J^U(f) + Rc_I \geq J^{SI}(f) + Rc_M,$$

for all f. When $\lambda = 0$, the utility of the uninformed dominates that of the informed. When $\lambda \in (0, 1)$, the utility of the uninformed equals that of the informed. In our calibration $\lambda < 1$ for all f so we do not need to consider the case when the informed are better off. Therefore if the uninformed would never prefer to be semi-informed, neither would the informed agents.

In the context of our microfoundation, we can think of the variance of m_o as being $(f_t - b_f \lambda_{t-1})\phi\mathsf{var}(M)$ (recall that $\mathsf{var}(m_t) = f_t\phi\mathsf{var}(M)$). The loss in precision from not paying c_M for the information sector's output is $b_f \lambda_{t-1}$, where $\lambda_{t-1} = \lambda^*(f_{t-1})$ is the time $t-1$ endogenous fraction resulting from the prior information state f_{t-1}.

Given the information sets $\mathcal{I}^U$ and $\mathcal{I}^I$ from Section 1.2.4, agents observe f_t. They are unaware of f_{t-1}. We assume agents make decisions consistent with the steady-state f_t distribution from Figure 1.5. Writing $\tilde{J}^{SI}(f_t, f_{t-1})$ for the utility the semi-informed would attain if they observed both f_t and f_{t-1}, we then have $J^{SI}(f_t) = \mathsf{E}[\tilde{J}^{SI}(f_t, f_{t-1})|f_t]$. To compare $J_U(f) + Rc_I$ and $J^{SI}(f) + Rc_M$, agents check the condition

$$J_U(f) + Rc_I - \sum_{f'} \tilde{J}^{SI}(f, f')\mathsf{P}(f_{t-1} = f'|f_t = f) > Rc_M, \tag{A.52}$$

where

$$\mathsf{P}(f_{t-1} = f' | f_t = f) = \frac{\mathsf{P}(f_t = f | f_{t-1} = f') \times \mathsf{P}(f_{t-1} = f')}{\mathsf{P}(f_t = f)}.$$

In this expression, $\mathsf{P}(f_{t-1} = f')$ and $\mathsf{P}(f_t = f)$ are the steady-state probabilities of f' and f, respectively. For fixed f, the distribution of $\epsilon_{f,t}$ in Table A.1 implies that $\mathsf{P}(f_t = f | f_{t-1} = f')$ is nonzero for three values of f' associated with the three possible shocks, ϵ_f, 0, and $-\epsilon_f$. The sum in (A.52) therefore reduces to a sum over these three possible shocks.

Figure A.2 shows the left side of (A.52) for all values of f. We see that this difference is always positive, with a minimum value of 0.0072. Therefore for any $c_M < 0.0072/R$, neither the uninformed nor informed agents in our equilibrium would choose to deviate from their equilibrium information choice.

We note from Figure A.2 the welfare benefit to the informed relative to the semi-informed is low (this is the part of the curve to the right of the vertical red line which is the threshold f_t at which the fraction informed turns positive). In our calibration $c_I = 0.2627$ and the maximum benefit of being informed relative to semi-informed is roughly 0.04. The intuition is that the increase in information precision from the $b_f \lambda_t$ part of the f_{t+1} dynamics in (1.5) is small. And yet, despite the small welfare contribution of this feedback term, the feedback has a pronounced effect on the market equilibrium.

A.6 Proof of Proposition 6 (Existence of Information Equilibrium)

To prove the result, we apply Kakutani's fixed point theorem, which states the following (see, for example, p.72 of Border 1989). Let the domain S be a non-empty, compact and convex set, and let $F : S \mapsto 2^S$ be a set-valued function on S. Suppose $F(x)$ is non-empty and convex for all $x \in S$, and suppose that F has a closed graph (as defined shortly). Then F has a fixed point, meaning a point $x \in S$ for which $x \in F(x)$.

Let $G(b, c, \lambda)$ be the mapping that sends initial coefficients (b, c) to updated coefficients (b', c') using λ through (A.21)–(A.26). Let

$$F(b, c, \lambda) = (b', c', \Lambda(b', c')) = (G(b, c, \lambda), \Lambda(G(b, c, \lambda)))$$

be the mapping that returns the updated (b', c') and the set of λs in $\Lambda(b', c')$. This representation is consistent with our algorithm: we first update (b, c) given λ; we then solve for the endogenous λ given the new (b', c'). The mapping F returns all of $\Lambda(b', c')$, rather than a single element.

Based on the proof of Proposition 4, we can restrict (b, c) to a domain $[0, \bar{b}]^n \times [0, \bar{c}]^n$, $n = |\mathcal{D}|$, in the sense that $G(\cdot, \cdot, \lambda)$ maps this set into itself for any λ. We may therefore take the domain of F to be $S = [0, \bar{b}]^n \times [0, \bar{c}]^n \times [0, 1]^n$, which is compact and convex. Moreover, for any $(b, c, \lambda) \in S$, $F(b, c, \lambda)$ is non-empty (by Propositions 3 and 5), and it is convex by the definition of $\Lambda(b, c)$.

It only remains to show that F has a closed graph. The closed graph property states that for any sequences $x_n \to x$, $y_n \to y$ with $y_n \in F(x_n)$ we have $y \in F(x)$. Because G is single-valued and continuous (as in the proof of Proposition 4), it suffices to show the following:

$$\text{if } \lambda' \notin \Lambda(b', c'), \text{ then } \lambda \notin \Lambda(b, c) \text{ for all } (b, c, \lambda) \text{ in a neighborhood of } (b', c', \lambda'). \quad \text{(A.53)}$$

We detail the case of model (1.6). Recall from the discussion surrounding (A.49) that when we solve for a $\lambda \in \Lambda_o(b, c)$, we may solve for each $\lambda(f)$, $f \in \mathcal{D}$, separately; the conditions on $\lambda(f)$ for different values of f do not interact. For each f, we look for a point at which

$$\Delta_f(\ell) \equiv \Delta_{b,c,f}(\ell) = \mathsf{E}[J_t^I - Rc | f_t = f] - \mathsf{E}[J_t^U | f_t = f]$$

crosses zero and set $\lambda(f) = \ell$; if zero is never crossed, we get a boundary case of $\lambda(f) = 0$ or 1. We have written $\Delta_{b,c,f}$ to emphasize that the utilities on the right are evaluated using (b, c).

We know from Section 5 that $\Delta_{b,c,f}(\ell)$ is continuous in ℓ for each f, and continuity in (b, c) follows similarly from (A.49). If $\Delta_{b,c,f}(\cdot)$ crosses zero, then we may define the first and last zero crossings by

$$\ell_{\min}(f) = \min\{\ell \in [0, 1] : \Delta_{b,c,f}(\ell) = 0\}, \quad \ell_{\max}(f) = \max\{\ell \in [0, 1] : \Delta_{b,c,f}(\ell) = 0\};$$

otherwise, set $\ell_{\min}(f) = \ell_{\max}(f) = 0$ if $\Delta_{b,c,f}(\ell) < 0$ for all ℓ, and $\ell_{\min}(f) = \ell_{\max}(f) = 1$ if $\Delta_{b,c,f}(\ell) > 0$ for all ℓ.

Returning to (A.53), it now follows that if $\lambda' \notin \Lambda(b', c')$ then it must be that $\lambda'(f) \notin [\ell_{\min}(f), \ell_{\max}(f)]$ for some f, again because the constraints on each $\lambda'(f)$ depend only on that f. In particular, then, it must be that $\Delta_{b',c',f}(\ell) \neq 0$ for all $\ell \in [0, \lambda'(f)]$ or for all $\ell \in [\lambda'(f), 1]$; it suffices to consider the first case because a symmetric argument works for the second case. Suppose $\Delta_{b',c',f}(\lambda'(f)) < 0$; a symmetric argument applies if $\Delta_{b',c',f}(\lambda'(f)) > 0$. Then

$$\Delta_{b',c',f}(\ell) < 0 \text{ for all } \ell \in [0, \lambda'(f)]. \tag{A.54}$$

We claim that this holds in a neighborhood of (b', c', λ'). To argue by contradiction, suppose not. In other words, suppose that in any ϵ neighborhood of (b', c', λ') we can find a point $(b_\epsilon, c_\epsilon, \lambda_\epsilon)$ and an $\ell_\epsilon \in [0, \lambda_\epsilon(f)]$ with $\Delta_{b_\epsilon,c_\epsilon,f}(\ell_\epsilon) = 0$. Taking a sequence of ϵs decreasing to zero, gives us a sequence of such $(b_\epsilon, c_\epsilon, \lambda_\epsilon)$ and ℓ_ϵ, with $(b_\epsilon, c_\epsilon, \lambda_\epsilon) \to (b', c', \lambda')$. As the ℓ_ϵ take values in the compact set $[0, 1]$, they have a convergent subsequence. So, by taking a subsequence ϵ' of the original ϵ values, we get, for some ℓ_0, $(b_{\epsilon'}, c_{\epsilon'}, \lambda_{\epsilon'}, \ell_{\epsilon'}) \to (b', c', \lambda', \ell_0)$. And since $\ell_{\epsilon'} \leq \lambda_{\epsilon'}(f)$ for all ϵ', we have $\ell_0 \leq \lambda'(f)$. By the continuity of Δ,

$$\Delta_{b',c',f}(\ell_0) = \lim_{\epsilon' \to 0} \Delta_{b_{\epsilon'},c_{\epsilon'},f}(\ell_{\epsilon'}) = \lim_{\epsilon' \to 0} 0 = 0,$$

which contradicts (A.54). We have thus shown that (A.54) holds in a neighborhood of (b', c', λ'). But then $\lambda \notin \Lambda(b, c)$, for all (b, c, λ) in a neighborhood of (b', c', λ'), which is what we needed to show to prove the closed graph property.

A.7 Derivation of Utility of Semi-Informed

This section derives the semi-informed expected utility discussed in Section A.5.3 of the paper. Such agents assume that the economy remains in equilibrium, and evaluate the expected utility of a deviation from the equilibrium behavior of other informed agents. We assume that semi-informed agents observe m_o of $m_t = m_o + m_u$ where the variable χ_t is the amount of the variance of m_t that is observable via m_o, i.e.

$$\mathsf{var}(m_o) = \chi_t \mathsf{var}(m_t) = \chi_t f_t \phi \mathsf{var}(M).$$

Here $1-\chi_t$ reflects the amount of foregone information by not buying the output of the information producers. In the context of (A.51) from the main paper, we have

$$\chi = \frac{f_t - b_f \lambda_{t-1}}{f_t},$$

where λ_{t-1} can have one of three values depending on whether $\epsilon_{f,t}$ equals ϵ_f, 0, or $-\epsilon_f$, as discussed in Section A.5.3 in the paper.

An agent evaluating the semi-informed information set $\mathcal{I}^{SI}$ solves the following problem:

$$\tilde{J}(f_t, f_{t-1}) = \max_q \mathsf{E}_{\mathcal{I}} \left[\mathsf{E}[W_{t+1}|\mathcal{I}, f_{t+1}] - \frac{\gamma}{2} \mathsf{var}(W_{t+1}|\mathcal{I}, f_{t+1}) \right],$$

where $\mathcal{I}$ represent the information set under consideration. Here f_{t-1} is determines the value of λ_{t-1}, and therefore χ. In light of (1.14) and (1.16) this can be written as

$$J_{\mathcal{I}} = \max_q \mathsf{E}_{\mathcal{I}} \left[q\textcircled{a} - \frac{\gamma}{2} q^2 \textcircled{b} \right],$$

where

$$\textcircled{a} = (1+d)(\mu_D + \rho D_t + \theta_t + \mathsf{E}[m_t|\mathcal{I}]) + \mathsf{E}_{f_t} a(f_{t+1}) - RP_t,$$
$$\textcircled{b} = (1+d)^2 \left[\mathsf{var}(m_t|\mathcal{I}) + (1-f_t)\sigma_M^2 \right] + \mathsf{E}_{f_t} V_B(f_{t+1}),$$

and $\textcircled{b}$ is known given $\mathcal{I}$. Note that the $a(\cdot)$ and $V_B(\cdot)$ functions are those from the equilibrium of the paper. The first-order condition is $q = \textcircled{a}/(\gamma\textcircled{b})$. Plugging this back into the value function we get

$$J_{\mathcal{I}} = \mathsf{E}_{\mathcal{I}} \left[\frac{\textcircled{a}}{\gamma\textcircled{b}} \textcircled{a} - \frac{\gamma}{2} \frac{\textcircled{a}^2}{\gamma^2 \textcircled{b}^2} \textcircled{b} \right] = \frac{1}{2\gamma} \mathsf{E}_t \frac{\textcircled{a}^2}{\textcircled{b}} = \frac{1}{2\gamma} \frac{\mathsf{E}_t \textcircled{a}^2}{\textcircled{b}}, \tag{A.55}$$

In light of the value of the linear price form in (1.11) and d, g in (A.17), the D_t and θ_t terms in $\textcircled{a}$ and in RP_t cancel and we get

$$\textcircled{a} = (1+d)(\mu_D + \mathsf{E}[m_t|\mathcal{I}]) + \mathsf{E}_{f_t} a(f_{t+1}) - R(a_t + b_t m_t + c_t X_t),$$

From (A.28) we have $Ra_t = (1+d)\mu_D - \mathsf{E}_t[\pi_{t+1}] + \mathsf{E}_{f_t} a(f_{t+1})$. And therefore we can write

$$\textcircled{a} = (1+d)\mathsf{E}[m_t|\mathcal{I}] + \mathsf{E}_t[\pi_{t+1}] - R(b_t m_t + c_t X_t),$$

For the semi-informed agent, the signal contained in the price is $\hat{P}_t = b_t m_u + c_t X_t$. If he also conditions on the price to learn about m_u then his information set is $\mathcal{I} = \{m_o, \hat{P}_t\}$ and

$$\begin{aligned}
\mathsf{E}[m_t|\mathcal{I}] &= m_o + \mathsf{E}[m_u|\hat{P}_t] \\
\mathsf{E}[m_u|\hat{P}_t] &= K(b_t m_u + c_t X_t) \\
K &= \frac{b_t(1-\chi_t) f_t \phi \mathsf{var}(M)}{b_t^2(1-\chi_t) f_t \phi \mathsf{var}(M) + c_t^2 \mathsf{var}(X)} \\
\mathsf{var}(m_t|\mathcal{I}) &= (1-\mathcal{R}^2)(1-\chi) f_t \phi \mathsf{var}(M) \\
\mathcal{R}^2 &= b_t K
\end{aligned}$$

In this case we have

$$\begin{aligned}
ⓐ &= (1+d)(m_o + K b_t m_u + K c_t X_t) + \mu_{\pi,t} - R(b_t m_t + c_t X_t) \\
&= (1 + d - R b_t) m_o + ((1+d)K - R) b_t m_u + ((1+d)K - R) c_t X_t + \mu_{\pi,t} \\
&= (1 + d - R b_t) m_o + ((1+d)K - R)(b_t m_u + c_t X_t) + \mu_{\pi,t}.
\end{aligned}$$

From this we calculate

$$\begin{aligned}
\mathsf{E}_t ⓐ^2 = \mu_{\pi,t}^2 &+ (1 + d - R b_t)^2 \chi_t f_t \phi \mathsf{var}(M) \\
&+ ((1+d)K - R)^2 [b_t^2 (1-\chi_t) f_t \phi \mathsf{var}(M) + c_t^2 \mathsf{var}(X)].
\end{aligned}$$

A.8 Correlated Shocks Case

We prove that our model solution for the correlated case in (1.20) is essentially unchanged under the assumption that the noise term $\epsilon_{f,t+1}$ is independent of time-t variables and has a zero mean. We modify the definition of f_t to

$$f_t = \frac{\mathsf{var}(\tilde{m}_t)}{\mathsf{var}(\tilde{m}_t + \epsilon_{t+1})}.$$

That is, f_t is the fraction of the knowable part of the dividend innovation that is orthogonal to the information shock.

