Essays in Applied Microeconomics

Ajin Lee

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

COLUMBIA UNIVERSITY

2017

ABSTRACT

Essays in Applied Microeconomics

Ajin Lee

This dissertation consists of three essays in applied microeconomics. Each chapter covers a large category of public spending in the US: (1) health care; (2) social insurance; and (3) education. This dissertation aims to understand the determinants of efficient delivery of public programs, focusing on disadvantaged subpopulations.

The first chapter looks at the effectiveness of health care systems. Medicaid, the largest public health insurance program in the US, has transitioned from a fee-for-service system (FFS) primarily administered by the government to a managed care system (MMC) administered by private insurers over the last few decades. I examine how hospitals' responses to financial incentives under these two systems affect hospital costs and newborn health outcomes. I analyze the universe of inpatient discharge records across New York State from 1995-2013, totaling 4.5 million births. First, I exploit an arbitrary determinant of MMC enrollment: infants weighing less than 1,200 grams were excluded from MMC and were instead served through FFS. Using a regression discontinuity design, I find that newborns enrolled in MMC stayed fewer days in hospitals and thus had less expensive visits relative to newborns enrolled in FFS. The cost difference is driven by birth hospitals retaining more newborns enrolled in FFS while transferring away those enrolled in MMC. I find that MMC had limited impacts on newborn health, measured by in-hospital mortality and hospital readmission. Hospitals tended to transfer out MMC newborns only when a high-quality hospital was nearby, which resulted in these infants receiving uncompromised care. Second, I exploit county-level rollout of the MMC mandate to examine impacts on the full population of infants using a difference-in-difference design. I find that hospitals achieved a similar rate of cost savings as for infants over the 1,200-gram threshold, while length of stay, the probability of transfer, and mortality did not change following the mandate. This finding suggests that there are alternative, successful methods by which hospitals reduce costs under MMC, including for high-risk deliveries.

The second chapter argues that wealth uncertainty influences when couples choose to retire. Using data from the Health and Retirement Study, I show that wives delay retirement when their husbands retire following a job loss. This effect is stronger when husbands are the primary earners, and couples are relatively poorer. This provides evidence of intra-household insurance that mitigates the impact of an unexpected earnings shock. I find that wives tend to delay retirement only until they become eligible for Social Security. This suggests that Social Security benefits can relax households' budget constraints and allow wives to join their husbands in retirement.

The third chapter focuses on heterogeneity in grade retention decisions in New York City public schools. Performance on proficiency exams can be a key determinant of whether students are retained or "held back" in their grade. We find female students in New York City are 25% more likely to be retained in their grade due to exam failure than boys. Hispanic students are 60% more likely and Black students 120% more likely to be retained due to exam failure (relative to White students). Poverty and previous poor performance also increase the likelihood of retention, while being young for grade or short does not. We conclude that "patterned discretion" exists in how standardized test results are utilized.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to my advisors for their generous support during my time at Columbia. Douglas Almond provided helpful suggestions and invaluable support throughout the whole process. I was lucky to work with him on several projects, and my work with him was crucial in my developments as an economist. Wojciech Kopczuk always had thoughtful comments and provided helpful guidance from the start of my time at Columbia. He advised me on my first project at Columbia and provided support to my work ever since. Tal Gross, Kate Ho, and Amy Ellen Schwartz provided helpful comments that broadened my perspective and greatly improved my dissertation.

I want to thank my family and friends who were always there for me every step of the way. I could not have done this without them. My family gave me endless support with their love, encouragement, and comfort throughout the whole process. I am grateful to my classmates and friends who shared the journey with me for six years at Columbia—especially Yogita Shamdasani for being a great friend through my ups and downs.

I am grateful to the seminar participants at Columbia University for their helpful suggestions and comments. This dissertation also benefited from the kind support of the Institute for Education and Social Policy at New York University. I thank Jean Roth at the National Bureau of Economic Research for assistance with the data for my first chapter. I also thank Siddhartha Aneja and Meryle Weinstein for assistance with the data for my third chapter.

Dedication

To my family,

for your love and support

Chapter 1. How Do Hospitals Respond to Managed Care?
Evidence from At-Risk Newborns

1.1 Introduction

Health care spending in the US is notoriously high. In 2014, the US government spent $1.1 trillion on public health insurance programs. 40% of US children are covered by Medicaid, the means-tested health insurance program funded by states and the federal government. To reduce costs, Medicaid has transitioned from the traditional fee-for-service (FFS) system administered by the government to the Medicaid Managed Care (MMC) system administered by private insurers – up from 10% of Medicaid enrollees in the early 1990s to 74% by 2013 (Duggan and Hayford, 2013; CMS, 2015a). This transition is expected to continue as several states expand their MMC programs. Despite this systematic change, the existing literature finds mixed impacts of MMC on both cost and health outcomes, providing little support for the transition to MMC.

This paper examines whether MMC can incentivize hospitals to reduce costs without compromising patient health. I exploit variation in the probability of MMC enrollment at a birth weight cutoff: infants weighing less than 1,200 grams (2 pounds, 10 ounces) were excluded from mandatory enrollment in MMC in New York State and were instead served through the traditional FFS system (NYSDOH, 2000, 2001). I compare infants whose birth weight falls just below the threshold and thus enroll in FFS with infants whose birth weight falls just above the threshold and thus enroll in MMC in a regression discontinuity (RD) design. While local, my estimates are important because they focus on the most expensive newborn deliveries. Infants that weigh below 1,200 grams account for one percent of the total newborn population but incur approximately one-third of total newborn hospital costs. This suggests that potential cost savings relative to FFS are large. Moreover, infants around the cutoff are at-risk newborns whose health outcomes are highly dependent on the quality of care. The mortality rate of infants near the threshold is ten times higher than the overall rate. If MMC compromises the quality of care, cost savings might be traded off against health outcomes.

Under the traditional FFS system, Medicaid reimburses hospitals directly for each service that they provide. The fact that costs were not seen by hospitals may have encouraged over-provision of care with dubious health benefits (Hackbarth et al., 2008; Arrow et al., 2009). Under MMC, Medicaid pays a fixed fee per month per enrollee to intermediary health plans that reimburse hospitals. This fixed fee structure under MMC incentivizes health plans to: (1) cut down unnecessary care in order to minimize cost; and (2) keep their enrollees healthy so as to avoid incurring future costs. A priori, MMC's incentive structure might restrain the excesses of FFS. In practice, MMC may fail to achieve its intended goals for several reasons. First, MMC may lead to under-provision of care. "Churning," the phenomenon of beneficiaries cycling in and out of Medicaid, reduces the incentive of health plans to promote the long-term health of their enrollees. The reduced incentive to manage the quality of care can result in adverse health outcomes. Second, the success of MMC is contingent on *hospitals'* financial incentives. Since MMC does not govern contracts between health plans and hospitals, it is unclear how the actual providers of care would respond to the incentives of MMC.

Focusing on hospital discharge records from New York City, I find that infants above the 1,200-gram threshold are 23 percentage points more likely to participate in MMC compared to infants below the threshold. I also find that they have discontinuously shorter lengths of stay and thereby have less expensive visits compared to infants below the threshold. The cost difference is driven by birth hospitals transferring more infants above the threshold to other short-term hospitals while holding onto lucrative infants below the threshold. Tracking infants across hospitals, I find that the cumulative length of stay and hospital costs are still lower above the threshold. These differences suggest that hospitals internalize financial incentives to reduce costs for MMC infants. I provide additional evidence that financial incentives do indeed drive these hospital responses. Consistent with a profit maximization problem of hospitals, effects are stronger when hospitals' spatial constraints bind (i.e., when they have few Neonatal Intensive Care Unit (NICU) beds available) and when potential receiving hospitals have spare capacity. In addition, the effects are stronger for infants with high expected costs of treatment.

Although costs and care change, I find limited evidence that the reduced amount of care provided to infants above the threshold results in worse health outcomes, as measured by individual-level mortality during hospitalization and the incidence of hospital readmission following the birth episode. I show that receiving hospitals are on average bigger and better-equipped than birth hospitals. Consequently, infants enrolled in MMC are likely to be transferred away; however, these transfers occur to higher-quality hospitals, resulting in minimum harm to health. These results suggest that MMC reallocates at-risk newborns from a lower-quality hospital to a higher-quality hospital.

I propose a mechanism through which hospitals might engage in such behavior in response to MMC: efficient coordination of care between local hospitals. In contrast to the above findings in New York City, I show that there are no differences between MMC and FFS in counties outside of New York City. This suggests that the structure of local health care markets may impact how hospitals respond to MMC. In particular, I consider distance from a birth hospital to a high-quality hospital with a NICU as a possible factor driving the differences between New York City and upstate counties. I find that hospitals are in fact more responsive to MMC when they have a high-quality hospital nearby, even within New York City. This suggests that even if MMC motivates hospitals to selectively transfer infants to maximize their profits, the cost of timely transfers may outweigh the financial benefit for some hospitals due to the lack of an efficient coordination system.

As is well known, RD estimates apply to those with a high probability of being near the threshold (Lee and Lemieux, 2010) and may not apply to other subpopulations. To address this, I exploit the rollout of the MMC mandate across counties in New York State in a difference-in-difference (DD) framework. I find that the DD estimates are comparable to my RD estimates for low birth weight infants. For infants with higher birth weight, I also find that hospitals achieve a similar level of cost reductions without affecting mortality. However, length of stay and the probability of transfer do not change for this group following the MMC mandate, suggesting that hospitals adjust the amount of care conditional on retaining these infants.

I also consider the average characteristics of "compliers" for both RD and DD models. Compliers for the RD model are infants who are induced to enroll in MMC due to exceeding the birth weight threshold at 1,200 grams. Compliers for the DD model are infants who are induced to enroll in MMC due to living in a county at the time of the MMC mandate rollout. I find that two groups of compliers are quite different. For example, compliers in the RD model stay in hospitals that have more beds, staff, and equipment compared to compliers in the DD model, who also have much higher birth weight. This suggests that treatment effects for these two models could differ since hospitals with varying observable characteristics may respond differently to incentives associated with MMC. Indeed, the means by which cost reductions are achieved differ. Nevertheless, the overarching finding of lower cost but similar health outcomes under MMC persists.

The remainder of the paper is organized as follows. Section 1.2 discusses my contributions to the related literature. Section 1.3 provides relevant institutional details. Section 1.4 describes my data and presents descriptive statistics. Section 1.5 describes the main empirical strategy, while Section 1.6 presents the main RD estimates and discusses the mechanism. To further understand hospitals' financial incentives, Section 1.7 explores three sources of heterogeneity: capacity at birth hospitals, capacity at potential receiving hospitals, and expected costs of treatment. Section 1.8 discusses several specification and robustness checks of the main results. Section 1.9 presents the DD estimates and compares complier characteristics between the DD and RD estimates. Section 1.10 discusses conceptual framework and cost implications. Section 1.11 concludes.

1.2 Contributions to the Relevant Literature

This section summarizes the relevant literature and discusses my contributions. The current literature on MMC has three limitations. First, there is no consensus on the effects of MMC as the findings in the literature are mixed. Second, few papers focus directly on provider-level responses, thus limiting our understanding of the mechanisms. Third, most papers focus on relatively healthier subpopulations who might have little room for cost reductions and health improvements. This paper attempts to address each of these three points.

First, I utilize a type of variation that has not been previously explored to identify the effects of MMC. I exploit a discontinuous exclusion from MMC enrollment based on birth weight in an RD framework. To complement my RD strategy, I also estimate a DD model using county-level rollout of the MMC mandate in New York State. Moreover, I compute mean characteristics of compliers for both RD and DD models to further understand the differences between these two models.

Several papers use local MMC mandates as an exogenous source of variation in a DD framework, but the findings are mixed. For instance, Duggan (2004) focuses on the impact on Medicaid expenditures using a local MMC mandate in California as a source of variation. He finds that an MMC mandate in California led to an *increase* in government spending with no health improvement, suggesting that MMC in fact decreased the program efficiency. His findings, however, do not always apply to a similar study in other states. For example, Harman et al. (2014) show that the MMC mandate in Florida led to a *reduction* in Medicaid expenditures. On the other hand, using datasets that represent a national sample, Herring and Adams (2011) and Duggan and Hayford (2013) find no overall effects on expenditures.

Similarly, the findings on the effects of MMC on health outcomes are also inconclusive. Several papers focus on pregnant women and infants as they account for a large share of Medicaid beneficiaries. Aizer et al. (2007) examine prenatal care and birth outcomes in California and find that MMC actually decreased the quality of prenatal care and increased the incidence of low birth weight, pre-term births, and neonatal mortality.[1] Their findings suggest that providers can respond to MMC by limiting care for certain subpopulations, resulting in adverse effects on health.[2] On the

[1] Conover et al. (2001) also find that MMC led to poor prenatal care and negative birth outcomes (lower Apgar scores, but no effect on infant mortality). In addition, Kaestner et al. (2005) document similar findings—poor prenatal care and birth outcomes—but show that their estimates are unlikely to be causal.

[2] Kuziemko et al. (2013) provide evidence on risk-selection under MMC. They find that the transition from FFS to

contrary, some of the earlier findings suggest improvements in prenatal care (Krieger et al., 1992; Levinson and Ullman, 1998; Howell et al., 2004).

Second, I focus on hospital responses to MMC and propose a hospital-level mechanism through which MMC can achieve its goals. Few papers in the literature directly discuss mechanisms and most focus on health plans' incentives. Duggan and Hayford (2013) show that states with high baseline Medicaid reimbursement rates achieved savings, suggesting the government's ability to negotiate lower prices with health plans as a mechanism for reducing health care expenditures under MMC.[3] In addition, Van Parys (2015) examines Florida's 2006 Medicaid reform and discusses that the types of competing health plans in regional health care markets affect how health plans reduce costs. Although it is useful to understand plan-level incentives, the lack of attention on provider-level incentives limits our understanding of how MMC can influence actual provider practice.[4]

Third, I focus on a high-cost subpopulation - low birth weight infants. Newborns are one of the costliest populations treated in US hospitals. In 2011, aggregate hospital costs on newborns were ranked on top among those billed to Medicaid and private insurance (HCUP, 2013). In particular, as Figure 1.1 shows, only around 1% of infants weighed less than 1,200 grams at birth, but they accounted for 22.3% of total costs between 1995 and 2013 in New York State. The literature focuses on relatively healthier subpopulations because most of the local MMC mandates exclude disabled subpopulations and high-cost procedures are often carved out of benefit packages.[5] As

MMC widened black-Hispanic (i.e., high- and low-cost infants) disparities in birth outcomes, suggesting that health plans shift their resources towards low-cost enrollees.

[3] Their findings are consistent with the literature on managed care in the private insurance market. For example, Cutler et al. (2000) examine the effects of managed care on price and quantity of health care for the privately insured, focusing on patients with heart disease. They show that unit prices (i.e., reimbursement payments) are lower under managed care than the traditional indemnity insurance, while they find relative modest differences in quantity (i.e., treatment patterns) and health outcomes.

[4] Marton et al. (2014) discusses how plans reimburse providers greatly affects the reduction in utilization and spending, suggesting that provider-level incentives play a key role in the success of MMC.

[5] One exception is the Florida's Medicaid reform that Van Parys (2015) studies. Florida required disabled beneficiaries who received Medicaid through Supplemental Security Income (SSI) to enroll in MMC. However, Van Parys

a number of states have begun to expand MMC to those with critical conditions (Iglehart, 2011; Libersky et al., 2013; KFF, 2015), however, it is timely and policy-relevant to understand whether MMC can successfully deliver medical care to these populations.

This paper is also related to the literature on hospital responses to a change in prices.[6] Dafny (2005) shows that hospitals "upcode" patients to take advantage of large price increases for certain diagnoses.[7] Acemoglu and Finkelstein (2008) find a large increase in capital-labor ratios following a reform that decreased reimbursement for labor input. Shigeoka and Fushimi (2014) find an increase in NICU utilization following a reform that made it more profitable in Japan. I contribute to this literature by examining how hospitals respond to a change in reimbursement rates for severely ill patients.

Moreover, this paper is related to the literature on returns to early life medical care. Almond et al. (2010) estimate marginal returns to medical care in early life using the very low birth weight classification at 1,500 grams and find that the higher level of medical care below the threshold results in lower mortality. Bharadwaj et al. (2013) use the same identification strategy and find that more medical care in early life leads to higher test scores in the long-term. I focus on a different cutoff at 1,200 grams to examine how different reimbursement methods affect hospitals and early life health care.

1.3 Background

In this section, I provide institutional details on MMC in New York State focusing on newborns. Section 1.3.1 describes mandatory enrollment in MMC in New York State and discusses imperfect compliance with the mandate. Section 1.3.2 describes the exclusion of newborns from mandatory enrollment in MMC based on birth weight. Section 1.3.3 discusses hospital payments under FFS

(2015) does not separately focus on examining the effects of MMC on this disabled subpopulation.

[6] Some papers focus on physicians' financial incentives. For example, see Clemens and Gottlieb (2014).

[7] See also Sacarny (2014) & Geruso and Layton (2015).

versus MMC in treating low birth weight infants.