First, we follow the proof in Section D to show the market equilibrium still holds with the price process taking the form in (1.11). Under the correlated case, the conditional mean

of terminal wealth $W_{t+1} = RW_t + q(D_{t+1} + P_{t+1} - P_t)$ can be solved as

$$\mathsf{E}[W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}] = q[(1+d)(\mu_D+\rho D_t+\theta_t+\mathsf{E}[m_t|\mathcal{I}_t^\iota]-\epsilon_{f,t+1})+a(f_{t+1})-RP_t]+RW_t, \quad \iota \in \{I, U\}. \tag{A.56}$$

For the conditional variance, suppose the conjectured variance V_B is self-consistent, we have

$$\hat{\mathsf{var}}(W_{t+1}|\mathcal{I}_t^\iota, f_{t+1}) = q^2(1+d)^2\left[\mathsf{var}(m_t|\mathcal{I}_t^\iota) + (1-f_t)\sigma_M^2\right] + q^2 V_B(f_{t+1}). \tag{A.57}$$

Here we use the fact that the value of $\epsilon_{f,t+1}$ can be fully determined conditional on $\mathcal{I}_t$ and f_{t+1}. The first-order condition for investor's utility maximization problem is

$$q_t^\iota = \frac{\mathsf{E}\{\mathsf{E}[\pi_{t+1}|\mathcal{I}_t^\iota, f_{t+1}]|\mathcal{I}_t^\iota\}}{\mathsf{E}\{\hat{\mathsf{var}}(\pi_{t+1}|\mathcal{I}_t^\iota, f_{t+1})|\mathcal{I}_t^\iota\}} = \frac{1}{\gamma}\frac{q_N^\iota}{q_D^\iota},$$

where the net profit π_{t+1} is defined by (1.8). By (A.57), it is easy to verify the conditional variances q_D^I and q_D^U are still given by (A.12) as for the independent shocks case. For conditional means, we have by (A.56) that

$$\begin{aligned} q_N^\iota &= \mathsf{E}\{\mathsf{E}[\pi_{t+1}|\mathcal{I}_t^\iota, f_{t+1}]\mathcal{I}_t^\iota\} \\ &= (1+d)(\mu_D + \rho D_t + \theta_t + \mathsf{E}[m_t|\mathcal{I}_t^\iota] - \mathsf{E}[\epsilon_{f,t+1}|\mathcal{I}_t^\iota] + \mathsf{E}_t a(f_{t+1})) - RP_t. \end{aligned}$$

With the assumption that the noise term $\epsilon_{f,t+1}$ is independent of time-t variables and has a zero mean, the additional term satisfies $\mathsf{E}[\epsilon_{f,t+1}|\mathcal{I}_t^\iota] = 0$. Thus, the expressions of conditional means q_N^I and q_N^U also coincide with their counterparts in the independent shocks case.

As both conditional means and variances are not changed in the correlated shocks case, the investor demands and the market clearing condition are still given by (A.13) and (A.16) respectively. Then we can follow the steps in (A.17) – (A.26) to prove the market equilibrium holds with the price process taking the form in (1.11). Similarly, for the net profit π_{t+1}, we can show $\mathsf{E}_t[\pi_{t+1}]$ still follows by (A.27) under the assumption $\mathsf{E}_t[\epsilon_{f,t+1}] = 0$. Thus, the $a()$ curve does not change in the correlated shocks case by (A.18), (A.27), and (A.28).

Finally, it is easy to see the information equilibrium is not impacted in the correlated shocks case, i.e., the endogenous $\lambda()$ curve is the same as its counterpart in the independent shocks case. By (A.45) and (A.46), the agent's value function conditional on $\mathcal{I}_t^\iota$ is given by

$$J_t^\iota = RW_t + \frac{1}{2\gamma}\frac{(q_N^\iota)^2}{q_D^\iota}, \quad \iota \in \{I, U\}.$$

As both q_N^ι and q_D^ι remain unchanged in the correlated shocks case, so does the value function J_t^ι for $\iota \in \{I, U\}$. This leads to the same endogenous $\lambda()$ as in the independent shocks case.

Combining the above results, we show the model solution for correlated shocks case is essentially the same as the independent shocks case. The case of $\epsilon_{f,t+1} = \epsilon_{f1,t+1} + \epsilon_{f2,t+1}$ and $M_{t+1} = m_t + \theta_t + \epsilon_{t+1} - h \times \epsilon_{f1,t+1}$ can be proved in a similar manner.

A.9 Numerical Implementation

To solve the model numerically, we take the state space $\mathcal{D}$ for f_t to be the grid $\{0, 1/(n-1), 2/(n-1), \ldots, 1\}$, with $n = 101$. We can then represent functions of f as n-dimensional vectors. Much of the algebra behind these calculations can be found in Appendixes A.4.1 and A.5.1 of the main paper.

Market equilibrium. For this step, we need to find a self-consistent variance belief, meaning the fixed point in Proposition 4. We start from a flat variance belief V_B with a small and constant value on $\mathcal{D}$ — smaller than the smaller of the two roots in footnote 7. We then iteratively apply equations (A.21)–(A.26) until the difference between consecutive variance beliefs becomes small throughout the state space. We have not proved convergence of this iterative procedure, but in all of our experiments we have found numerically that the method converges very quickly, in roughly seven iterations, provided the initial variance belief is small.

In each iteration, we need to evaluate the conditional expectation of V_B over f_{t+1} given f_t in (A.21). To reduce the impact of our discretization of the state space and approximate the results we would obtain with larger n, we use linear interpolation. In more detail, with $\lambda(\cdot)$ given, we approximate the evolution of f_t through an n-state Markov chain with a transition matrix P, where P_{ij} is the probability of transitioning from state $f_{[i]} \in \mathcal{D}$ to state $f_{[j]} \in \mathcal{D}$. With $f_t = f_{[i]}$ on the grid $\mathcal{D}$, a shock $\epsilon_{f,t+1}$ may map f_{t+1} off the grid without the projection $\Pi_{\mathcal{D}}$. Rather than round to the nearest grid point, we interpolate to reduce discretization

error.[12] If a shock $\epsilon_{f,t+1}$ puts probability p on a point $f_{t+1} = \alpha f_{[j]} + (1-\alpha) f_{[j+1]}$, $0 \le \alpha \le 1$, between two grid points, we assign probability αp to P_{ij} and $(1-\alpha)p$ to $P_{i,j+1}$. We use these interpolated probabilities in calculating the conditional expectation $\mathsf{E}[V_B(f_{t+1})|f_t = f_{[i]}]$ and similar expressions. The steady-state distribution of f_t in Figure 1.5 is calculated from the transition matrix P. We also use linear interpolation the plots in the paper.

Prices. In calculating self-consistent variance beliefs, we get the price coefficients $b(\cdot)$ and $c(\cdot)$, and the coefficients d and g are constants. It only remains to find $a(\cdot)$, for which we use (A.28) and the matrix P of transition probabilities. The vector of m-step ahead expected net profits satisfies $\mathsf{E}_t[\pi_{t+1+m}] = P^m \mathsf{E}_t[\pi_{t+1}]$. The series in (A.28) can therefore be written in vector form as

$$\begin{aligned} \frac{1}{R}\sum_{i=0}^{\infty} \frac{1}{R^i} \mathsf{E}_t[\pi_{t+i+1}] &= \frac{1}{R}\left(I + \frac{1}{R}P + \frac{1}{R^2}P^2 + \cdots\right)\mathsf{E}_t[\pi_{t+1}] \\ &= \frac{1}{R}\left(I - \frac{1}{R}P\right)^{-1} \mathsf{E}_t[\pi_{t+1}]. \end{aligned} \tag{A.58}$$

Endogenous λ. For each f we search numerically for a point $\lambda(f)$ at which the difference of expected utilities in (A.49) is zero, using linear interpolation between grid points. That difference does not depend on $a(\cdot)$, so we can find $\lambda(\cdot)$ without calculating $a(\cdot)$. If the difference in (A.49) is always positive, we set $\lambda(f) = 1$; if it is always negative, we set $\lambda(f) = 0$.

Information equilibrium. For a complete model solution we proceed as follows. We start with a small flat variance belief V_B and an arbitrary λ; e.g., $\lambda \equiv 0$. We do one update of V_B using (A.21)–(A.26) and then calculate the endogenous λ. We repeat these updates iteratively, each time updating V_B and then solving for the new λ. Once V_B has converged, we evaluate $a(\cdot)$.

[12] This mechanism differs slightly from that used in Proposition 5. Assuming the shocks $\epsilon_{f,t}$ have a density simplifies the continuity argument needed there, but for numerical calculations it is simpler to assume they have finite support.

A.9.1 Approximating Return Moments

Returns are given by $r_{t+1} = (P_{t+1} + x_D D_{t+1})/P_t - 1$, where $x_D \in \{0, 1\}$ indicates whether we want total (with dividends) or net returns. Because prices can become negative, we cannot calculate return moments in the usual sense. We instead normalize by $P_0 = a(f) + d\bar{D}$ and calculate returns moments conditional on $\mathcal{I} = \{D_t = \bar{D}, f_t = f, m_t = \theta_t = X_t = 0\}$; P_0 is the mean price given $\mathcal{I}$. The corresponding expected excess return for a given level of f is given by $12 \times \mathsf{E}_t[\pi_{t+1}]/P_0$.

Using (1.11), we make the approximation

$$r_{t+1} \approx \frac{a_{t+1} + b_{t+1} m_{t+1} + g\theta_{t+1} - c_{t+1} X_{t+1} + (x_D + d)(\bar{D} + \epsilon_{t+1})}{a(f) + d\bar{D}} - 1,$$

where all quantities are conditioned on $\mathcal{I}$. The variability in the numerator (we refer to it as ν) above will be driven by shocks m_{t+1}, θ_{t+1}, X_{t+1} and ϵ_{t+1}, but also by changes in the coefficients $a_{t+1}, b_{t+1}, c_{t+1}$ which are functions of f_{t+1}. We will use the relationship that

$$\mathsf{var}(\nu) = \mathsf{E}[\mathsf{var}(\nu | f_{t+1})] + \mathsf{var}(\mathsf{E}[\nu | f_{t+1}]). \tag{A.59}$$

The moments in (A.59) are conditional on $\mathcal{I}$, but we omit $\mathcal{I}$ to avoid clutter. We see that

$$\mathsf{var}(\nu | f') = b(f')^2 \phi f' \sigma_M^2 + c(f')^2 \sigma_X^2 + g^2 (1 - \phi) f' \sigma_M^2 + (x_D + d)^2 (1 - f') \sigma_M^2.$$

The first expectation in (A.59) is

$$\sum_{f'} ¶(f_{t+1} = f' | f_t = f) \times \mathsf{var}(\nu | f'),$$

using the transition probabilities of f_t.

We note that $\mathsf{E}[\nu | f'] = a(f') + (1 + d)\bar{D}$ and therefore

$$\mathsf{var}(\mathsf{E}[\nu | f']) = \left\{ \sum_{f'} ¶(f_{t+1} = f' | f_t = f) \times a(f')^2 \right\} - \bar{a}^2,$$

where $\bar{a} = \sum_{f'} ¶(f_{t+1} = f' | f_t = f) \times a(f')$.

The return volatility at every f, conditional on $\mathcal{I}$ is given by

$$vol(r_{t+1}) \approx \sqrt{12} \times \frac{\mathsf{sd}(\nu)}{a(f) + d\bar{D}}. \tag{A.60}$$

The return moments derived in this section entail two approximations required by the possibility of negative prices: we evaluate moments conditional on $\mathcal{I}$ and we normalize by the conditional mean price in calculating returns. We have verified via simulation (where negative prices are very rare) that the conditional return volatility above is close to the realized return volatility from simulated data.