1.3.1 Mandatory MMC Enrollment in New York State

Medicaid beneficiaries in New York State are generally required to enroll in a managed care plan. The mandatory enrollment in MMC was phased in starting October 1997 in Albany and four other upstate counties. In New York City, the MMC mandate was introduced in August 1999 and was fully implemented in September 2002. As of November 2012, MMC was mandated in all 62 counties. However, the actual share of Medicaid recipients enrolled in MMC falls short of 100%. In July 2015, two and a half years after the full implementation, only 78% of the New York State Medicaid population were enrolled in MMC while the rest were still enrolled in FFS.[8]

Figure 1.2 shows the trends in the share of infants covered by Medicaid in New York State using inpatient discharge records. Medicaid coverage has increased over time, and around half of all births were financed through Medicaid in 2013. The composition of Medicaid coverage has changed dramatically over the study period. In 1995, only about 5% of total Medicaid infants were covered by Health Maintenance Organizations (HMOs), a type of managed care organizations (MCOs), while the rest 95% were covered by non-HMO. By 2013, 83% of total Medicaid infants were enrolled in HMOs, and the rest 17% were served through non-HMO. I use Medicaid HMO and MMC interchangeably in the remainder of the paper based on the comparison between the managed care penetration published by Centers for Medicare & Medicaid Services (CMS) and the share of Medicaid infants enrolled in HMO in my sample.[9]

The share covered by HMO is not 100% even after the statewide implementation of the mandate due to three reasons. First and foremost, there are a few infants who are still covered by

[8] http://kff.org/medicaid/state-indicator/share-of-medicaid-population-covered-under-different-delivery-systems/

[9] According to CMS (2015b), the Medicaid managed care penetration rate in New York State increased from 61.5% in 2005 to 76.7% in 2011. In my sample of infants in New York State, the share of Medicaid infants enrolled in HMO increased from 62.1% in 2005 to 76.2% in 2011. This suggests that Medicaid HMO is a good measure of the total MMC participation in New York State.

Medicaid FFS due to exclusions and exemptions from the MMC enrollment. I exploit one of the exclusions for my identification strategy, which I describe further in the following section. Second, some infants who are newly enrolled in Medicaid might show up as having the FFS coverage in the discharge records at birth, in case their parents fail to enroll their child in a managed care plan in a timely manner.[10] Third, even for infants who are subject to mandatory enrollment, the implementation might not be perfect or immediate due to some administrative shortcomings.

1.3.2 Exclusion Below the 1,200 Grams Birth Weight Threshold

Infants born to pregnant women who are receiving Medicaid on the date of delivery are automatically eligible for Medicaid for one year. If the mother is enrolled in a health plan that provides an MMC option, the child is automatically enrolled in the mother's plan in most cases. When the infant weighs less than 1,200 grams, however, the system receives an alert with an indicator from the hospital noting that the infant should not be enrolled with an MCO for the first six months of their lives. They are instead served through the FFS system. This creates a discontinuous exclusion from MMC based on birth weight, which I exploit in an RD framework to estimate the causal effects of MMC in comparison to FFS.

These infants with very low birth weight were excluded from MMC enrollment along with other subpopulations that are medically complicated and expensive to treat. For example, nursing home residents and people residing in state psychiatric facilities were also excluded from MMC enrollment during the study period (Sparer, 2008). Given the high costs of treatment and clinical complications, these groups were excluded initially due to several concerns raised by both health plans and beneficiaries. Health plans had little experience with severely ill subpopulations and lacked the coordinated delivery system for them. Beneficiaries were also concerned about inadequate provider networks under MMC.

However, the state has been gradually phasing in mandatory enrollment into MMC for these

[10] Newly enrolled Medicaid beneficiaries are given 90 days to choose a health plan.

subpopulations, mainly for greater cost savings. As part of the Medicaid Redesign Team (MRT) initiatives, infants weighing less than 1,200 grams at birth have been no longer excluded from MMC enrollment since April 2012.[11] Therefore, this paper has direct policy implications on whether MMC can achieve cost reductions without harming health outcomes of critically ill newborns.

1.3.3 Hospital Payments Under FFS Versus MMC

Under FFS, hospitals are directly reimbursed by Medicaid in a uniform manner. In New York State, the Medicaid program uses a prospective payment system using Diagnosis Related Groups (DRGs) to reimburse health care providers for inpatient services they provide to FFS enrollees. Each inpatient visit is classified into a DRG based on patient conditions, and Medicaid pays a fixed rate to hospitals according to the DRG assigned to the patient (Quinn, 2008).

Under MMC, Medicaid pays health plans a flat fee per month per enrollee (i.e., capitation) and health plans are responsible for reimbursing hospitals for inpatient services. Therefore, hospital payments under MMC vary depending on contractual details between health plans and hospitals. Health plans choose a wide range of methods in reimbursing providers, from a fee-for-service method to capitation. For inpatient services associated with newborn medical care, however, most health plans in New York State also use a prospective payment system using DRGs.

Since health plans have an incentive to reduce costs given the fixed revenue structure, prospective payment to hospitals under MMC are likely lower than the hospital payments under FFS. According to the New York State Department of Health (NYSDOH), the actual hospital payments under FFS are in fact higher than the suggested hospital payments under MMC.[12] Refer to Appendix Section A for further details on hospital payments. I discuss a conceptual framework of

[11] http://www.health.ny.gov/health_care/medicaid/program/update/2012/2012-02.htm#infants

[12] The suggested hospital payments are intended to be used as base rates where adjustments can be made based on the contracts between health plans and hospitals (http://www.health.ny.gov/facilities/hospital/reimbursement/apr-drg/rates/ffs/index.htm).

hospital responses to different levels of prospective payment in Section 1.10.1.

1.4 Data

For my main analysis, I use inpatient discharge records from State Inpatient Databases (SID) of Healthcare Cost and Utilization Project (HCUP) for New York State from 1995-2013.[13] This dataset contains the universe of inpatient discharge records, thus almost all births. This dataset contains critical information for my identification strategy such as birth weight in grams and primary expected payer. I examine the effects of MMC on various measures of inpatient care including total charges, length of stay (LOS), transfer, and mortality during hospitalization. Starting 2003, New York State Inpatient Databases include encrypted person identifiers that enable researchers to identify multiple hospital visits of the same patient over time. This allows me to distinguish births, transfers, and subsequent visits.

In addition, I use American Hospital Association (AHA) Annual Survey of Hospitals from 1995-2013.[14] This dataset contains detailed information on hospitals such as hospital names, location, staff, and facilities. I use these various hospital characteristics to understand the mechanism through which MMC affects hospital practice.

Table 1.1 provides summary statistics of my main analysis sample, infants in New York State from 2003-2011. I focus on periods between 2003 and 2011 to exploit encrypted person identifiers to track patients over time and to exclude the periods when the exclusion was no longer valid. Among the full sample of newborns in the first column, 43% of the total 2 million discharge records are financed by Medicaid. Within Medicaid, 62% of infants are covered by HMO.

Total charges are list prices for all services provided at the facility to each discharge record. The list price for a given service is the same for all patients regardless of their insurance status. Discounts are applied to list prices for actual payments based on contractual details between each

[13] Data access to HCUP was provided by the National Bureau of Economic Research (NBER).

[14] Access to AHA was also granted by NBER.

insurer and hospital. Although total charges are not the exact payments made by insurers, they are a good proxy for the amount of services provided to a given patient. Total costs are total charges multiplied by hospital-year-specific cost-to-charge ratios. This measure is considered to better reflect how much hospital services actually cost. Total costs are considerably lower than total charges, \$3,500 compared to \$9,609 on average.[15] In the full sample, infants stay on average four days in the hospital. Death is a rare event, around 0.3%. Around 1% of the total newborns experience transfers, and 10% stay in a NICU facility.

The last two columns show means for the sample near the 1,200-gram threshold. Below the threshold, 95% of Medicaid beneficiaries are enrolled in a non-HMO category, which indicates that the exclusion is implemented fairly well. Hospital visits are highly expensive for these very low birth weight infants. Total charges are over \$200,000 below and \$145,000 above the threshold. Total costs are also high, \$75,758 below and \$52,670 above the threshold. These infants stay hospitalized for more than a month on average. Mortality is also greater than the full sample, which is around 5% below the threshold and 2% above the threshold. Transfers occur for more than 10% of these infants, and the majority of them utilize NICU (74-75%).

1.5 Empirical Strategy

To examine the effects of MMC in comparison to FFS, I exploit the 1,200-gram threshold in a regression discontinuity design. That is, I compare infants whose birth weight falls just below the 1,200-gram threshold and thus are served through Medicaid FFS to infants whose birth weight falls just above the threshold and thus are enrolled in MMC. I estimate the following regression to examine the first stage effect of exceeding the threshold on MMC participation. Then, I proceed to examine the reduced-form effects on several discharge outcomes Y_i:

$$Y_i = \alpha + \beta D_i + f(X_i) + \phi_y + \phi_m + \psi_c + u_i \tag{1}$$

[15] All monetary values are in 2011 dollars adjusted by CPI-U.

where i denotes a discharge record. D_i is a binary variable that takes one if the birth weight of a record i is greater than or equal to 1,200 grams. X_i indicates a running variable, which is birth weight centered at 1,200 grams. I control for a trend in birth weight with a linear spline, $f(X_i) = X_i + D_i X_i$. Additionally, to increase precision, I control for admission year fixed effects (ϕ_y), admission month fixed effects (ϕ_m), and hospital county fixed effects (ψ_c). Excluding these additional controls has little impact on the results.

For bandwidth selection, I employ a bandwidth selection method proposed by Calonico et al. (2014) for each outcome. This method suggests a bandwidth ranging from 100 to 200 grams for my main outcome variables. I estimate these models with Ordinary Least Squares (i.e., local linear regressions with a uniform kernel). In the tables, I specify the bandwidth used for each estimation and report the RD estimate β with robust standard errors.[16] As a robustness check, I additionally examine whether the estimates are sensitive to a range of bandwidth choices and functional forms of $f(X_i)$.

The main identifying assumption of my RD design is that control over birth weight is imprecise (Lee and Lemieux, 2010). Figure 1.3 shows the frequency of discharge records by birth weight. Panel (a) plots the histogram using one-gram bins. There are large heaps at multiples of 10 and smaller heaps at multiples of 5, most likely due to rounding in reporting. Other than that, however, there is little evidence of irregular heaps around 1,200 grams. Panel (b) plots the same information using 20-gram bins along with local linear regression fitted lines. For figures, I estimate local linear regressions using the triangular kernel and a bandwidth of 150, separately for below and above the threshold. Again, it shows that the mean frequency is smooth across the threshold. McCrary (2008) test also indicates that the discontinuity estimate is not statistically significant at the 5% level.

In addition, I test whether birth weight is manipulated for infants with high expected costs. Specifically, I compute predicted list prices from regressing total charges on principal diagnosis and principal procedure fixed effects. I then divide the sample by quartiles of the predicted list

[16] Clustering standard errors at the birth weight level does not affect the results (Card and Lee, 2008).

prices. I find no evidence of heaping across the distribution, even for infants in the top quartile of expected costs (Appendix figure C.1). Taken together, I find no evidence of manipulation around the 1,200-gram threshold.

Additionally, I repeat the estimations dropping infants at 1,200 grams ("donut RD") to test whether the tendency to round to 1,200 grams is correlated with other characteristics that are also correlated with my outcomes (Barreca et al., 2011). I find that my results are robust to this restriction, suggesting that the observed heaps are likely random and thus do not interfere with identification.[17]

To further test the validity of the RD design, I examine whether observed predetermined characteristics are similar around the threshold. Since it is difficult to accurately predict birth weight prior to delivery, predetermined characteristics of patients and birth hospitals are unlikely to change discontinuously across the threshold. Table 1.2 summarizes the RD estimates for these baseline characteristics. As expected, none of the estimates are statistically significant, indicating that the exclusion in fact created random variation in enrollment into MMC.

1.6 Main Results

In this section, I present main results separately for New York City in Section 1.6.1 and for counties outside of New York City in Section 1.6.4. Section 1.6.5 considers proximity to a high-quality hospital as a potential mechanism behind the main findings.

1.6.1 New York City

1.6.2 Provider Practice Outcomes

Since treatment at birth can change the course of subsequent hospital care, I distinguish visits at birth from subsequent visits. Panel A of Table 1.3 shows the RD estimates at birth hospitals

[17] My results are also robust to excluding other large heaps and to restricting the estimations to large heaps only.

and Figure 1.4 presents the corresponding figures. Consistent with the policy, panel (a) of Figure 1.4 shows that the MMC participation rate discontinuously increases above the threshold. This corresponds to an increase of 23 percentage points, which constructs a fuzzy RD design.[18] The MMC participation rate below the threshold is close to zero, which indicates that the exclusion from MMC enrollment based on birth weight is strictly implemented.

I show that the higher MMC rate is associated with shorter length of stay, lower charges and costs, consistent with hospitals' incentives to reduce the amount of care for infants enrolled in MMC. Column 2 of Table 1.3 shows that length of stay drops by 12% above the threshold.[19] The large reduction in length of stay results in lower charges and lower costs by similar magnitudes.

The reduction in length of stay could be driven by (1) faster routine discharges from a birth hospital or (2) transfers from a birth hospital to another facility for additional care. I first examine the transfer decision. An inter-hospital transfer is an option for infants who require specialized or intensive care if they are born in inadequately-equipped facilities. For infants below the threshold, hospitals have an incentive to retain them to extract higher payments. However, since the risk of treating the infants at relatively inadequate facilities may be too great further below the threshold, hospitals would keep the healthiest among the infants enrolled in FFS, those right below the threshold. For infants above the threshold, this incentive essentially disappears, and hospitals would rather have an incentive to transfer them. I find that the probability of transfer to another short-term hospital in fact increases by 2.4 percentage points above the threshold. In addition, panel (e) of Figure 1.4 shows that the effect is driven by the lower likelihood of transfer right

[18] The composition of Medicaid beneficiaries might be affected due to differential selection into Medicaid following the MMC mandate. The managed care mandate can make Medicaid participation more appealing for infants above the threshold, while it does not affect those below the threshold as they are excluded from the mandate. For instance, assuming the quality of care is higher under managed care, some families who otherwise would not participate in Medicaid might decide to enroll in Medicaid (Currie and Fahr, 2005). In addition, given that families covered by MMC are given time to choose a health plan, timing of Medicaid enrollment might vary at the threshold. To minimize selection, I do not restrict my estimation to Medicaid participants. In the RD estimation window 52% of the sample have Medicaid, 43% have private insurance, 5% are uninsured.

[19] To be specific, I use log(length of stay+1) as the outcome. Using the inverse hyperbolic sine transformation to avoid adding an arbitrary number one yields the same result.

below the threshold.[20]

I examine whether the shorter length of stay is driven by faster routine discharges by focusing on infants who are routinely discharged from birth hospitals. I find no effects on length of stay or cost measures for this group of infants (e.g., RD estimate for log(length of stay): -0.017; standard error: 0.026). Note that infants who are routinely discharged below the threshold are not comparable to those above the threshold due to the differential probability of transfer across the threshold. Nevertheless, a smooth linear fit around the threshold (Appendix Figures C.2) suggests that transfers are likely the main driver of the reduction in length of stay at birth hospitals.

The majority of transfers occur soon after birth. In my sample, 70% of transfers occur within the first three days after birth (Figure 1.6). Additionally, health plans have limited control over hospitals' decisions on neonatal transfers. Due to the emergency of neonatal transfers, prior authorization by insurers is not required (NYSDOH, 2016). This suggests that transfer decisions are essentially made by hospitals. Moreover, hospitals that receive transferred infants in my sample are "higher-quality" hospitals. Figure 1.7 compares mean characteristics of birth hospitals and receiving hospitals. Receiving hospitals on average have more beds, physicians, and nurses. They are more likely to be teaching hospitals and more likely to have a NICU facility. These hospital characteristics suggest that infants in my sample are generally transferred to higher-quality hospitals that are bigger and better-equipped.

Exploiting the encrypted person identifiers, I further examine how MMC affects subsequent care provided to infants around the birth weight threshold. Panel B of Table 1.3 shows the effects on individual-level outcomes that aggregate outcomes at birth hospitals with outcomes at subsequent visits including transfers (if transferred). The corresponding figures are shown in Figure 1.5.

I find that the magnitudes of the shorter length of stay, lower charges, and costs are smaller when aggregating the amount of care provided at subsequent visits. However, length of stay is

[20] I examine other dispositions such as transfer to other facilities (e.g., skilled nursing facility, intermediate care facility) and home health care, but I find no effects on these measures.

still shorter above the threshold by 9% and the estimate is marginally significant at the 10% level. When including hospital fixed effects (panel C of Table 1.3), the point estimates barely change but precision increases. This suggests that the effects in fact come from within-hospital differences in treatment depending on the infant's insurance status. With hospital fixed effects, the 9% reduction in total costs becomes marginally significant.

1.6.3 Health Outcomes

In this section, I test whether the reduced amount of care provided to infants above the threshold results in worse health outcomes. First, I examine mortality at birth hospitals. If FFS infants receive more resources than MMC infants even among those who remain at birth hospitals, there may be negative health consequences for infants enrolled in MMC at birth hospitals. I find that the point estimate is positive but insignificant (RD estimate: 0.019; robust standard error: 0.016). However, since the probability of transfer changes at the threshold, there may be selection into who remains at birth hospitals, which can differentially affect the probability of death across the threshold.

Subsequently, I track the infants over time and estimate the probability of hospital readmission and individual-level mortality during hospitalization (columns 5 and 6 of Table 1.3 panel B). If the reduced amount of care provided to infants above the threshold at birth was inadequate, the probability of hospital readmission might be higher above the threshold. I find no evidence of that: the point estimate on hospital readmission is zero and statistically insignificant (RD estimate: -0.000; robust standard error: 0.021). This suggests that the reduction in total length of stay at birth may have improved efficiency by cutting down unnecessarily long stays.