A.10 Calibration Details for f_t Model

A.10.1 Dynamics of S&P 500 Dividends

We interpret D_t as an unobserved dividend process that would obtain if all S&P 500 companies paid a monthly dividend in each month. This D_t does not correspond to the actual, observed monthly S&P 500 dividend which represents the quarterly dividend payments of only a subset of the S&P 500 companies. To be consistent with (1.1), we model D_t as an AR(1) process where

$$\begin{aligned}
D_{t+1} &= \mu_D + \rho D_t + M_{t+1} \\
D_{t+2} &= \mu_D + \rho\mu_D + \rho^2 D_t + \rho M_{t+1} + M_{t+2} \\
D_{t+3} &= \mu_D + \rho\mu_D + \rho^2\mu_D + \rho^3 D_t + \underbrace{\rho^2 M_{t+1} + \rho M_{t+2} + M_{t+3}}_{\eta_{t+3}}.
\end{aligned}$$

If we interpret a time period as a single month then the quarterly dividend $Q_i = D_{t+1} + D_{t+2} + D_{t+3}$ can be written (expressing each D as in the last equation above) as an ARMA(1,1) process:

$$Q_{i+1} = \mu_Q + \rho_Q Q_i + u_{i+1} + \theta u_i \tag{A.61}$$

where $E[u_{i+1}u_i] = 0$ and

$$\begin{aligned}
\mu_Q &= 3\mu_D(1 + \rho + \rho^2) \\
\rho_Q &= \rho^3 \\
u_{i+1} + \theta u_i &= \eta_{t+3} + \eta_{t+2} + \eta_{t+1} \\
&= \rho^2 M_{t+1} + \rho M_{t+2} + M_{t+3} + \rho^2 M_t + \rho M_{t+1} + M_{t+2} + \rho^2 M_{t-1} + \rho M_t + M_{t+1}.
\end{aligned} \tag{A.62}$$

The parameters θ and σ_u^2 are chosen to match the variance and autocorrelation of the error term $\eta_{t+3} + \eta_{t+2} + \eta_{t+1}$. The variance of the error term yields the restriction that

$$\begin{aligned}\mathsf{var}(u_{i+1} + \theta u_i) &= \sigma_u^2 + \theta^2 \sigma_u^2 \\ &= \sigma_M^2 \left[1 + (1+\rho)^2 + (1+\rho+\rho^2)^2 + (\rho+\rho^2)^2 + \rho^4\right] \\ &= \sigma_M^2 [3 + 4\rho + 5\rho^2 + 4\rho^3 + 3\rho^4]. \end{aligned} \tag{A.63}$$

Note that the quarter i dividend contains innovations from months $t-4, t-3, t-2, t-1, t$, and therefore the correlation of the error terms from quarter $i+1$ and i yields the restriction that

$$\begin{aligned}\mathsf{cov}(u_{i+1} + \theta u_i, u_i + \theta u_{i-1}) &= \theta \sigma_u^2 \\ &= (\rho + \rho^2)\sigma_M^2 + \rho^2(1+\rho)\sigma_M^2 = \sigma_M^2(\rho^3 + 2\rho^2 + \rho). \end{aligned} \tag{A.64}$$

We use (A.62) and (A.63) to pin down σ_M. Equation (A.64) is then an overidentifying restriction on the model parameters, which together with (A.63) implies

$$\frac{\theta}{1+\theta^2} = \frac{\rho + 2\rho^2 + \rho^3}{3 + 4\rho + 5\rho^2 + 4\rho^3 + 3\rho^4} \tag{A.65}$$

A.10.1.1 Estimating AR(1) Parameters

Estimating[13] (A.61) yields four parameter estimates: $\hat{\mu}_Q, \hat{\rho}_Q, \hat{\theta}, \hat{\sigma}_u^2$. Our monthly dividend process has parameters μ_D, ρ, σ_M^2. From (A.62) and (A.63) we see that

$$\begin{aligned}\hat{\rho} &= \hat{\rho}_Q^{1/3}, \\ \hat{\sigma}_M^2 &= \frac{\hat{\sigma}_u^2 + \hat{\theta}^2 \hat{\sigma}_u^2}{3 + 4\hat{\rho} + 5\hat{\rho}^2 + 4\hat{\rho}^3 + 3\hat{\rho}^4}. \end{aligned}$$

Also the long-run level of the quarterly dividend is related to the long-run level of the monthly dividend via

$$\bar{Q} \equiv \frac{\mu_Q}{1-\rho_Q} = \frac{3\mu_D(1+\rho+\rho^2)}{1-\rho^3} = 3\frac{\mu_D}{1-\rho},$$

where the last step follows from $(1+\rho+\rho^2)(1-\rho) = 1-\rho^3$. Therefore the long-run level of the monthly dividend is equal to $\bar{D} = \bar{Q}/3$. Since we are interested in a monthly dividend

[13] We use the `arima` function in R.

Model	ρ_Q	θ	σ_u^2	$\bar{Q}$	$\bar{D}$	ρ	σ_M
Qtr (det)	0.9056	-0.2632	0.0043	1.0261	0.3420	0.9675	0.0470
Qrt (det,SA)	0.9160	-0.1808	0.0032	1.0310	0.3437	0.9712	0.0394

Table A.2: Calibration of the S&P500 dividend model.

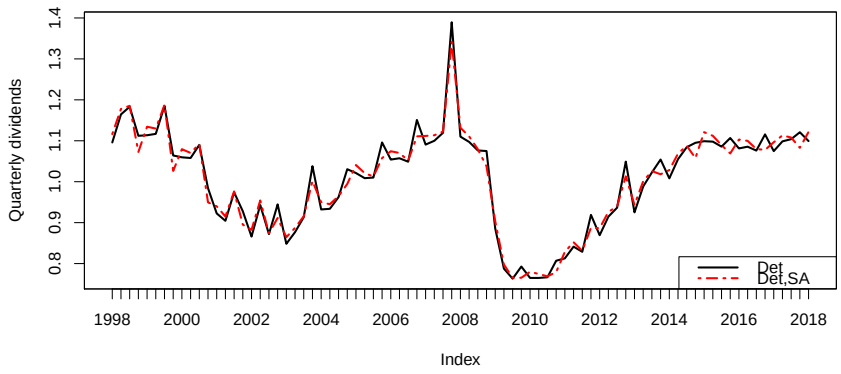


Figure A.3: Detrended and seasonally-adjusted S&P 500 quarterly dividend series.

process with a mean of 1, we set $\mu_D = 1 - \hat{\rho}$ and normalize the innovation volatility by $\hat{\sigma}_M / \hat{\bar{D}}$.

Our system of equations (involving $\mu_Q, \rho_Q, \sigma_u^2, \theta$) for μ_D, ρ, σ_D^2 is overidentified. In fact, (A.65) places an additional restriction on $\hat{\rho}$ coming from $\hat{\theta}$. In our estimates (below) we find $\hat{\theta} < 0$ suggesting the model is misspecified.

To estimate the model we use an exponentially detrended real (deflated using the PCE ex food and energy seasonally adjusted index) quarterly dividend series from January 1998 (when our daily S&P 500 dividend series starts) until March 2018. We also do an estimation with a seasonally adjusted series that involves regressing quarterly dummies out of the log dividend. Estimates from these two models are given in Table A.2. The two rows correspond to the detrended and the detrended-seasonally-adjusted series respectively. The first four parameters are from the estimated ARMA(1,1) model. σ_M is the volatility of the monthly dividend innovation, normalized for $\bar{D} = 1$.

Figure A.3 shows the S&P 500 dividend series used in this estimation.

A.10.2 Dynamics of Turnover

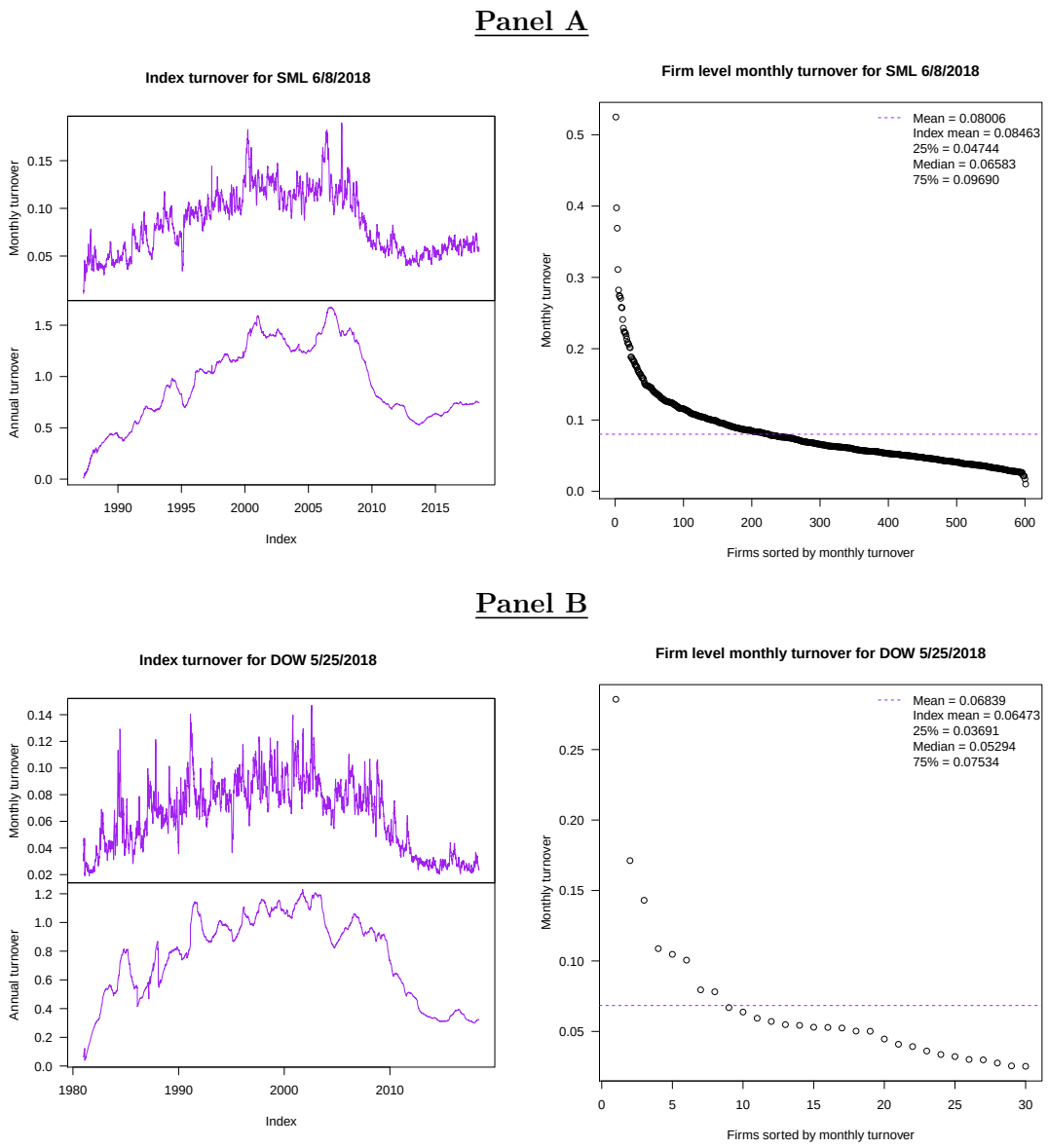


Figure A.4: Dynamics of turnovers

Turnover, defined as share trading volume divided by the number of shares outstanding, for the Dow Industrial and the S&P 600 SmallCap (SML) indexes.

Appendix B

Chapter 2 Supplemental Information

B.1 Derivation of Portfolio-implied Realized Correlation

In this section, we provide the derivation for the portfolio-implied realized correlation (2.6). Denote the two mutually exclusive sets of stocks by A and B. We construct three portfolios. The first two include stocks in A and B, respectively. Their returns can be computed by

$$R_A = \sum_{i \in A} w_i R_i \text{ and } R_B = \sum_{i \in B} w_i R_i,$$

where w_i and R_i denote the weight and return of stock i, respectively. For ease of representing the portfolio return, here we use R_i to denote the simple return instead of log return of individual stocks. In our high-frequency setting, these two are generally very close. Note that we do not require $\sum_i w_i = 1$ as long as w_i are fixed. The third portfolio S consists of stocks from both A and B, with return given by $R_S = \sum_{i \in A \cup B} w_i R_i$.

The return variances of the three portfolios can be computed by

$$\sigma_A^2 = \sum_{i,j \in A} w_i w_j \rho_{i,j} \sigma_i \sigma_j, \quad \sigma_B^2 = \sum_{i,j \in B} w_i w_j \rho_{i,j} \sigma_i \sigma_j,$$

and

$$\sigma_S^2 = \sigma_A^2 + \sigma_B^2 + 2 \sum_{i \in A, j \in B} w_i w_j \rho_{i,j} \sigma_i \sigma_j, \tag{B.1}$$

where σ_i denotes the standard deviation of the returns of stock i. The last summation term in σ_S^2 captures the covariance between the stocks from the two sets. With average correlation $\bar{\rho}_{A,B}$ defined in (2.5), the last term can be expressed by

$$\sum_{i \in A, j \in B} w_i w_j \rho_{i,j} \sigma_i \sigma_j = \bar{\rho}_{A,B} \sum_{i \in A, j \in B} w_i w_j \sigma_i \sigma_j.$$

Combining this equation with (B.1), we can solve for $\bar{\rho}_{A,B}$ from the return variances as

$$\bar{\rho}_{A,B} = \frac{\sigma_S^2 - \sigma_A^2 - \sigma_B^2}{2 \sum_{i \in A, j \in B} w_i w_j \sigma_i \sigma_j}. \tag{B.2}$$

To estimate the realized version of $\bar{\rho}_{A,B}$ with high-frequency data, we plug the realized variances into (B.2) to obtain

$$\mathrm{RCorr}_{A,B}^{(J,K)} = \frac{\mathrm{RV}_{S}^{(J,K)} - \mathrm{RV}_{A}^{(J,K)} - \mathrm{RV}_{B}^{(J,K)}}{2\sum_{i\in A, j\in B} w_i w_j \sqrt{\mathrm{RV}_{i}^{(J,K)}} \cdot \sqrt{\mathrm{RV}_{j}^{(J,K)}}}. \tag{B.3}$$

This leads to the portfolio-implied realized correlation estimator in (2.6).