The point estimate on individual-level mortality, however, is positive and large although statistically insignificant (RD estimate: 0.015; robust standard error: 0.016). In addition, it is only slightly lower than the estimate at birth hospitals, suggesting that the difference in mortality at birth hospitals are unlikely driven by selection. This is not surprising since more than half of all deaths I observe occur within the first three days following birth. This result suggests a potential shift in resources at birth hospitals from infants above the threshold towards infants below the threshold. Nevertheless, given limited precision, it is hard to conclude that MMC had significant impacts on health outcomes.[21]

Additionally, I examine various outcomes associated with the quality of care and patient health, including hospital readmission due to preventable conditions,[22] level IV NICU stays, any NICU stays, utilization of chest X-rays, ultrasounds, and implants, as well as various therapy services (Appendix Table D.1). I do not detect any statistically significant effect on these measures except for one outcome. For utilization of physical therapy services, I find an increase of 4 percentage points above the threshold, suggesting that if anything MMC may be associated the higher quality of care.

1.6.4 Rest of the State

In this section, I repeat the estimations for counties outside of New York City. Table 1.4 summarizes the effects on discharge outcomes at birth hospitals (panel A) and aggregated outcomes at the individual level (panel B). Appendix Figures C.3 and C.4 show the corresponding figures.

In counties outside of New York City, I find few differences between MMC and FFS. The probability of MMC participation increases discontinuously at the threshold by 15 percentage points, which is slightly lower than the New York City estimate. Panel (a) of Appendix Figure C.3 shows that the Medicaid HMO participation is close to zero below the threshold, while it jumps discontinuously to around 20% above the threshold. Unlike New York City, however, I find no effects on all other discharge outcomes in this sample. The estimates are positive and imprecise. Figures also show little evidence of discontinuous changes in outcomes across the threshold.

The lack of effects on discharge outcomes outside of New York City suggests that local health

[21] Adding various controls (e.g., diagnosis fixed effects) does not reduce standard errors of my mortality outcomes.

[22] I follow the definition of avoidable hospitalizations in Parker and Schoendorf (2000) and Dafny and Gruber (2005).

care markets may play a role in hospital responses to MMC. Since New York City is unique in many aspects compared to the rest of the state, there could be numerous channels through which MMC affects hospitals. For instance, the number of plans is much larger in New York City compared to the rest of the state, which could affect the level of competition in local health care markets and thus the strength of incentives to reduce costs and improve quality.[23] The density of local health care markets can also have an impact on hospital practice style by allowing hospitals to coordinate the provision of care to local patients. In Section 1.6.5, I pay particular attention to the role of proximity between hospitals in understanding this geographical heterogeneity.

1.6.5 Potential Mechanism

In this section, I consider proximity to a potential receiving hospital as a potential mechanism that drives the differences between New York City and the rest of the state. The idea is that costs of transfer may be lower in New York City due to shorter distances between hospitals. The costs may include transportation costs, transaction costs between originating and receiving hospitals, and potential harm to infants' health. There are risks associated neonatal transfers,[24] and the literature documents that the longer duration of transport is associated with increased neonatal mortality (Mori et al., 2007) and poor physiologic status of newborns (Arora et al., 2014).

In particular, I focus on the distance from a birth hospital to a hospital with a NICU as a potential receiving hospital. Focusing on hospitals with a NICU is a natural choice since the majority of infants near the threshold utilize NICU. To illustrate the geographical difference between New York City and the rest of the state, I first measure straight-line distances. Specifically, I geocode the center point of each hospital zip code and compute the distance from a birth hospital to the

[23] Unfortunately, simple comparisons by the number of plans are fraught with the endogeneity of plan entry and exit, and I do not have a valid instrument for the number of plans to further investigate this mechanism in the current project.

[24] For instance, Arad et al. (1999), Mohamed and Aly (2010), Nasr and Langer (2011) & Nasr and Langer (2012) document neonatal transfers are associated with higher mortality and more complications. However, since transfers are not randomly assigned, the resulting outcomes are confounded by selection into transfers.

nearest hospital that provides a NICU facility. The distance between hospitals is much shorter in New York City compared to other counties outside of New York City (Appendix Figure C.5). The median distance is 1.3 miles in New York City and 18 miles outside of New York City.

To examine whether proximity predicts hospitals' practice style, I compare hospitals that have a hospital with a NICU close by with hospitals that have a hospital with a NICU far away relative to the median driving time *within* New York City. Driving time between hospitals is the relevant measure of proximity since the main mode of neonatal transport is ground ambulance (Ohning, 2015). Specifically, I compute driving time using Google Map APIs from each birth hospital to the nearest hospital with a NICU.

Table 1.5 shows that even within New York City, the reduction in length of stay and the increase in the probability of transfer are driven by hospitals with shorter driving time to the nearest hospital with a NICU. This suggests that proximity to a potential destination hospital plays an important role in birth hospitals' decision-making process. Given the longer driving distance between hospitals outside of New York City, transfer decisions might depend less on financial incentives but more on medical needs, which are unlikely to change discontinuously at the threshold.

This finding suggests that hospitals engage in profit-seeking behavior in response to financial incentives associated with MMC, but only when they can minimize the potential harm and costs through expedient transfer to a high-quality hospital. This finding is consistent with the growing literature that documents that health care providers respond to financial incentives but they are not willing to sacrifice the health of their patients in doing so (Ho and Pakes, 2014).

1.7 Heterogeneity in New York City

To further understand how hospitals respond to MMC in New York City, I conduct three heterogeneity analyses. In Section 1.7.1, I examine the role of capacity at birth hospitals. Section 1.7.2 examines the role of capacity at potential receiving hospitals. In Section 1.7.3, I examine predicted list prices of newborns to evaluate whether hospitals are especially responsive to infants who are costly to treat.

1.7.1 Capacity at Birth Hospitals

Here, I further explore hospitals' incentives to transfer away infants with less generous payments. Suppose that the number of NICU beds is fixed, and the hospital decides whether to retain a low birth weight infant at its own NICU facility or to transfer the infant to another hospital following birth. Although entering the NICU market has a large fixed cost, marginal costs of providing neonatal intensive care is relatively low. Therefore, the hospital has an incentive to utilize empty beds.[25] That is, as long as the reimbursement payments are higher than the relatively moderate marginal costs, the hospital can increase its profits by retaining infants enrolled in both MMC and FFS. When the hospital is spatially constrained, however, the hospital can benefit more from holding onto infants enrolled in FFS than those enrolled in MMC. Therefore, incentives to transfer infants enrolled in MMC are likely pronounced when the hospital has few NICU beds available.

To test this hypothesis, I exploit variation in monthly NICU utilization. Specifically, I define the NICU occupancy in a given month as the number of infants admitted last month and stayed in a NICU facility for at least 10 days.[26] I use the number of infants admitted last month to avoid counting the endogenous number of NICU stays in the contemporaneous month as a measure of how crowded NICU is. To ensure that infants who leave the hospital soon after birth are not included in the occupancy measure, I restrict length of stay to be at least 10 days. Given that the mean length of stay for very low birth weight infants is longer than a month, 10 days is unlikely to be a binding restriction.

I compare months when the NICU occupancy is below the median with months when the NICU occupancy is above the median at a given hospital in a given year. Within hospital-year

[25] Freedman (2016) tests this hypothesis and finds that empty beds increase NICU utilization.

[26] Appendix Figure C.6 plots this NICU occupancy measure for each month for an example hospital in a given year. It shows that there is large variation in NICU utilization across months.

comparisons ensure that the comparison is made at fixed capacity since the number of NICU beds is unlikely to change dramatically for a given hospital in a given year. The results are shown in panels A and B of Table 1.6. When the NICU occupancy is above the median, the reduction in length of stay, total charges, and total costs are large and significant around 20%; and the probability of transfer also increases by 4 percentage points. When the NICU occupancy is below the median (i.e., hospitals have enough number of beds), I find little impact of MMC on all outcomes, consistent with the spatial constraint playing an important role.

Since the NICU occupancy at the month level[27] cannot directly be compared to the number of NICU beds, high NICU occupancy may not indicate that the hospital is close to capacity. To address this issue, I create a crowdedness measure that is relative to hospital capacity. The mean length of stay for infants who stayed in a NICU facility for at least 10 days is 34 days. Thus, dividing the NICU occupancy, which is computed at the month level, by the number of beds yields a crude measure of the daily NICU occupancy rate. I compare below- and above-median months using this measure and find similar results (Appendix Table D.2). This supports the above finding that hospitals' incentives become stronger when they are spatially constrained.

1.7.2 Capacity at Potential Receiving Hospitals

Since hospitals have a financial incentive to utilize empty beds, I examine the role of crowdedness at potential destination hospitals. I consider two types of potential destination hospitals: (1) the nearest hospital with a NICU facility following Section 1.6.5; and (2) a "typical destination" hospital, which I define as the receiving hospital of the majority of (any) neonatal transfers from a given hospital.[28]

[27] I observe the admission month and the discharge month but do not observe the exact date of admission or discharge. Due to this data limitation, I am unable to identify exactly how many of NICU beds are occupied on a given day.

[28] In my sample, around 32% of total transfers occur to the nearest hospital with a NICU; and around 51% of total transfers end up at typical destination hospitals.

As in Section 1.7.1, I use the NICU occupancy to measure how crowded the potential destination hospital is. Table 1.7 shows that the effects are stronger in months when the nearest hospital with a NICU is relatively less crowded. Similarly, I find that the birth hospital is more likely to differentially treat infants across the threshold when its typical destination is relatively less crowded (Appendix Table D.3). This suggests that MMC may have induced hospitals to engage in reallocation of at-risk infants from a crowded hospital to a less crowded hospital via transfers.

In addition to the incentive to utilize empty beds, receiving hospitals with high-quality may have another incentive to accept the transferred infants. When health plans and hospitals negotiate over hospital payments for Medicaid patients, hospital quality plays a crucial role in determining the bargaining power of hospitals. That is, higher-quality hospitals likely have more bargaining power and thus command higher prices (Gaynor et al., 2015). In my sample, receiving hospitals are generally bigger and better-equipped, suggesting that they may face relatively modest incentives to differentially treat infants enrolled in FFS versus MMC.

1.7.3 Expected Costs of Treatment

In this section, I examine which group of infants is most affected by hospitals' financial incentives. Unless the reimbursement payments are perfectly adjusted for severity, infants with high predicted costs of treatment are especially costly to hospitals. Therefore, profit-maximizing hospitals are more likely to respond to infants whose marginal costs are high. To test this hypothesis, I create a measure of predicted costs of treatment. Specifically, I compute predicted list prices by regressing total charges on principal diagnosis fixed effects and principal procedure fixed effects. This measure thus estimates the expected total charges solely based on the severity of patients' conditions.

Consistent with the hypothesis, I find that hospital responses are stronger for infants with higher predicted list prices (Table 1.8). For infants with below-median predicted list prices, MMC reimbursement payments may still exceed the marginal costs and hospitals are unlikely to treat infants differentially across the threshold on the extensive margin (i.e., the retention versus transfer margin). For infants with above-median predicted list prices, the lower reimbursement payments under MMC may not cover the expected costs of treatment for these infants and thus birth hospitals are more likely transfer out infants above the threshold. Consequently, infants with severe conditions may be transferred to higher-quality hospitals, which suggests a potential improvement in the match between the patient and hospital.

For infants with above-median predicted list prices, however, I find that mortality during hospitalization at birth hospitals increases above the threshold and the estimate is marginally significant at the 10% level. This suggests that hospitals may shift resources towards infants under FFS with higher reimbursement payments, resulting in harming health among the most high-risk subpopulations under MMC. When I follow the patients over time, the individual-level mortality during hospitalization for this subgroup is still large, although insignificant (RD estimate for individual-level mortality: 0.032; robust standard error: 0.023). Albeit with limited precision, this suggests that MMC may adversely affect health for infants with the most severe conditions.

1.8 Specification and Robustness Checks

As a specification check, I test whether the estimates are robust to the choice of bandwidth and the degree of polynomials. I repeat the estimations varying bandwidths from 100 grams to 500 grams in 50-gram increments for each outcome. I use quadratic and cubic polynomials in addition to the linear polynomial to control for trends in birth weight. Appendix Figure C.7 shows the RD estimates by bandwidth for different degrees of polynomials. Overall, all panels show that the RD estimates are not sensitive to the choice of bandwidth and the degree of polynomials. In particular, the estimates for log(length of stay) and the probability of transfer are stable across different choices of bandwidths and polynomials, supporting my main specification.

One issue associated with identification using the birth weight threshold at 1,200 grams is that it coincides with one of the conditions that qualify children for the Supplemental Security Income

(SSI) program, which provides monthly cash payments and Medicaid to beneficiaries. However, I argue that SSI participation is likely to have a limited impact on medical care of newborns. First, monthly cash payments are unlikely to affect families' health care utilization conditional on Medicaid participation. When the child is in a medical facility, monthly cash payments are limited to $30. Since the amount of cash payments is fairly small and services provided to newborns enrolled in Medicaid are exempt from copayment, SSI payments are unlikely to alter families' incentives to utilize health care conditional on Medicaid participation. Additionally, the average monthly benefit for children was $633 in December 2014 (Duggan et al., 2015). Given the substantial amount of income transfer low-income families can expect outside of a medical facility, there may be an incentive for families to leave the facility early. However, this would go *against* finding a reduction in length of stay above the threshold.

If SSI participation based on the birth weight qualification induces people to participate in Medicaid who otherwise would not, it can affect both families and health care providers by substantially changing the cost of health care services. I examine whether the probability of receiving Medicaid discontinuously increases below the threshold. I find that the probability of Medicaid participation is in fact higher above the threshold and the estimate is not statistically significant (RD estimate: 0.024; robust standard error: 0.023). Little impact on Medicaid participation is likely due to a high baseline insured rate among very low birth weight infants, independent of SSI participation. Given the high costs of treatment, hospitals have a strong incentive to enroll all infants who qualify for a public health insurance program, if they do not already have one through the mother. This finding suggests that SSI has limited impacts on medical care of newborns around the 1,200-gram threshold.

Nevertheless, I conduct two exercises to test whether my results are robust to SSI participation. First, I repeat the estimations for two other states (New Jersey and Maryland) over the same period where the federal SSI rule applies but the exclusion from MMC does not, and I find no effects on discharge outcomes for this sample (panel A of Table 1.9).[29] This suggests that SSI has little impact on my findings. Second, I use the inclusion of infants weighing less than 1,200 grams into mandatory MMC enrollment in April 2012 to test the robustness of my results. I repeat my estimations using the discharge records of infants born after April 2012 in New York City and I find no effects on discharge outcomes during this period (panel B of Table 1.9),[30] suggesting that my results are not driven by something other than the exclusion from MMC.

1.9 Difference-in-Difference Estimation

In this section, I employ a difference-in-difference approach using the MMC mandate rollout across counties in New York State. The mandate was phased in starting October 1997 and was fully implemented in November 2012. To compare DD estimates with my RD estimates, I restrict the estimation up to 2011 since the exclusion of low birth weight infants was lifted in April 2012. Thus, the sample consists of inpatient visits of all newborns born between 1995-2011. In a DD framework, I estimate the effects of the MMC mandate on MMC participation and various discharge outcomes.[31] I report the coefficient of interest δ from the following regression:

$$Y_{ict} = \lambda_c + \gamma_t + \delta D_{ct} + \theta_c t + \epsilon_{ict} \tag{2}$$

where i denotes a discharge record, c denotes county, and t denotes year. I consider various outcomes Y_{ict} such as the probability of having Medicaid HMO as the primary expected payer, log(length of stay), log(total charges), log(total costs), the probability of transfer, and mortality during hospitalization. I include county fixed effects (λ_c) and year fixed effects (γ_t). D_{ct} indicates

[29] Additionally, I restrict the estimation to large urban areas in these two states and still find no differences below and above the threshold.

[30] I also use the periods before the mandate was introduced in New York City (prior to 1999) and find no differences at the threshold.

[31] Appendix Table **??** examines the change in sample composition in a DD framework. It shows that Medicaid participation decreases following the MMC mandate, suggesting that the introduction of MMC negatively affected the overall Medicaid enrollment.

the years after the mandate for each county. I include county-specific time trends ($\theta_c t$) in some specifications as a specification check. Standard errors are clustered at the county level.

Panel A of Table 1.10 shows the estimates from the baseline DD model excluding the county-specific time trends. The probability of participating in Medicaid HMO increases by 11 percentage points among infants following the mandate. This is smaller than the RD estimate which is around 23 percentage points, mainly due to heterogeneous compliance across counties. Column 2 shows that the DD estimate on length of stay is negative, but the magnitude is much smaller than my RD estimate. The DD estimates for total charges and total costs are negative and fairly close to my RD estimates. There is no change in the probability of transfer and mortality during hospitalization following the mandate in the whole sample of newborns.

As a check on the DD identification strategy, I estimate the model including the county-specific time trends. Panel B shows that including the time trends has little impact on the estimates, supporting the parallel trends assumption. Moreover, I employ an event study approach to examine pre-trends. Appendix Figure C.8 shows that there is little evidence of pre-trends in the probability of Medicaid HMO participation. These results suggest that differential time trends across counties are unlikely to drive my findings.

The comparison between the two sets of estimates emphasizes how hospital responses can vary across different subpopulations, suggesting that my RD estimates may have little external validity. To further understand the differences between the two models, I take two approaches. First, I repeat the DD estimations by birth weight groups in Section 1.9.1. Second, I compute and compare complier characteristics in Section 1.9.2.