B.2 Supplementary Tables

Table B.1: Average passive ownership for the highest and lowest passive owenership bins

Bins	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Low	0.04	0.04	0.03	0.04	0.05	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05
High	0.09	0.09	0.09	0.10	0.13	0.14	0.15	0.15	0.17	0.19	0.20	0.22	0.23	0.26	0.26

Table B.2: Number of stocks in the S&P 500 Index in each entire year

2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
475	478	458	458	461	469	483	474	481	476	483	472	471	472	478

Table B.3: Average daily correlation of stock pairs in each daily correlation bin

Years	1	2	3	4	5	6	7	8	9	10	11	12
2003	-0.01	0.09	0.15	0.20	0.24	0.28	0.31	0.34	0.38	0.42	0.47	0.57
2004	-0.02	0.05	0.09	0.13	0.17	0.20	0.22	0.25	0.28	0.31	0.35	0.45
2005	-0.02	0.04	0.09	0.12	0.16	0.19	0.22	0.24	0.27	0.30	0.34	0.45
2006	-0.04	0.02	0.06	0.10	0.13	0.16	0.19	0.22	0.25	0.28	0.32	0.42
2007	0.06	0.14	0.19	0.23	0.27	0.30	0.33	0.36	0.39	0.42	0.47	0.57
2008	0.18	0.28	0.34	0.40	0.44	0.48	0.51	0.53	0.56	0.59	0.63	0.70
2009	0.04	0.16	0.23	0.29	0.34	0.38	0.42	0.46	0.49	0.53	0.58	0.67
2010	0.08	0.19	0.27	0.33	0.38	0.42	0.45	0.49	0.52	0.56	0.60	0.68
2011	0.23	0.33	0.39	0.44	0.49	0.53	0.56	0.59	0.62	0.66	0.69	0.76
2012	0.01	0.09	0.14	0.19	0.23	0.27	0.30	0.34	0.37	0.41	0.47	0.56
2013	0.01	0.09	0.14	0.19	0.23	0.27	0.30	0.33	0.36	0.40	0.44	0.54
2014	-0.02	0.05	0.11	0.16	0.20	0.24	0.28	0.32	0.35	0.39	0.45	0.56
2015	-0.01	0.11	0.18	0.23	0.28	0.32	0.36	0.39	0.43	0.47	0.52	0.63
2016	-0.14	-0.01	0.07	0.13	0.19	0.23	0.27	0.31	0.35	0.40	0.46	0.58
2017	-0.20	-0.09	-0.04	0.00	0.04	0.07	0.10	0.13	0.16	0.20	0.25	0.39
2018	-0.06	0.03	0.10	0.17	0.23	0.27	0.31	0.35	0.39	0.43	0.48	0.60

Table B.4: GICS codes and sector names

GICS code	Sector	Abbreviation
10	Energy	Eng
15	Materials	Mat
20	Industrials	Ind
25	Consumer discretionary	Disc
30	Consumer staples	Stap
35	Healthcare	Heal
40	Financial	Fin
45	Information technology	Tech
50	Communication services	Comm
55	Utilities	Util
60	Real estate	Estat

Table B.5: Number of stocks in different sectors

Years	Comm	Disc	Stap	Eng	Fin	Heal	Ind	Tech	Mat	Util	Estat
2003	23	71	36	25	78	50	55	78	35	33	3
2004	21	70	35	25	71	51	52	78	34	33	5
2005	18	67	34	26	76	58	53	77	32	32	5
2006	19	63	35	26	72	56	51	68	30	31	7
2007	23	64	38	30	73	52	50	61	28	29	10
2008	18	63	37	34	68	52	53	64	27	31	14
2009	19	61	40	37	64	54	54	67	28	31	14
2010	19	64	41	37	65	52	56	70	31	31	17
2011	18	65	39	40	65	51	57	62	30	31	16
2012	21	66	38	40	66	53	57	63	30	30	17
2013	18	65	39	44	62	52	59	58	31	30	18
2014	20	69	37	42	62	54	64	57	28	30	20
2015	20	69	35	39	62	51	63	56	26	28	21
2016	21	67	35	35	58	56	63	58	24	26	24
2017	19	63	33	29	63	57	67	56	23	28	29
2018	20	63	32	28	64	58	66	60	23	28	31

Table B.6: Selected sector pairs with high and low daily correlations in each year

Years	Lowest	2rd lowest	3rd lowest	3rd highest	2rd highest	Highest
2004	(Eng, Stap)	(Eng, Tech)	(Eng, Heal)	(Ind, Fin)	(Fin, Mat)	(Mat, Ind)
2005	(Heal, Eng)	(Eng, Stap)	(Eng, Tech)	(Ind, Mat)	(Fin, Util)	(Eng, Util)
2006	(Stap, Eng)	(Heal, Eng)	(Disc, Eng)	(Mat, Eng)	(Fin, Ind)	(Ind, Mat)
2007	(Heal, Tech)	(Heal, Disc)	(Heal, Eng)	(Eng, Mat)	(Fin, Mat)	(Mat, Ind)
2008	(Eng, Disc)	(Eng, Fin)	(Fin, Util)	(Util, Eng)	(Eng, Mat)	(Mat, Ind)
2009	(Disc, Heal)	(Stap, Heal)	(Tech, Heal)	(Ind, Mat)	(Eng, Ind)	(Eng, Mat)
2010	(Stap, Disc)	(Tech, Stap)	(Heal, Stap)	(Ind, Mat)	(Ind, Fin)	(Eng, Ind)
2011	(Tech, Stap)	(Disc, Stap)	(Comm, Stap)	(Mat, Fin)	(Ind, Mat)	(Fin, Ind)
2012	(Disc, Util)	(Util, Tech)	(Util, Comm)	(Mat, Ind)	(Fin, Mat)	(Fin, Ind)
2013	(Tech, Util)	(Tech, Stap)	(Eng, Heal)	(Mat, Ind)	(Fin, Mat)	(Ind, Fin)
2014	(Util, Disc)	(Util, Tech)	(Comm, Util)	(Fin, Mat)	(Tech, Fin)	(Fin, Ind)
2015	(Util, Eng)	(Real, Eng)	(Disc, Util)	(Fin, Tech)	(Util, Real)	(Fin, Ind)
2016	(Fin, Util)	(Disc, Util)	(Util, Eng)	(Fin, Tech)	(Mat, Fin)	(Fin, Ind)
2017	(Fin, Util)	(Eng, Util)	(Util, Comm)	(Mat, Ind)	(Mat, Fin)	(Ind, Fin)
2018	(Disc, Util)	(Util, Eng)	(Fin, Util)	(Fin, Tech)	(Mat, Ind)	(Fin, Ind)

Table B.7: Average daily beta of stocks in each daily beta bin

Years	1	2	3	4	5	6	7	8	9	10	11
2003	0.38	0.52	0.63	0.77	0.86	0.93	1.02	1.13	1.25	1.45	1.88
2004	0.42	0.53	0.64	0.73	0.82	0.90	1.00	1.10	1.25	1.50	2.03
2005	0.50	0.63	0.73	0.82	0.91	0.99	1.06	1.12	1.20	1.33	1.62
2006	0.38	0.51	0.60	0.71	0.80	0.89	1.00	1.11	1.25	1.46	1.94
2007	0.48	0.61	0.69	0.79	0.87	0.94	1.01	1.09	1.21	1.36	1.71
2008	0.50	0.62	0.71	0.80	0.87	0.95	1.03	1.13	1.29	1.48	1.87
2009	0.32	0.47	0.61	0.75	0.89	1.03	1.21	1.36	1.60	1.96	2.74
2010	0.46	0.56	0.69	0.81	0.93	1.03	1.14	1.26	1.38	1.53	1.82
2011	0.46	0.58	0.73	0.86	0.97	1.05	1.15	1.25	1.37	1.53	1.80
2012	0.36	0.48	0.62	0.78	0.90	1.01	1.11	1.23	1.38	1.54	1.88
2013	0.58	0.72	0.78	0.84	0.92	0.99	1.07	1.15	1.25	1.37	1.60
2014	0.38	0.55	0.68	0.80	0.90	0.99	1.07	1.14	1.22	1.34	1.61
2015	0.53	0.67	0.75	0.85	0.91	0.96	1.01	1.05	1.12	1.20	1.38
2016	0.30	0.51	0.67	0.82	0.91	1.01	1.14	1.24	1.35	1.53	1.96
2017	0.06	0.27	0.50	0.66	0.77	0.90	1.03	1.17	1.29	1.46	1.80
2018	0.19	0.40	0.57	0.72	0.82	0.89	0.94	1.01	1.08	1.19	1.44

Appendix C

Chapter 3 Supplemental Information

C.1 Reduced-form Evidence for Discounting Behavior in ICU Admissions

In this section, we conduct reduced-form regressions to analyze the main determinants of the system's ICU admission decisions. We show the ICU admission decisions are indeed impacted by the system state, i.e., the number of patients in ED and ICU. The results from the reduced-form regression provide consistent evidence for the discounting behaviors captured by our structural model.

C.1.1 Model

We apply a multinomial logit model to estimate the ICU admission decisions. In each period, the hospital chooses one of the three options for each patient: admit the patient into the ICU; admit the patient to non-ICU units like medical/surgical ward; or make the patient wait in the ED. Note that the decision to keep the patient waiting in the ED is often necessitated by system-level considerations; in the absence of capacity constraints, a patient would not be kept waiting for admission to the hospital. We include patient characteristics, system state variables, and seasonality effects as the potential determinants of these decisions.

As with the structural model, we set a period to be two hours in the logit model. At the start of each period, we construct system "snapshots" which includes detailed information (e.g., gender, age, and severity scores) for each patient in the ED, as well as the total number of patients boarding in the ED and the number of patients in the ICU.

The system's decision on patient i in period t is determined by two types of variables

in the model: patient i's characteristics $\mathbf{X}_i$ and system state variables in period t, $\mathbf{S}_t$. $\mathbf{X}_i$ includes patient i's gender, age, as well as three severity scores–i.e., LAPS2, COPS2, and CHMR. To account for potential differences between hospitals, we also include a categorical variable to represent the hospital in which patient i is treated. In summary, we have

$$\mathbf{X}_i = \{\text{Gender}_i, \text{Age}_i, \text{LAPS2}_i, \text{COPS2}_i, \text{CHMR}_i, \text{Hosp}_i\}.$$

The system state vector $\mathbf{S}_t$ includes the following variables

$$\mathbf{S}_t = \{\text{ICUOccu}_t, \text{EDNum}_t, \text{DepPre}_t, \text{AvgLAPS2}_t, \text{DayOfWeek}_t, \text{HourOfDay}_t, \text{MonthDummy}_t\},$$

where ICUOccu_t denotes the current ICU occupancy level. As the ICU sizes vary dramatically across the hospitals, we use the ICU percentile rank to measure occupancy. EDNum_t denotes the number of current ED patients for whom a decision to admit them to the hospital has been made but for which a decision about when and where to admit them needs to be made, and DepPre_t denotes the number of patients who left the ICU (i.e. discharged from the ICU or died) in the previous period. AvgLAPS2_t denotes the average severity level measured by the LAPS2 score of the current ICU patients in period t. Finally, the categorical variables DayOfWeek_t, HourOfDay_t, and MonthDummy_t capture the potential seasonality and time trend in the decisions: DayOfWeek_t and HourOfDay_t denote the day of week and hour of day respectively; MonthDummy_t is the dummy variable representing the month in the sample (total 23 months).

For patient i who is in the ED at the start of period t, we estimate the system's decision d_{it} using a multinomial logit model:

$$\begin{aligned}\ln\left[\frac{\Pr(d_{it}|\mathbf{X}_i,\mathbf{S}_t)}{\Pr(\text{nonICU}_{it}|\mathbf{X}_i,\mathbf{S}_t)}\right] &= \gamma_{0,d} + \gamma_{G,d}\text{Gender}_i + \gamma_{A,d}\text{Age}_i + \gamma_{L,d}\text{LAPS2}_i + \gamma_{CP,d}\text{COPS2}_i \\ &+ \gamma_{CH,d}\text{CHMR}_i + \gamma_{H,d}\text{Hosp}_i + \gamma_{ICU,d}\text{ICUOccu}_t + \gamma_{ED,d}\text{EDNum}_t \\ &+ \gamma_{Dep,d}\text{DepPre}_t + \gamma_{AL,d}\text{AvgLAPS2}_t + \gamma_{DW,d}\text{DayOfWeek}_t \\ &+ \gamma_{HD,d}\text{HourOfDay}_t + \gamma_{Mon,d}\text{MonthDummy}_t + \epsilon_{it}, \qquad \text{(C.1)}\end{aligned}$$

where $d_{it} \in \{Wait_{it}, Adm_{it}\}$. $\Pr(d_{it}|\mathbf{X}_i\mathbf{S}_t)$ is the probability of d_{it} conditional on $(\mathbf{X}_i, \mathbf{S}_t)$. $\Pr(\text{nonICU}_{it}|\mathbf{X}_i\mathbf{S}_t)$ is the probability of admitting patient i to non-ICU units in period t conditional on $(\mathbf{X}_i, \mathbf{S}_t)$. We use the non-ICU admission decision as the base case, and estimate

the probabilities of the ICU admission (d = ICUAdm) and waiting (d = Wait) decisions relative to the non-ICU admission decision respectively. To account for the heteroskedasticity, we cluster standard errors by hospitals in the regression.

We use the McFadden's pseudo R-squared to measure the goodness of fit of the multinomial logit model (C.1). It is defined as:

$$\text{Pseudo } R^2 = 1 - \frac{\ln l^{mod}}{\ln l^{null}}, \tag{C.2}$$

where l^{mod} is the likelihood from the estimated model, and l^{null} is the likelihood from the "null" model that only includes the intercept and categorical variable for each hospital, i.e.,

$$\ln\left[\frac{\Pr(d_{it}|\mathbf{X}_i, \mathbf{S}_t)}{\Pr(\text{nonICU}_{it}|\mathbf{X}_i, \mathbf{S}_t)}\right] = \gamma_{0,d} + \gamma_{H,d}\text{Hosp}_i + \epsilon_{it},$$

where $d_{it} \in \{\text{Wait}_{it}, \text{ICUAdm}_{it}\}$; Hosp_i is the hospital categorical variable.

We first estimate the model by combining the patient data from all hospitals. Then, considering the heterogeneity across hospitals, we also estimate the model for individual hospitals separately after dropping the categorical variable term $\gamma_{H,d}\text{Hosp}_i$ in (C.1). We discuss select coefficients from the multinomial logistic regression in next section; other regression results are available upon request.