1.9.1 Difference-in-Difference Estimation by Birth Weight Groups

To compare the DD estimates with my RD estimates for very low birth weight infants, I repeat the DD estimations (equation (4)) by birth weight groups. Given the small number of infants, I aggregate all infants weighing between 600 and 1,200 grams for the DD estimation below the threshold. Above the threshold, I repeat the estimation for each birth weight group in 150-gram increments. In addition, I repeat the RD estimations using 150 grams as the bandwidth for all outcomes and compare them with the DD estimates for infants whose birth weight is between 1,200 grams and 1,350 grams. In Figure 1.8, I plot the DD estimates for each birth weight group along with 95% confidence intervals. I plot the RD estimates along with 95% confidence intervals from New York City in 2003-2011 in red bars for the 1,200-1,350 gram bin.

Panel (a) of Figure 1.8 shows that the probability of being enrolled in Medicaid HMO is not affected by the mandate for infants with birth weight below 1,200 grams, which confirms that the exclusion from the mandate is implemented well. The increase in the probability of having Medicaid HMO is around 7 percentage points for all birth weight groups above the threshold.

Panels (b)-(f) show that for infants with birth weight between 1,200 and 1,350 grams the DD estimates are similar to the RD estimates. The DD estimates are imprecise for these low birth weight infants, but the RD estimates for the 1,200-1,350 gram group are generally within the confidence intervals of the DD estimates. Since both DD and RD models identify the effects using infants with the same range of birth weight, the similarity between these estimates supports my main RD estimates.

The DD estimates for infants with higher birth weight suggest that hospitals do engage in some cost-reduction measures in response to the MMC mandate for infants across the whole range of birth weight, but potentially using different methods. Both total charges and total costs decline, while length of stay and the probability of transfer barely change following the mandate among heavier infants. This suggests that hospitals may achieve cost reductions for these infants by adjusting the amount of care on the intensive margin (i.e., conditional on retaining at birth hospitals). Specifically, I consider other measures of health and the quality of care as outcomes (Appendix Table D.4) and find reductions in the utilization of chest X-rays and ultrasounds. I also find suggestive evidence that the utilization of respiratory and speech therapy services declines following the MMC mandate.

1.9.2 Complier Characteristics

To further gain insights on the differences between the RD and DD estimates, I examine hospital and patient characteristics for "compliers" who comply with each of the two instruments and compare them to the overall characteristics. Compliers in my RD context refer to those who are induced to enroll in MMC due to exceeding the 1,200-gram threshold. Compliers under the DD specification are those who are induced to enroll in MMC due to county-level rollout of the MMC mandate. It is impossible to identify compliers since counterfactual outcomes are not observable, but it is possible to describe the distribution of their characteristics (Abadie, 2003). I compute mean characteristics of the compliers following Angrist and Pischke (2009) and Almond and Doyle (2011).[32] Refer to Appendix Section B for details.

Table 1.11 presents the mean complier characteristics for both RD and DD samples. Panel A summarizes hospital characteristics and panel B compares patient characteristics. Column 1 shows the complier mean for the RD framework in the estimation window using the bandwidth of 150 grams, while column 2 shows the overall mean characteristics within the estimation window. Column 3 shows the complier mean for the DD specification, and column 4 shows the full sample mean of all infants.

Comparing columns 1 and 2 in panel A, compliers and the overall sample within the RD estimation window are relatively similar regarding the number of beds, staff, and admissions. A few notable differences, however, include the number of lives covered in capitated services arrangement and the share of infants covered by Medicaid. I use the 1995 values (before the mandate was in place) for the capitated lives covered since compliers by definition have more patients covered in a capitated payment structure contemporaneously. The number of lives covered in capitated services arrangement is lower for compliers than for the overall sample within the estimation window. This suggests that hospitals who previously served fewer patients covered in capitation were more

[32] See also Kim and Lee (2016).

compliant to the birth weight exclusion, which is as expected since more patients in these hospitals were *induced* to enroll in MMC following the mandate compared to patients in hospitals with high baseline participation in some capitated services.

In addition, compliers tend to stay in hospitals that serve more infants covered by Medicaid. This could be the case if hospitals with a high volume of Medicaid infants are more aware of the policy and thus more compliant to it. Moreover, assuming there is a cost associated with treating Medicaid managed care patients differently from traditional Medicaid patients (e.g., hiring a managed care manager), hospitals might do so only when there are enough number of patients affected by the adoption of MMC. Panel B shows that compliers are likely less advantaged subgroups. They are more likely to be racial minorities, and they tend to live in zip codes in the bottom quartile of the median income distribution. Consistent with this finding, Appendix Table D.5 shows that the effects are driven by counties with the lowest median household income where the share of Medicaid participation is likely high.

Similar to the compliers in the RD framework, column 3 shows that hospitals that comply with the MMC mandate have fewer lives covered in capitated services arrangement and more infants covered by Medicaid compared to the full sample. The DD compliers are also more likely to be racial minorities and poor compared to the full sample. However, compliers in the DD framework are different in many dimensions from compliers in the RD framework. They have much higher birth weight and stay in hospitals that are less likely to have a NICU facility or to be a teaching hospital. They also tend to have fewer beds, staff, and patients compared to the RD compliers. This suggests that compliers in the DD framework stay in hospitals that may employ alternative methods in achieving cost savings. Consequently, the treatment effects likely vary across these two instruments, consistent with the differences between the RD and DD estimates.

1.10 Discussion

1.10.1 Conceptual Framework

I develop a simple framework of provider responses to FFS and MMC under the prospective payment system.[33] I discuss under which conditions the chosen level of quantity is likely lower for an infant enrolled in MMC than for an infant enrolled in FFS. Suppose that the hospital receives prospective payment for providing inpatient services to a given infant based on the infant's DRG. I define the hospital's profit to be revenue minus total costs.

$$\pi(q) = R - C(q) = a \cdot \omega - C(q) \tag{3}$$

where a denotes the hospital base payment and ω denotes the service intensity weight for DRG classification. Total costs depend on q, the quantity of inpatient services provided. Note that revenue is constant under the prospective payment system.

Intuitively, since the hospital revenue R does not depend on q, the hospital's choice of q would not change once R changes (from FFS to MMC). Here, I assume that the *physician* is the key decision-maker who chooses the level of services provided to the infant. Additionally, I assume that the physician's utility depends both on the hospital's profit and the benefits to the infant:

$$U(\pi(q), B(q))$$

where $B(q)$ denotes the infant's total benefits from hospitalization. Let $b(q)$ denote marginal benefit. The first order condition from the physician's utility maximization problem is as follows.

$$\frac{\partial U}{\partial \pi}\frac{d\pi}{dq} + \frac{\partial U}{\partial B}\frac{dB}{dq} = 0 \tag{4}$$

[33] I follow the basic setup from Ellis and McGuire (1986).

Using equation (3), equation (4) can be written as

$$\frac{\partial U/\partial B}{\partial U/\partial \pi} b(q) = c(q) \tag{5}$$

where $c(q) > 0$ denotes marginal cost. Equation (5) suggests that the physician chooses the level of quantity that sets the weighted marginal benefit to the infant equal to the marginal cost to the hospital. The weight $MRS_{B,\pi} = \frac{\partial U/\partial B}{\partial U/\partial \pi}$ measures the rate at which the physician is willing to trade off marginal profit to the hospital for marginal benefit to the patient. In other words, $MRS_{B,\pi}$ measures how much the physician values the benefits to the patient relative to the hospital profit. I consider two cases depending on whether $MRS_{B,\pi}$ is a function of π.

Case 1 *$MRS_{B,\pi}$ depends on π, i.e., there are "income effects" in the physician's preferences.*

In this case, the prospective payment amount a affects the choice of q. For instance, consider $U(\pi(q), B(q)) = \pi(q)B(q)$. Then the equation (5) becomes $\frac{a \cdot \omega - C(q)}{B(q)} b(q) = c(q)$. If a is lower for an MMC infant than for a FFS infant, the slope of the physician's indifference curve will be flatter and the chosen level of q will be lower for an MMC infant.

However, the physician's choice of q is unlikely to depend on the level of hospital revenue since the amount of care provided to a single infant would have a very small effect on the total revenue of the hospital. I consider a case where the physician's choice of q is independent of hospital revenue below.

Case 2 *$MRS_{B,\pi}$ does not depend on π, i.e., there are no "income effects" in the physician's preferences.*

Since the slope of the physician's indifference curve does not depend on π, the amount of prospective payment does not affect the choice of q. For instance, consider $U(\pi(q), B(q)) = \pi(q) + B(q)^2$. Then the equation (5) becomes $2B(q)b(q) = c(q)$. Under this scenario, this simple model predicts that infants enrolled in both FFS and MMC will receive the *same* amount of care even with a different level of prospective payment.