C.1.2 Results

In this section, we report the estimation results for the multinomial logit model (C.1). We report the estimated coefficients for three main variables: LAPS2_i, ICUOccu_t, and EDNum_t. Table C.1 below shows the estimation results for model (C.1) with all hospitals combined. Note that in the estimation, each hospitalization may be counted multiple times if the patient waits in the ED for more than one period. Thus, the sample size (183,691) is larger than the number of total hospitalizations (164,167).

The results in Table C.1 show that all the coefficients except γ_{ED} for ICU admission decision are statistically significant at 0.1% level and have the expected sign. The γ_{ED} for ICU admission decision is significant at 5% level with the expected sign. In particular, higher LAPS2 score increases the probability of admission to ICU relative to other units, as these patients are more critically ill. For these severe patients, the system may also need

Table C.1: Estimation results for Multinomial-Logistic Regression (C.1), N = 183,691, R-squared = 0.16

	LAPS2_i	ICUOccu_t	EDNum_t
	γ_L	γ_{ICU}	γ_{ED}
Waiting	0.008^{***}	1.178^{***}	0.241^{***}
	(0.000)	(0.031)	(0.006)
Admission	0.028^{***}	-0.396^{***}	-0.014^{*}
	(0.000)	(0.030)	(0.006)

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Select coefficients (LAPS2 score, ICU occupancy level, and number of ED boarding patients) from (C.1) for waiting and admission decisions respectively.

to keep them waiting in the ED when the ICU is congested. Thus, higher LAPS2 score also increases the probability of waiting, although the magnitude of the impact is smaller. More importantly, the estimates of γ_{ICU} and γ_{ED} suggest that, even after controlling for patient characteristics and fixed effects, a busier system state (more congested ICU or more congested ED) decreases the probability of ICU admission and increases the probability of waiting, respectively. Such evidence suggests that the system indeed internalizes the intertemporal externalities on the ICU admission decisions by adjusting their behaviors according to the current system state: When ICU or ED is more congested, they are more likely to delay the admission of current patients to save ICU beds for future patients.

We also report the McFadden's pseudo R^2 in Table C.1. While 0.16 is relatively low, we emphasize that the pseudo R^2 is computed using the null model where hospital fixed effects are included. Moreover, it is consistent with the magnitude seen for models of operational decisions in healthcare systems (see, e.g., Kim et al. (2015), Chan et al. (2016), and Song et al. (2019) among others).

In Table C.2, we report the coefficients for the three main covariates when we estimate model (C.1) for each hospital separately. In the last column, we also provide the McFadden's R^2 of the model for each hospital. The results are qualitatively similar to that for all hospitals combined in Table C.1. Most of the coefficients have the expected signs for individual hospitals, although some are not statistically significant as the sample size of each individual hospital is much smaller than all hospitals combined. Full estimation results are available from authors upon request.

Table C.2: Estimation results for multinomial logistic regression (C.1): Individual hospitals

		for Waiting decision			for ICU Admission decision			
Hosp	Size	LAPS2 γ_L	ICUOccu γ_{ICU}	EDNum γ_{ED}	LAPS2 γ_L	ICUOccu γ_{ICU}	EDNum γ_{ED}	R^2
1	13,964	0.006^{***}	0.699^{***}	0.230^{***}	0.022^{***}	-0.372^{**}	-0.019	0.19
		(0.001)	(0.113)	(0.018)	(0.001)	(0.114)	(0.020)	
2	12,871	0.007^{***}	1.205^{***}	0.170^{***}	0.026^{***}	-0.642^{***}	-0.007	0.14
		(0.001)	(0.093)	(0.015)	(0.001)	(0.116)	(0.021)	
3	8,391	0.005^{*}	0.605^{**}	0.465^{***}	0.031^{***}	-0.191	0.022	0.15
		(0.002)	(0.228)	(0.053)	(0.002)	(0.193)	(0.048)	
4	16,162	0.013^{***}	1.830^{***}	0.209^{***}	0.031^{***}	-0.344^{*}	0.022	0.22
		(0.001)	(0.149)	(0.019)	(0.001)	(0.140)	(0.023)	
5	5,499	0.013^{***}	1.475^{***}	0.404^{***}	0.031^{***}	-0.782^{***}	0.002	0.16
		(0.002)	(0.222)	(0.051)	(0.002)	(0.183)	(0.051)	
6	11,698	0.012^{***}	1.591^{***}	0.280^{***}	0.029^{***}	-0.270	-0.015	0.17
		(0.001)	(0.162)	(0.027)	(0.001)	(0.146)	(0.028)	
7	5,200	0.010^{***}	0.892^{***}	0.262^{***}	0.020^{***}	-1.489^{***}	0.092	0.14
		(0.002)	(0.211)	(0.057)	(0.002)	(0.230)	(0.069)	
8	9,382	0.011^{***}	2.227^{***}	0.203^{***}	0.024^{***}	-0.591^{***}	-0.012	0.14
		(0.001)	(0.153)	(0.029)	(0.001)	(0.117)	(0.027)	
9	14,774	0.007^{***}	1.454^{***}	0.220^{***}	0.029^{***}	-0.419^{***}	-0.010	0.20
		(0.001)	(0.104)	(0.014)	(0.001)	(0.126)	(0.020)	
10	6,032	0.009^{**}	1.560^{***}	0.234^{**}	0.032^{***}	-0.434^{*}	-0.014	0.16
		(0.003)	(0.333)	(0.073)	(0.002)	(0.201)	(0.055)	
11	3,334	0.008^{***}	0.719^{***}	0.112	0.029^{***}	-0.969^{***}	-0.095	0.15
		(0.002)	(0.184)	(0.070)	(0.002)	(0.216)	(0.093)	
12	8,413	0.012^{***}	1.161^{***}	0.168^{***}	0.034^{***}	-1.025^{***}	-0.023	0.14
		(0.001)	(0.146)	(0.020)	(0.002)	(0.219)	(0.033)	
13	8,620	0.008^{***}	1.336^{***}	0.331^{***}	0.032^{***}	-0.214	0.011	0.18
		(0.002)	(0.229)	(0.042)	(0.002)	(0.184)	(0.040)	
14	14,209	0.011^{***}	1.523^{***}	0.140^{***}	0.022^{***}	-0.269^{*}	0.025	0.14
		(0.001)	(0.129)	(0.021)	(0.001)	(0.132)	(0.024)	
15	7,073	0.006	0.509	0.181^{*}	0.031^{***}	-0.083	0.024	0.14
		(0.003)	(0.365)	(0.083)	(0.002)	(0.182)	(0.045)	
16	7,517	0.008^{***}	0.893^{***}	0.179^{**}	0.032^{***}	-0.393^{*}	-0.035	0.20
		(0.002)	(0.229)	(0.057)	(0.002)	(0.171)	(0.048)	
17	3,896	0.014^{***}	0.736^{**}	0.404^{***}	0.027^{***}	-1.444^{***}	-0.043	0.21
		(0.002)	(0.256)	(0.071)	(0.002)	(0.237)	(0.083)	
18	7,550	0.004^{*}	0.266	0.272^{***}	0.029^{***}	0.001	0.046	0.17
		(0.002)	(0.238)	(0.043)	(0.002)	(0.218)	(0.045)	
19	7,732	0.008^{***}	1.015^{***}	0.271^{***}	0.028^{***}	-0.341^{*}	-0.071	0.16
		(0.002)	(0.289)	(0.054)	(0.001)	(0.172)	(0.036)	
20	5,678	0.001	1.208^{***}	0.345^{***}	0.030^{***}	-0.744^{***}	-0.003	0.24
		(0.002)	(0.215)	(0.053)	(0.002)	(0.208)	(0.062)	
21	3,154	0.007^{***}	0.999^{***}	0.142^{**}	0.032^{***}	-1.091^{***}	0.091	0.16
		(0.002)	(0.175)	(0.044)	(0.003)	(0.253)	(0.077)	
22	2,542	0.012^{***}	0.006	0.205^{***}	0.027^{***}	-0.751^{**}	-0.104	0.14
		(0.002)	(0.256)	(0.061)	(0.003)	(0.276)	(0.078)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Select coefficients (LAPS2 score, ICU occupancy level, and number of ED boarding patients) from (C.1) for waiting and admission decisions respectively.

C.2 Formulae and Supplementary Tables

C.2.1 Explicit expressions for state transition probability $g(s'|s)$

In this section, we provide explicit expressions for the function $g(s'|s)$ used in Proposition 1, which is the transition probability from the intermediate state s (after action is taken) to the state s' at the start of the next period. Denote $s = (\tilde{n}^E_{l,t}, \tilde{n}^E_{h,t}, \tilde{n}^I_t)$ and $s' = (n^E_{l,t+1}, n^E_{h,t+1}, n^I_{t+1})$, which are the state after action is taken in period t and the state at the beginning of period $t+1$ (before action) respectively. By (3.7), the transition from s to s' includes the ED arrivals from the two classes of patients, external arrivals to the ICU, and departures from the ICU. As the ED arrivals are independent of the ICU external arrivals and departures, we have

$$g(s'|s) = g_{Q,l}\left(n^E_{l,t+1}|\tilde{n}^E_{l,t}\right) g_{Q,h}\left(n^E_{h,t+1}|\tilde{n}^E_{h,t}\right) g_I\left(n^I_{t+1}|\tilde{n}^I_t\right), \tag{C.3}$$

where $g_{Q,l}$ and $g_{Q,h}$ denote the transition probabilities for the numbers of ED patients from low and high classes respectively, and g_I denotes the transition probability for the number of ICU patients.

For the ED transition probabilities $g_{Q,l}$ and $g_{Q,h}$, we only need to consider the new arrivals for class $i \in \{l, h\}$, which follow truncated Poisson distributions with rate $\lambda_{Q,i}$ and truncation by M_{A_i} from above. Additionally accounting for the ED capacity constraint, the number of ED arrivals is capped by $\max\{M_{A_i}, Q_i - \tilde{n}^E_{i,t}\}$, i.e., the bigger of the maximum arrival per period and the remaining ED capacity. Thus, the transition probability can be computed as

$$g_{Q,i}(m|n) = \begin{cases} \left(\lambda_{Q,i}\right)^{m-n} \exp(-\lambda_{Q,i})/(m-n)! & \text{if } \ n \le m < n + \max\{M_{A_i}, Q_i - n\} \\ \\ \sum_{j=\bar{A}_i}^{+\infty} \left(\lambda_{Q,i}\right)^j \exp(-\lambda_{Q,i})/j! & \text{if } m = n + \max\{M_{A_i}, Q_i - n\} \end{cases} \tag{C.4}$$

and $g_{Q,i}(m|n) = 0$ elsewhere. The second line in (C.4) considers the case where the upper bound of the number of arrivals, $\max\{M_{A_i}, Q_i - \tilde{n}^E_{i,t}\}$, is reached.

For the ICU transition probability g_I, we need to consider both external arrivals and departures. The number of external arrivals E_t follows a Poisson distribution with rate λ_E,

and is capped by the remaining ICU capacity $B-\tilde{n}_t^I$. With E_t external arrivals in the period, the ICU would have total $\tilde{n}_t^I + E_t$ patients. Then, the number of departures, D_t, follows a Binomial-$(\tilde{n}_t^I + E_t, \mu_I)$ distribution. We have following relationship:

$$n_{t+1}^I + D_t = \tilde{n}_t^I + E_t.$$

Thus, given the number of external arrivals, E_t, the number of departures follows by $D_t = \tilde{n}_t^I + E_t - n_{t+1}^I$. We note that E_t can be greater than $\max\{n_{t+1}^I - \tilde{n}_t^I, 0\}$. Summing up the probability of all possible choices of the Poisson-distributed E_t (and the Binomial-distributed D_t accordingly), we can derive transition probability $g_I(m|n)$ as

$$\begin{aligned} g_I(m|n) \quad = \quad & \sum_{j=\max\{m-n,0\}}^{B-n-1} \lambda_I^j \frac{\exp(-\lambda_I)}{j!} \frac{(n+j)!}{(n+j-m)!m!} \mu_I^{n+j-m} (1-\mu_I)^m \\ & + \left(\sum_{j=B-n}^{+\infty} \lambda_I^k \frac{\exp(-\lambda_I)}{j!} \right) \frac{B!}{(B-m)!m!} \mu_I^{B-m} (1-\mu_I)^m, \text{ for } 0 \le m \le B, \quad \text{(C.5)} \end{aligned}$$

where the second line considers the case that E_t is truncated by the remaining capacity $B-n$. Under this case, the ICU reaches the full capacity and D_t is a Binomial-(B, μ_I) variable. Combining (C.4) and (C.5), we obtain the explicit expression for state transition probability $g(s'|s)$ by (C.3).

C.2.2 Model Assumptions for Identification

This section documents the assumptions made in Komarova et al. (2018), which are also satisfied by our model. In their paper, x and a denote the system state and action, respectively; ε represents the random perturbation in utility.

Assumption 2. *(i) (Additive Separability) For all a, x, ε, the per-period utility follows:*

$$u(a, x, \varepsilon) = \pi(a, x) + \varepsilon(a).$$

(ii) (Conditional Independence) The transition distribution of the states has the following factorization for all x', ε', x, ε, a:

$$P(x', \varepsilon'|x, \varepsilon, a) = Q(\varepsilon')G(x'|x, a),$$

where Q is the cumulative distribution function of ε and G denotes the transition law of x_{t+1} conditioning on x_t and a_t. Furthermore, ε_t has finite first moments, and a positive, continuous, and bounded density.

(iii) (Finite Observed State) $X = \{1, \ldots, K\}$.

Assumption 3 (Linear-in-Parameter)**.** *For all a, x:*

$$\pi(a, x; \theta) = \pi_0(a, x) + \theta^\top \pi_1(a, x),$$

where π_0 is a known real value function, π_1 is a known p-dimensional vector value function and θ is the p-dimensional unknown parameter.

In our setting, we have per-period cost given by (3.2) as

$$c(s_t, d_t) = c_{r,l} r_{l,t} + c_{w,l} \left(n^E_{l,t} - a_{l,t} - r_{l,t} \right) + c_{r,h} r_{h,t} + c_{w,h} \left(n^E_{h,t} - a_{h,t} - r_{h,t} \right).$$

The deterministic part of the per-period utility follows by $u(s_t, d_t) = -c(s_t, d_t)$. Thus, it is indeed linear in parameters $\theta = \{c_{r,l}, c_{w,l}, c_{r,h}, c_{w,h}\}$. In our model, the functions π_0 and π_1 specify to

$$\pi_0(d_t, s_t) \equiv 0,$$

and

$$\pi_1(d_t, s_t) = - \begin{bmatrix} r_{l,t} \\ n^E_{l,t} - a_{l,t} - r_{l,t} \\ r_{h,t} \\ n^E_{h,t} - a_{h,t} - r_{h,t} \end{bmatrix}.$$

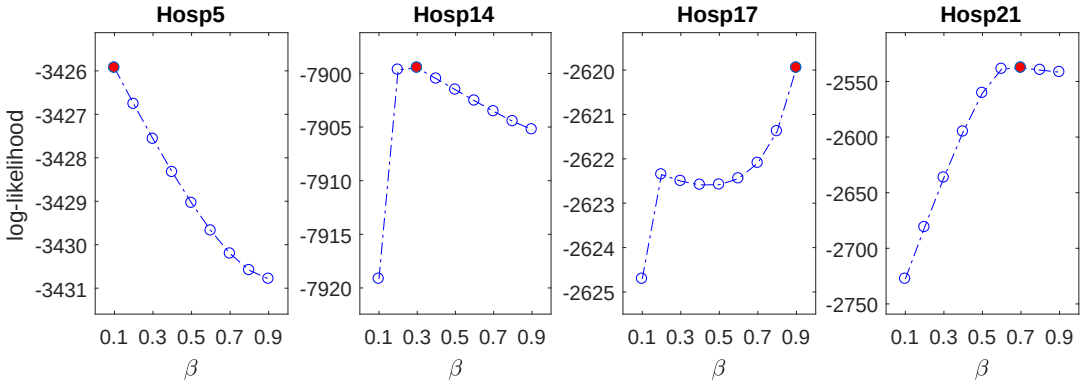
where $s_t = \left(n^E_{l,t}, n^E_{h,t}, n^I_t \right)$ and $d_t = \{a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}\}$.

C.2.3 Supplementary Tables

Table C.3: Sample Size: Numbers of observed days and hospitalizations for each hospital

Hosp	Num. of days	Num. of hospitalizations
1	667	11,676
2	500	9,902
3	667	8,039
4	667	14,595
5	576	5,082
6	667	10,577
7	578	4,915
8	653	8,400
9	514	12,355
10	609	5,978
11	667	2,655
12	388	6,751
13	667	8,061
14	667	12,841
15	575	7,208
16	667	7,190
17	548	3,511
18	667	7,702
19	547	7,476
20	666	5,096
21	333	2,109
22	333	2,048

Figure C.1: Examples of log-likelihood versus discount factor for a subset of hospital



Note: The estimated likelihood at $\beta = 0.1, 0.2, \ldots, 0.9$ for four different hospitals (5, 14, 17, and 21). The β with the best log-likelihood is highlighted in red.

Table C.4: Drop in ICU admission probability as ICU gets congested: e.g. increasing from 50% occupancy to having only 1 available bed

Hosp	$\hat{\beta}$	AdmDrop	Rel. AdmDrop
1	0.3	0.047	0.142
2	0.5	0.047	0.170
3	0.1	0.013	0.060
4	0.4	0.021	0.082
5	0.1	0.036	0.099
6	0.2	0.031	0.099
7	0.9	0.017	0.105
8	0.1	0.047	0.139
9	0.5	0.036	0.140
10	0.1	0.018	0.068
11	0.9	0.055	0.150
12	0.6	0.025	0.104
13	0.1	0.021	0.075
14	0.3	0.029	0.101
15	0.1	0.024	0.078
16	0.1	0.018	0.072
17	0.9	0.047	0.171
18	0.4	0.018	0.074
19	0.1	0.035	0.100
20	0.3	0.021	0.081
21	0.7	0.030	0.106
22	0.3	0.047	0.127

The estimated discount factor and drop in admission probability: AdmDrop and Rel.AdmDrop denote the absolute and relative admission probability drop given in (3.19) and (3.20), respectively.

Table C.5: Counterfactual estimates of impact when λ_E decreases by 10%

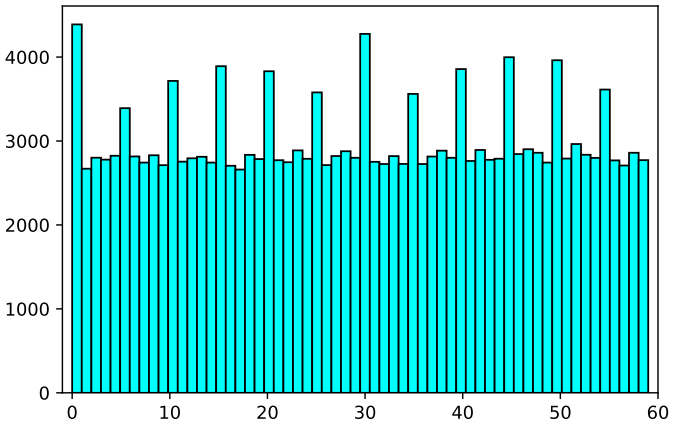
Hosp	$\hat{\beta}$	$\Delta\mathsf{Pr}$(HighCgstn) (in % points)	Rel $\Delta\mathsf{Pr}$(HighCgstn) (in %)	$\Delta\mathsf{Pr}$(Balk) (in % points)	Rel $\Delta\mathsf{Pr}$(Balk) (in %)	ΔDays HighCgstn (in # days)	ΔPats HighCgstn (in # patients)	ΔPats Balk (in # patients)
1	0.3	1.35	0.28	0.47	0.30	4.93	41.11	6.37
2	0.5	2.16	0.29	0.76	0.29	7.90	61.43	11.13
3	0.1	0.94	0.31	0.29	0.34	3.44	13.51	1.54
4	0.4	1.95	0.53	0.70	0.58	7.12	79.53	17.11
5	0.1	2.21	0.16	0.82	0.17	8.06	29.28	5.78
6	0.2	0.79	0.44	0.27	0.50	2.87	25.05	3.55
7	0.9	4.41	0.24	1.95	0.29	16.08	58.18	20.06
8	0.1	1.80	0.23	0.65	0.24	6.57	39.08	5.98
9	0.5	2.75	0.37	0.96	0.39	10.02	91.62	17.29
10	0.1	1.56	0.36	0.56	0.44	5.69	35.84	6.31
11	0.9	2.84	0.19	1.15	0.23	10.37	21.77	3.75
12	0.6	2.45	0.45	0.96	0.50	8.95	63.72	13.84
13	0.1	0.39	0.31	0.09	0.25	1.41	7.02	0.54
14	0.3	1.46	0.54	0.50	0.56	5.33	58.57	11.58
15	0.1	0.14	0.36	0.00	0.01	0.51	2.83	0.05
16	0.1	0.52	0.42	0.19	0.52	1.91	9.15	0.97
17	0.9	2.60	0.18	1.15	0.23	9.48	22.27	4.50
18	0.4	0.31	0.76	0.09	0.78	1.12	14.18	2.23
19	0.1	0.01	0.27	0.00	0.08	0.05	0.48	0.03
20	0.3	0.10	0.25	0.04	0.39	0.37	1.53	0.13
21	0.7	2.66	0.24	1.20	0.29	9.72	34.54	9.17
22	0.3	1.98	0.34	0.85	0.41	7.22	36.97	7.73

Note: Counterfactual simulation result from decreasing external arrival rates λ_E by 10%: The third and fourth columns report the absolute and relative drops in high congestion probability in (3.21); the fifth and sixth columns report the absolute and relative drops in balking probability in (3.22). The last three columns report the equivalent numbers of days and patients affected in a year.

C.3 Choice of interval length in structural model

In this section, we briefly discuss our choice of time window in the structural model, i.e., 2hr in our study. In reality, admission decisions cannot occur instantaneously. Figure C.2 shows a histogram of the minutes for the admission actions (to ICU and non-ICU units) with data from all hospitals. From the figure, we observe that there are rounding of the admission times, as shown by the peaks at 5 minutes, 15 minutes, and at the half hour marks. Thus, there is already some quantization of time.

Figure C.2: Histogram of minutes for admission actions



To choose a reasonable time window, we report in Table C.6 the proportion of intervals with admission actions (i.e., non-waiting decisions) when we choose different lengths for the time window. The proportions are computed using all hospitals in our sample. We see from Table C.6 that if we choose very small time windows (e.g. less than 30 minutes), very few admission actions take place in each time window. This would render many state-action pairs useless for the identification purpose and lead to noisy estimates. Thus, it is more reasonable to use one-hour or two-hour time windows.

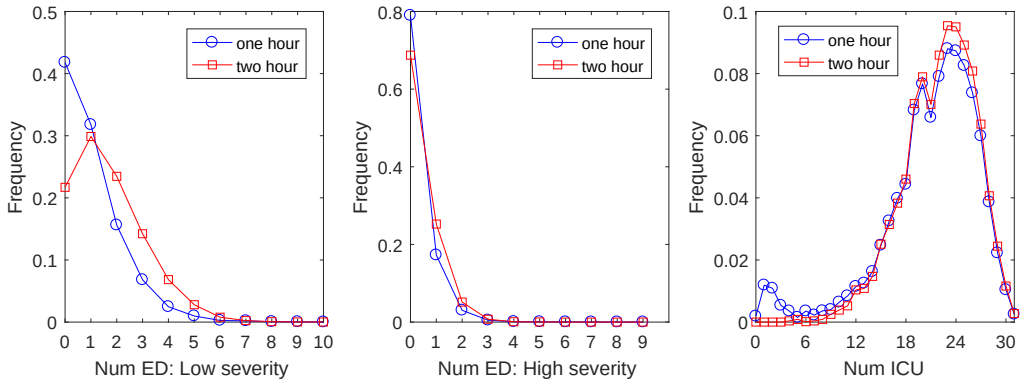
We further compare the choice of one hour versus two hours as the interval length. We construct the state-action pairs with a one-hour time window and compare their distribution with those from the two-hour time window as used in our estimation. The distribution of the state-action pairs is the main variation in the data that provide identification for the model parameters. We find that the distributions from the two time slots are close for the

number of ICU admissions of both classes (a_l and a_h), number of patients in ICU (n^I), and number of high severity patients in ED (n_h^E) who are much more likely to be admitted to ICU. This is not surprising as the ICU admission decisions and the ED arrivals of high severity patients happen infrequently in our data. Thus, while using one-hour time window may change the cost estimates and their interpretations, we believe it will not substantially change the estimated discount factor and the predictions in the counterfactual studies as the observed ICU occupancy levels and ICU admission decisions are very similar to those with the two-hour time window. For illustration, Figures C.3 and C.4 show the distributions of states and actions with one-hour and two-hour time windows for Hospital 4, which has 31 beds in ICU.

Table C.6: Proportion of intervals with admission actions

Interval	5 min	15 min	30 min	1 hour	2 hour
Ratio	4.2%	12.1%	22.5%	38.8%	59.3%

Figure C.3: Distribution of states for Hospital 4 with one-hour and two-hours windows

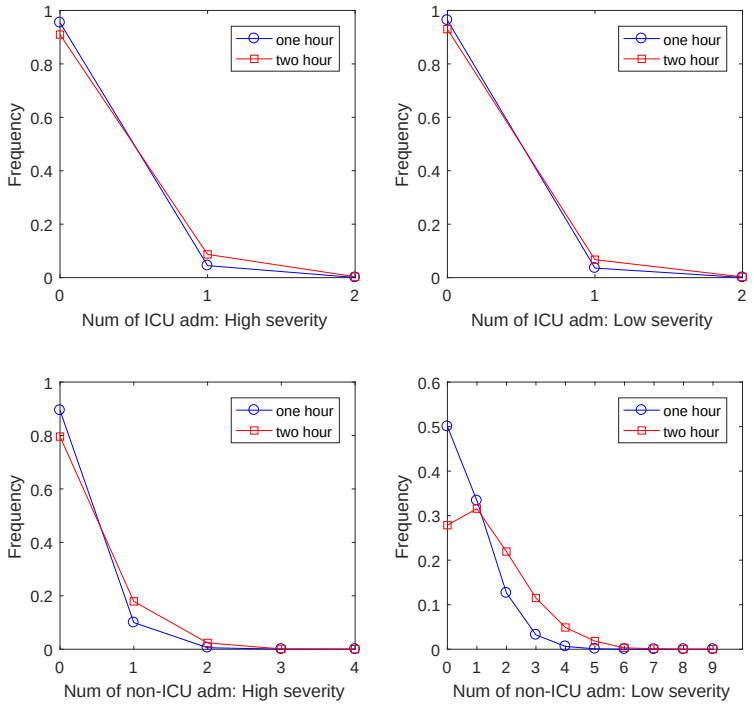


C.4 Proof for Proposition 2

Proof:

Preliminaries : From the definition of the value functions $V(s_t, \varepsilon_t)$ in (3.5) and $\tilde{V}(s)$ in

Figure C.4: Distribution of actions for Hospital 4 with one-hour and two-hours windows



(3.10), we have:

$$\begin{aligned}\tilde{V}(s) &= \sum_{s_t}\int_{\varepsilon_t} V(s_t,\varepsilon_t)g(s_t|s)q(\varepsilon_t|s_t)d\varepsilon_t \\ &= \sum_{s_t}\int_{\varepsilon_t}\sup_{d_t\in\Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty}\beta^{j-t}U(s_j,d_j,\varepsilon_j)|s_t,\varepsilon_t\right\} g(s_t|s)q(\varepsilon_t|s_t)d\varepsilon_t. \qquad \text{(C.6)}\end{aligned}$$

The expectation above is taken over the transition of (s_j,ε_j) starting from (s_t,ε_t).

Recall $g(s'|s)$ is the transition probability from the intermediate state s (after action is taken) to the state s' at the start of the next period, and $q(\varepsilon'|s')$ is the probability density of the random utility component in the next period. Thus, the function $\tilde{V}(s)$ represents the expected future utilities starting from intermediate state s and assuming the hospital always takes the optimal action.

We introduce following notations. We consider two systems s and s'. For every period t, we use s_t and s'_t to denote the system states at the start of period t, which are

$$s_t = \left(n^E_{l,t}, n^E_{h,t}, n^I_t\right) \text{ and } s'_t = \left(\left(n^E_{l,t}\right)', \left(n^E_{h,t}\right)', \left(n^I_t\right)'\right).$$

Additionally, we use

$$\tilde{s}_t = \left(\tilde{n}^E_{l,t}, \tilde{n}^E_{h,t}, \tilde{n}^I_t\right) \text{ and } \tilde{s}'_t = \left(\left(\tilde{n}^E_{l,t}\right)', \left(\tilde{n}^E_{h,t}\right)', \left(\tilde{n}^I_t\right)'\right)$$

to denote the intermediate states after actions d_t and d'_t are taken in period t, which are given by $\tilde{s}_t = \varphi(s_t, d_t)$ and $\tilde{s}'_t = \varphi(s'_t, d'_t)$ according to (3.6). Note that for notational compactness we suppress the dependence of the states on the action. Assume the two systems start from intermediate states $\tilde{s}_0$ and $\tilde{s}'_0$ with

$$\tilde{n}^E_{i,0} = \left(\tilde{n}^E_{i,0}\right)' \text{ for } i \in \{l, h\}, \text{ and } \tilde{n}^I_0 \le \left(\tilde{n}^I_0\right)', \tag{C.7}$$

then Proposition 2 translates to

$$\tilde{V}(\tilde{s}_0) \ge \tilde{V}(\tilde{s}'_0). \tag{C.8}$$

Coupling: Our proof is based on a coupling argument and induction in time. We first introduce the coupling of the two systems s and s' as follows: First, the two systems witness identical arrivals to ED and external arrivals to ICU in every period, i.e., $A_{l,t} = A'_{l,t}$, $A_{h,t} = A'_{h,t}$, and $E_t = E'_t$ for every t. Next, we couple the ICU departures from the two systems as follows. Denote the numbers of ICU patients before departures by $\bar{n}^I_t$ and $(\bar{n}^I_t)'$ respectively and assume $\bar{n}^I_t \le (\bar{n}^I_t)'$. Then, the departures D_t and D'_t are coupled as $D'_t = D_t + Z_t$, where D_t is a Binomial-$(\bar{n}^I_t, \mu_I)$ variable and Z_t is a Binomial-$((\bar{n}^I_t)' - \bar{n}^I_t, \mu_I)$ variable. That is, the number of departures in the s' system is always at least as many as the number in the s system. Finally, if an identical action d is taken in each system, the random utility components $\varepsilon_t(d)$ and $\varepsilon'_t(d)$ associated with that action d coincides for the two systems in every period t.

Under the coupling described above, we first prove following lemma that establishes the relationship between intermediate states and states at the start of next period.

Lemma 3. *Under the coupling, if the intermediate states in each system in period $t-1$ satisfy*

$$\tilde{n}^E_{i,t-1} = \left(\tilde{n}^E_{i,t-1}\right)' \text{ for } i \in \{l, h\}, \text{ and } \tilde{n}^I_{t-1} \le \left(\tilde{n}^I_{t-1}\right)', \tag{C.9}$$

then the states at the start of period t satisfy

$$n^E_{i,t} = \left(n^E_{i,t}\right)' \text{ for } i \in \{l, h\} \text{ and } n^I_t \le \left(n^I_t\right)'. \tag{C.10}$$

PROOF: See Appendix C.4.1. □

Mimicking Policy: We now define the policies used in each system. We assume the system s' always takes its optimal action which achieve the supremum in (C.6). For system s, we define a mimicking policy π which mimics the action taken in the s' system whenever possible; if it is not possible, it takes its own optimal action. We denote the value function associated with this policy by $V^{\pi}(s)$, which is defined by (C.6) with optimal action d_t replaced by the one under policy π. Such a policy is not necessarily optimal for system s and, by definition, we have

$$\tilde{V}(\tilde{s}_0) \geq V^{\pi}(\tilde{s}_0). \tag{C.11}$$

To prove the proposition, we will establish following two properties under our coupling and the policy π. First, two systems always have same number of patients in the ED, but system s has no more patients in the ICU:

$$n^E_{i,t} = \left(n^E_{i,t}\right)' \text{ for } i \in \{l, h\} \text{ and } n^I_t \leq \left(n^I_t\right)', \quad \forall t. \tag{C.12}$$

Second, the action taken in the s' system is always admissible for system s; thus system s always mimics the action of s' under π:

$$d_t = d'_t \in \Pi(s_t), \quad \forall t. \tag{C.13}$$

Note that (C.12) directly implies (C.13), as it follows from (3.1) that given the same number of patients in ED, the system with fewer ICU patients has a larger admissible action set, leading to $d'_t \in \Pi(s'_t) \subseteq \Pi(s_t)$.

Induction: We establish (C.12) for every t by induction.

Base Case: The base case follows directly from the relationship of the initial intermediate states $\tilde{s}_0$ and $\tilde{s}'_0$, which satisfy (C.7), and from Lemma 3.

$$n^E_{i,1} = \left(n^E_{i,1}\right)' \text{ for } i \in \{l, h\} \text{ and } n^I_1 \leq \left(n^I_1\right)'. \tag{C.14}$$

Inductive Step: We assume (C.12) holds for period j and show this implies it holds for period $j+1$.

In period j, under policy π, system s takes the same action of s' since by the inductive hypothesis the action is admissible, i.e., $d_j = d'_j$. Given the same action is taken in each

system, the intermediate states after action, satisfy the following relationship:

$$\begin{aligned}\tilde{n}^E_{i,j} &= n^E_{i,j} - a_{i,j} - r_{i,j} = \left(n^E_{i,j}\right)' - a'_{i,j} - r'_{i,j} = \left(\tilde{n}^E_{i,j}\right)' \text{ for } i \in \{l,h\},\\ \tilde{n}^I_j &= n^I_j + a_{l,j} + a_{h,j} \le \left(n^I_j\right)' + a'_{l,j} + a'_{h,j} = \left(\tilde{n}^I_j\right)'. \end{aligned} \tag{C.15}$$

Finally, we can apply Lemma 3 to prove the relationship (C.12) holds for period $j+1$. This completes the inductive step.

Per-Period Utilities: We have shown that under our coupling and the policy π for the s system, at the start of each period, the two systems always have same numbers of patients in the ED, and system s always has fewer patients in the ICU than that in system s'. Thus, the system s always mimics the action by s' under the policy π. We next prove the per-period utilities always coincide for the two systems, which follows by:

$$\begin{aligned} U(s_t, d_t, \varepsilon_t) &= -c(s_t, d_t) + \varepsilon_t(d_t) = -\sum_{i\in\{l,h\}} c_{r.i} r_{i,t} - \sum_{i\in\{l,h\}} c_{w,i}\left(n^E_{i,t} - a_{i,t} - r_{i,t}\right) + \varepsilon_t(d_t)\\ &= -\sum_{i\in\{l,h\}} c_{r.i} r'_{i,t} - \sum_{i\in\{l,h\}} c_{w,i}\left(\left(n^E_{i,t}\right)' - a'_{i,t} - r'_{i,t}\right) + \varepsilon'_t(d'_t) = U(s'_t, d'_t, \varepsilon'_t). \end{aligned}$$

This is because: (i) Both systems take the same action, thus they admit and reroute same numbers of patients, this leads to same non-ICU admission costs; (ii) As both systems have same numbers of patients in the ED, the number of patients remaining in the ED after actions are also the same, leading to the same waiting costs; (iii) By our coupling, the random utility components coincide for the same action $d_t = d'_t$, i.e., $\varepsilon_t(d_t) = \varepsilon'_t(d'_t)$.

As the per-period utilities coincide for every period given system s takes policy π and system s' takes its own optimal policy, we have

$$V^\pi(\tilde{s}_0) = \tilde{V}(\tilde{s}'_0),$$

then it follows by (C.11)

$$\tilde{V}(\tilde{s}_0) \ge V^\pi(\tilde{s}_0) = \tilde{V}(\tilde{s}'_0).$$

This proves the proposition. □

C.4.1 Proof of Lemma 3

PROOF: The result follow directly from the coupled arrivals and departures in the two systems. Since we start from the intermediate state, the system evolution to period t is only

dictated by the stochastic arrivals to the ED, external arrivals to the ICU, and departures from the ICU during period $t-1$.

It is trivial to see the relationship for the ED patients holds by our coupling of the ED arrival processes. Since, by our coupling $A_{i,t-1} = A'_{i,t-1}$, then $n^E_{i,t} = \left(n^E_{i,t}\right)'$ for $i \in \{l,h\}$ directly follows from our assumption that $\tilde{n}^E_{i,t-1} = \left(\tilde{n}^E_{i,t-1}\right)'$.

We now consider the ICU patients. By our coupling, the s and s' systems see the same number external arrivals, $E_{t-1} = E'_{t-1}$. Then, the total number of ICU patients in each system before departures satisfies the following relationship:

$$\bar{n}^I_t \quad = \quad \min\left\{\tilde{n}^I_{t-1} + E_{t-1}, B\right\} \leq \min\left\{(\tilde{n}^I_{t-1})' + E'_{t-1}, B\right\} = (\bar{n}^I_t)', \tag{C.16}$$

By our coupling, the number of departures from the ICU in the s system is related to the number of departures in the s' system as follows: $D_{t-1} + Z_{t-1} = D'_{t-1}$. Thus,

$$\begin{aligned}\left(n^I_t\right)' - n^I_t &= \left(\bar{n}^I_{t-1}\right)' - \bar{n}^I_{t-1} - \left(D'_{t-1} - D_{t-1}\right) \\ &= \left(\bar{n}^I_{t-1}\right)' - \bar{n}^I_{t-1} - Z_{t-1} \geq 0.\end{aligned}$$

The last inequality follows as $Z_{t-1} \leq \left(\bar{n}^I_{t-1}\right)' - \bar{n}^I_{t-1}$. This completes the proof for Lemma 3. □

C.5 Proof for Lemmas 1 and 2

We first provide the proof for Lemma 1.

PROOF: Recall the two states are given by $s_1 = (1,0,0)$ and $s_2 = (0,1,0)$, i.e., s_1 and s_2 have one patient in the ED from the low and high severity class, respectively . The probabilities of admitting the patient to ICU and non-ICU units under the two states are denoted by $\mathsf{Pr}(a|s_1)$, $\mathsf{Pr}(r|s_2)$, $\mathsf{Pr}(a|s_1)$, and $\mathsf{Pr}(r|s_2)$. Following Proposition 1, we can compute the log of these choice probabilities as

$$\ln \mathsf{Pr}(a|s_1) \quad = \quad \beta\tilde{V}(s_a) - C(s_1), \tag{C.17a}$$

$$\ln \mathsf{Pr}(r|s_1) \quad = \quad -c_{r,l} + \beta\tilde{V}(s_r) - C(s_1), \tag{C.17b}$$

and

$$\ln \mathsf{Pr}(a|s_2) = \beta\tilde{V}(s_a) - C(s_2), \tag{C.17c}$$

$$\ln \mathsf{Pr}(r|s_2) = -c_{r,h} + \beta\tilde{V}(s_r) - C(s_2), \tag{C.17d}$$

where $s_a = (0,0,1)$ denotes the system state after the patient is admitted to the ICU, and $s_r = (0,0,0)$ denotes the system state after the patient is admitted to non-ICU unit (ward). Note the choice probabilities of admission decisions in (C.17a) and (C.17c) used the assumption that the per-period cost of admitting a patient to the ICU is zero for both classes of patients. Taking the differences in the choice probabilities across the two states for ICU and non-ICU admission decisions respectively, we obtain

$$\ln \mathsf{Pr}(a|s_1) - \ln \mathsf{Pr}(a|s_2) = C(s_2) - C(s_1). \tag{C.18a}$$

and

$$\ln \mathsf{Pr}(r|s_1) - \ln \mathsf{Pr}(r|s_2) = c_{r,h} - c_{r,l} + C(s_2) - C(s_1). \tag{C.18b}$$

Here we see the terms involving the value function $\tilde{V}(\cdot)$ in (C.17a) – (C.17d) are cancelled out in the differences. Further subtracting (C.18a) from (C.18b), we get rid of the terms related to state-dependent function $C(\cdot)$ and identify the non-ICU admission cost difference as

$$\begin{aligned} c_{r,h} - c_{r,l} &= \ln P(r|s_1) - \ln \mathsf{Pr}(r|s_2) - (\ln \mathsf{Pr}(a|s_1) - \ln \mathsf{Pr}(a|s_2)) \\ &= \ln\left(\frac{\mathsf{Pr}(r|s_1)}{\mathsf{Pr}(a|s_1)}\right) - \ln\left(\frac{\mathsf{Pr}(r|s_2)}{\mathsf{Pr}(a|s_2)}\right). \end{aligned} \tag{C.19}$$

This proves Lemma 1. In particular, the steps in (C.18a), (C.18b), and (C.19) show how well-constructed state and action pairs can be used to disentangle the complex structure in the choice probability expression (3.12). By taking differences across the two states for ICU and non-ICU admission decisions in (C.18a) and (C.18b) respectively, we remove the term $\beta\tilde{V}(\cdot)$ that is related to future payoff. Then, by taking difference between the two decisions in (C.19), we further remove the state-dependent terms $C(\cdot)$ to expose the cost parameters. □

We then provide the proof for Lemma 2.

PROOF: In Lemma 2, the two states considered are $s_1 = (0, 1, 1)$ and $s_2 = (0, 1, 0)$. That is, there is one patient in the ICU in s_1, and one high severity patient in the ED in both states. Following Proposition 1, we can compute the log probabilities for admitting the patient to ICU and non-ICU unit under the two states as

$$\ln \mathsf{Pr}(a|s_1) = \beta\tilde{V}_2 - C(s_1), \tag{C.20a}$$

$$\ln \mathsf{Pr}(r|s_1) = -c_{r,h} + \beta\tilde{V}_1 - C(s_1), \tag{C.20b}$$

and

$$\ln \mathsf{Pr}(a|s_2) = \beta\tilde{V}_1 - C(s_2), \tag{C.20c}$$

$$\ln \mathsf{Pr}(r|s_2) = -c_{r,h} + \beta\tilde{V}_0 - C(s_2); \tag{C.20d}$$

here $\tilde{V}_k = \tilde{V}((0, 0, k))$ denotes the value function of the state with k patients in the ICU and no patients in the ED. Computing the difference in log probabilities across the two states for ICU and non-ICU admission decisions respectively, we get

$$\ln \mathsf{Pr}(a|s_1) - \ln \mathsf{Pr}(a|s_2) = \beta(\tilde{V}_2 - \tilde{V}_1) + C(s_2) - C(s_1), \tag{C.21a}$$

$$\ln \mathsf{Pr}(r|s_1) - \ln \mathsf{Pr}(r|s_2) = \beta(\tilde{V}_1 - \tilde{V}_0) + C(s_2) - C(s_1), \tag{C.21b}$$

Then we can remove the terms related to state-dependent function $C(\cdot)$ by subtracting the above two equations, which leads to

$$\beta\left[(\tilde{V}_2 - \tilde{V}_1) - (\tilde{V}_1 - \tilde{V}_0)\right] = \ln\left(\frac{\mathsf{Pr}(a|s_1)}{\mathsf{Pr}(r|s_1)}\right) - \ln\left(\frac{\mathsf{Pr}(a|s_2)}{\mathsf{Pr}(r|s_2)}\right). \tag{C.22}$$

This proves Lemma 2. Again, this well-constructed state and action pair enable us to unpack the terms in the log choice probability (3.12). By taking differences across the two states in (C.21a) and (C.21b), we get rid of the terms related to the per-period costs. Then, by taking differences between the two actions in (C.22), we further remove the terms related to $C(\cdot)$ to get the desired result of the value function. □

Appendix D

Chapter 4 Supplemental Information

D.1 Description and Summary Statistics of Independent Variables in (4.1) and (4.2)

To control for the effects of patient's characteristics and severity levels, we include a comprehensive list of demographic, risk, operative, and operational factors as independent variables in our estimation. Some of these factors are already discussed in Section 4.3. We now provide the description and summary statistics for other independent variables included in X_i for our models (4.1) and (4.2).

In Table D.1, we document the descriptions, types, and summary statistics of the independent variables. Besides, we provide their locations in the STS data collection form. We handle the missing values in the binary and categorical variables as follows: if the number of missing observations is smaller than 100 (1.8% of the sample), we impute their values using the majority from the cases in the same NYHA class. Otherwise, we add a new category "Unknown" to represent the missing values. The summary statistics of the categorical variables are reported in Table D.2. Note that the NYHA classification is not available (N/A) if the patient has not experienced a heart failure. The PA pressure is coded as "High" if it is higher than 55mg, and "Low" otherwise.

We classify the cases to different surgery types to control for the procedures performed by the surgeons. First, we have eight standard surgery types from the STS data. For the cases that do not fall into the standard types, we classify their surgery types by the following heuristic rule. We collect from the STS data which of the following four procedures are performed in the surgery: coronary artery bypass, valve, other cardiac procedure, and other non-cardiac procedure. If only one of the four procedures is performed, we classify the case as

a non-standard isolated type, e.g., "non-standard isolated valve" if only the valve procedure is conducted. If more than one of the procedures are performed, we classify the case as the "non-standard multiple" type. Finally, if none of the four procedures is performed, we classify it as "not identified." In total, we have six types for the non-standard cases, i.e., four non-standard isolated ones, non-standard multiple, and not identified. The numbers of cases of each type (both standard and non-standard ones) are summarized in Table D.3.

Table D.1: Description and Summary Statistics of Other Independent Variables in Models (4.1) and (4.2)

Variable	Description	Section in STS	Type	Mean
Race	Patient's race	Demographics	Categorical	-
Endocard	Endocarditis	Risk factor	Binary	0.053
PeriAD	Peripheral arterial disease	Risk factor	Binary	0.088
Lung	Lung disease with severity $\geq$ mild	Risk factor	Binary	0.192
Hypertension	Hypertension	Risk factor	Binary	0.777
CaroStenosis	Carotid Stenosis	Risk factor	Binary	0.054
Syncope	Syncope	Risk factor	Binary	0.031
Dialysis	Dialysis for renal failure	Risk factor	Binary	0.030
Diabetes	Insulin control for diabetes	Risk factor	Binary	0.111
Liver	Liver disease	Risk factor	Binary	0.022
Cancer	Cancer within five years	Risk factor	Binary	0.062
Thoracic	Thoracic aorta disease	Risk factor	Binary	0.094
DrugUse	Recent or remote drug use	Risk factor	Binary	0.088
Smoke	Smoke status of patient	Risk factor	Categorical	-
PrevCI	Previous cardiac intervention	Previous Intervention	Binary	0.431
CardShock	Cardiogenic shock	Preoperative	Binary	0.076
MI	Prior MI	Preoperative	Binary	0.120
NYHA	NYHA classification	Preoperative	Categorical	-
Aorta	Aorta procedure performed	Operative	Binary	0.123
Incidence	Non-initial cardiovascular surgery	Operative	Binary	0.188
PA_Pressure	PA systolic pressure	Hemodynamics	Categorical	-
TotCABG	Number of arteries bypassed	Coronary Bypass	Continuous	1.36

In summary, the independent variable X_i in (4.1) and (4.2) includes the factors in Table D.1, patient's gender and age, surgery status, patient's admission type, surgery type in Table D.3, surgeon's identifier, patient's pre-LOS, block schedule status, and dummies for weekday, month, and year of the surgery.

Table D.2: Summary Statistics of Categorical Variables in Table D.1

Variable	Category	Num Obs.	Ratio
NYHA	N/A	1933	0.361
	I	516	0.096
	II	998	0.186
	III	991	0.185
	IV	663	0.124
	Unknown	251	0.047
Race	White	4273	0.798
	Asian	590	0.110
	Black	274	0.051
	Other	215	0.040
Smoke	FALSE	2694	0.503
	TRUE	2429	0.454
	Unknown	229	0.043
PA Pressure	High	376	0.070
	Low	2247	0.420
	Unknown	2729	0.510

Table D.3: Numbers of Cases by Surgery Types

Surgery Type	Number of Cases	Ratio
CABG	1718	0.321
AVR	683	0.128
MVR	225	0.042
MVr	254	0.047
CABG + AVR	318	0.059
CABG + MVR	57	0.011
CABG + MVr	58	0.011
AVR + MVR	107	0.020
Non-standard isolated Valve	574	0.107
Non-standard isolated CAB	28	0.005
Non-standard isolated cardiac	369	0.069
Non-standard isolated non-cardiac	15	0.003
Non-standard multiple	690	0.129
Not identified	256	0.048

D.2 Definition of Independent Variables in the Schedule Imputation Model (4.3)

In this section, we document the independent variables included in $X'_{s,t}$ for the logistic regression model (4.3). To impute whether the surgeon is assigned a block schedule on a given day, we include multiple operational factors related to the workload of the focal and other surgeons. As we mentioned in the main body, future information can be included in $X'_{s,t}$ as we are imputing the block schedule instead of making any prediction.

Table D.4 summarizes the variables included in $X'_{s,t}$ (plus a constant term) for the logistic model (4.3). In particular, $ORTime_{s,t}$ denotes the sum of OR time of the cases by surgeon s on day t, ignoring overlapping due to surgery parallel. $StartHour_{s,t}$ and $EndHour_{s,t}$ are calculated using the OR entry and exit time of the cases by surgeon s on day t; $StartHour_{s,t}$ (resp. $EndHour_{s,t}$) corresponds to the OR entry (resp. OR exit) time of the earliest (resp. latest) case, rounded to the nearest hour. $PatRemain_{s,t}$ is the number of patients remaining in the hospital for surgeon s. This refers to the patients that (1) already admitted to the hospital by day $t-1$, (2) surgeries have not been performed by day $t-1$, and (3) surgeries are eventually performed by surgeon s. Besides, $WDElecRatio_{s,t}$ is the proportion of elective cases by surgeon s in $[t-180, t+180]$ that fall on the same weekday as t. This variable is included as surgeons' blocks tend to fall on specific weekdays to reduce the variation in surgeons' schedule. Besides, most of the elective cases are performed in their surgeons' block. We have in total 23 independent variables (plus a constant term) in the logistic model (4.3) for imputing the block schedule.

Table D.4: Definition of Independent Variables in the Schedule Imputation Model (4.3)

Variable	Definition
$Surg_s$	Dummy variable for surgeon identifier
$WeekDay_t$	Dummy variable for weekday of t
$ElecCur_{s,t}$, $UrgCur_{s,t}$, $EmergCur_{s,t}$	Number of elective/urgent/emergent cases by surgeon s on day t
$ElecOth_{s,t}$, $UrgOth_{s,t}$, $EmergOth_{s,t}$	Number of elective/urgent/emergent cases by other surgeons on day t
$ORTime_{s,t}$	Total OR time of cases by surgeon s on day t
$StartHour_{s,t}$, $EndHour_{s,t}$	Start and end of the cases by surgeon s on day t
$StartLate_{s,t}$, $EndEarly_{s,t}$	Indicators for $StartHour_{s,t} \geq$ 8AM and $EndHour_{s,t} \leq$ 3PM
$AdmCur_{s,t}$	Numbers of patients admitted by surgeon s on day t
$PatRemain_{s,t}$	Numbers of patients remaining in the hospital for surgeon s
$NumPreDay_{s,t}$, $NumPostDay_{s,t}$	Numbers of cases by surgeon s in the previous and next weekday
$WorkPreDay_{s,t}$, $WorkNextDay$	Indicators for $NumPreDay_{s,t} \geq 1$ and $NumPostDay_{s,t} \geq 1$
$NumCurWeek_{s,t}$	Numbers of days worked by surgeon s in current calendar week
$DistLast_{s,t}$, $DistNext_{s,t}$	Number of days from the previous and next working day of surgeon s
$WDElecRatio_{s,t}$	Proportion of elective cases by surgeon s in $[t-180, t+180]$ that are performed on the same weekday as t

D.3 Supplementary Tables

This section includes the supplementary tables. In particular, Tables D.5 and D.6 report the estimated effects on surgery duration and patient outcomes when we use total incision time of other cases (Sum_Inc_i) as surgeon's daily workload measure. Tables D.7 and D.8 show the estimated coefficient γ for the two workload measures in (4.2) for the three binary outcomes.

Table D.5: Estimated Effects of Daily Workload (Total Incision Time of Other Cases) on Surgery Duration and Patient Outcomes: Full Sample

	Continuous y_i: Coefficients			Binary y_i: AME		
	Incision time	Post-LOS	Total ICU time	Reoperation	Readmission	Mortality
Panel A: Full	0.083^{*}	0.269^{**}	0.194^{*}	0.005^{*}	$0.013^{\dagger}$	0.002
	(0.038)	(0.100)	(0.088)	(0.002)	(0.008)	(0.002)
Num Obs.	5345	5344	5319	5345	5116	5081
Panel B: Full (w/o IV)	−0.016	0.015	0.009	−0.000	−0.000	$0.002^{\dagger}$
	(0.010)	(0.034)	(0.016)	(0.001)	(0.001)	(0.001)
Num Obs.	5345	5344	5319	5345	5116	5081

The estimated effects of surgeon's daily workload (total incision time of other cases) on surgery duration and patient outcomes for the full sample. We report the estimated coefficients in (4.1) for the three continuous dependent variables, and the AME from (4.2) for the three binary dependent variables. Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table D.6: Estimated Effects of Daily Workload (Total Incision Time of Other Cases) on Surgery Duration and Patient Outcomes: Elective and Non-elective Sample

	Continuous y_i: Coefficients			Binary y_i: AME		
	Incision time	Post-LOS	Total ICU time	ReOp	Readmission	Mortality
Panel A: Elec	0.072^{**}	−0.063	0.005	0.002	0.004	0.012
	(0.025)	(0.182)	(0.135)	(0.007)	(0.012)	(0.009)
Num Obs.	2474	2474	2454	2394	2398	1897
Panel B: Non-elec	0.098	0.606^{*}	0.384^{*}	0.010^{*}	0.018^{*}	−0.001
	(0.064)	(0.264)	(0.192)	(0.004)	(0.008)	(0.002)
Num Obs.	2871	2870	2865	2871	2697	2769

The estimated effects of surgeon's daily workload (total incision time of other cases) on surgery duration and patient outcomes for the elective and non-elective sample. We report the estimated coefficients in (4.1) for the three continuous dependent variables, and the AME from (4.2) for the three binary dependent variables. Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table D.7: Estimated Coefficients of Daily Workload in (4.2) for Binary Outcomes: Full Sample

	Reoperation		Readmission		Mortality	
	$NumCases$	$SumInc_i$	$NumCases$	$SumInc_i$	$NumCases$	$SumInc_i$
Panel A: Full	0.268^{**}	0.049^{*}	$0.371^{\dagger}$	$0.075^{\dagger}$	0.197	0.039
	(0.089)	(0.021)	(0.203)	(0.041)	(0.213)	(0.039)
Num Obs.	5345	5345	5116	5116	5081	5081
Panel B: Full (w/o IV)	−0.019	−0.004	−0.002	−0.002	0.190^{*}	$0.034^{\dagger}$
	(0.038)	(0.006)	(0.050)	(0.008)	(0.083)	(0.018)
Num Obs.	5345	5345	5116	5116	5081	5081

Estimated coefficient γ in (4.2) for the full sample. Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table D.8: Estimated Coefficients of Daily Workload in (4.2) for Binary Outcomes: Elective and Non-elective Sample

	Reoperation		Readmission		Mortality	
	$NumCases$	$SumInc_i$	$NumCases$	$SumInc_i$	$NumCases$	$SumInc_i$
Panel A: Elec	0.273	0.040	0.183	0.024	1.035^{*}	0.202^{*}
	(0.524)	(0.096)	(0.411)	(0.077)	(0.422)	(0.082)
Num Obs.	2394	2394	2398	2398	1897	1897
Panel B: Non-elec	0.329^{*}	0.065^{*}	0.452^{*}	0.098^{*}	−0.060	−0.010
	(0.131)	(0.028)	(0.186)	(0.040)	(0.197)	(0.038)
Num Obs.	2871	2871	2697	2697	2769	2769

Estimated coefficient γ in (4.2) for the elective and non-elective samples. Standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.