Case 2 cannot explain my empirical result showing that hospitals provide less care to infants enrolled in MMC than to infants enrolled in FFS. This suggests the need for a theoretical model that would rationalize how the chosen level of inpatient services could be different between MMC and FFS when both systems use the prospective payment system.[34]

1.10.2 Cost Implications

In the New York City sample, I find that the overall costs aggregated at the individual level drop by 9% In the New York City sample, I find that the overall costs aggregated at the individual level drop by 9% for very low birth weight infants according to my preferred specification with hospital fixed effects (panel C of Table 1.3). This amounts to an average reduction of \$8,764 (=0.093×\$94,237) for an infant right below the threshold in 2011 values. For average infants, I find that the overall costs aggregated at the individual level decline by 6% (panel B of Table 1.10). This amounts to an average reduction of \$214 (=0.062×\$3,446). Note, however, that total costs are not actual payments made by insurers. With the caveat that the reduction in total costs may not translate into actual savings in health care spending and that the cost estimates are based on a particular sample, this suggests that hospitals indeed provide the less amount of care to infants enrolled in MMC.

The effect on individual-level mortality is positive but imprecisely estimated with the 95% confidence interval [-0.014, 0.048]. Given the wide confidence interval, it is hard to draw a conclusion on the value of a statistical life. When evaluated at the mean effect, the implied cost of saving a statistical life is \$515,529 (=\$8,764/0.017), which is fairly close to the estimate of \$550,000 (in 2006 dollars) for newborns with birth weight near 1,500 grams from Almond et al. (2010). Limited

[34] Competitive pressure from health plans may lead to a reduction in q for infants enrolled in MMC under two assumptions. First, bargaining between health plans and hospitals induces hospitals to choose physicians who provide less care. Second, those physicians differentially provide less care to infants enrolled in MMC than to infants enrolled in FFS. However, both of these two assumptions require testing and validation.

precision on health measures, however, suggests that the reduction in costs due to MMC may be efficient as it is achieved without harming health.

However, the current study has a few limitations in conducting a complete cost-benefit analysis. The health measures I examine are limited and imperfect as I only observe extreme measures such as death during hospitalization. I do not observe death or other health care utilization outside of the inpatient setting (e.g., outpatient visits).[35] In addition, there may be other forms of "costs" besides health consequences such as non-medical costs to hospitals (e.g., lawsuits) and parental disutility from separation/transfer, which I do not observe. For example, neonatal transfers can cause enormous stress and anxiety to parents (Hawthorne and Killen, 2006).

1.11 Conclusion

Recognizing limitations of the FFS system, the US health care market has increasingly adopted new payment systems that promote more efficient delivery of health care. These new systems are generally designed to reward improvement in the quality of care without unnecessarily increasing costs (Hackbarth et al., 2008; Arrow et al., 2009). Notably, the Affordable Care Act introduced accountable care organizations (ACOs) for Medicare populations that share similar incentives and goals as managed care organizations under Medicaid. This paper provides important implications for hospital responses to these incentives.

My findings suggest that hospitals respond to financial incentives stemming from different reimbursement models by adjusting their practice style. Hospitals reduce costs by transferring infants under MMC to other hospitals while holding onto infants enrolled in FFS. Hospital responses are particularly large when they are spatially constrained and for infants with high predicted list prices. However, I find no impact on hospital readmission and do not detect statistically significant impacts on mortality during hospitalization. Given my focus on very low birth weight infants, these findings suggest that MMC expansion to previously excluded high-cost and critically-ill subpopu-

[35] In future projects, I plan to examine the impact of MMC on outpatient and emergency department visits.

lations may be successful.

In addition, I find that the effects are driven by birth hospitals that have a hospital with a NICU nearby. This suggests that hospitals do not compromise the quality of care or patient health, by engaging in profit-seeking behavior only when they can minimize the potential harm through expedient coordination with a high-quality hospital. Notice that the *interaction* between financial incentives under MMC and short distances between local hospitals resulted in efficient delivery of MMC. This suggests that the structure of local health care markets may play an important role in successful delivery of a health care system, especially for critically ill patients who require coordination of care between local hospitals.

The overarching finding that MMC achieves cost reductions in ways that do not appear to compromise the quality of care is robust across the RD and DD models. This is surprising given the large differences in complier means between these two models, as shown in Table 1.11. There are two implications of this finding. First, my estimates are fairly representative and generalizable to the overall newborn population, as supported by the similarity between the DD complier mean and the overall sample mean. Second, even for the highest-risk infants, the RD results suggest a similar conclusion that costs go down while health does not seem to deteriorate. My finding of no adverse (postnatal) health effects, however, is in contrast to negative effects on prenatal care and worse birth outcomes found in Aizer et al. (2007). Whether there are differences between the response of prenatal versus postnatal care to MMC, both of which affect neonatal health but through distinct clinical channels, is an area for future research.

1.12 Figures

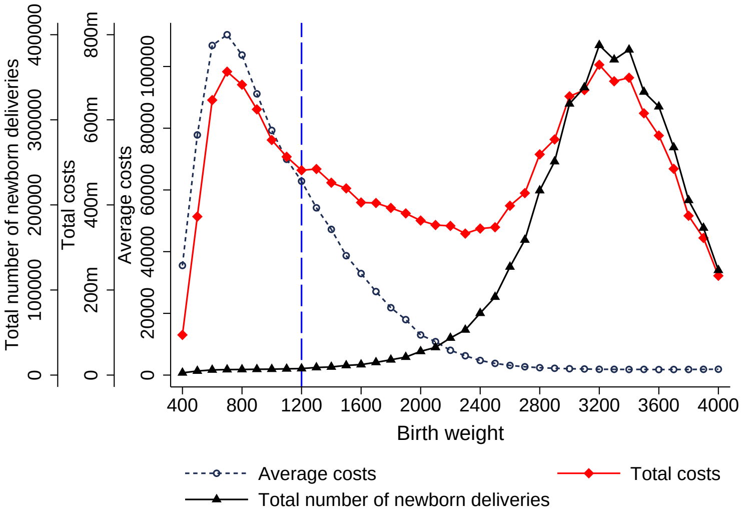


Figure 1.1: Average hospital costs and total discharges by birth weight, New York State, 1995-2013

Notes: Average costs are computed for each 100-gram bin using total charges multiplied by cost-to-charge ratio. The total number of discharges are computed for each 100-gram bin using the number of discharges with a birth weight record. Total costs are the product of these two: average costs times the total number of discharges.

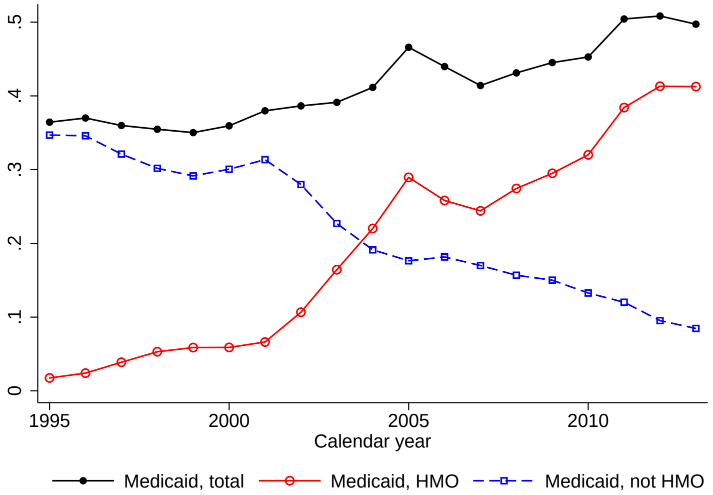
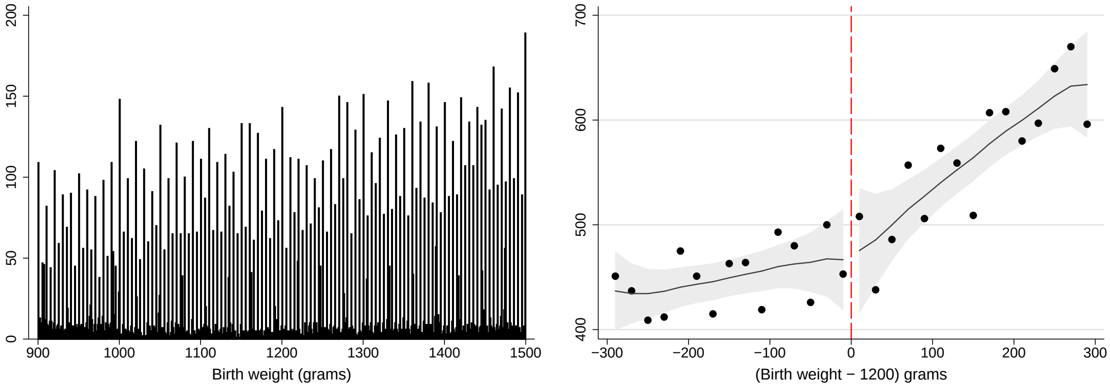


Figure 1.2: Share of infants covered by Medicaid, New York State, 1995-2013

Notes: HMO stands for Health Maintenance Organization, a type of managed care organizations (MCOs).

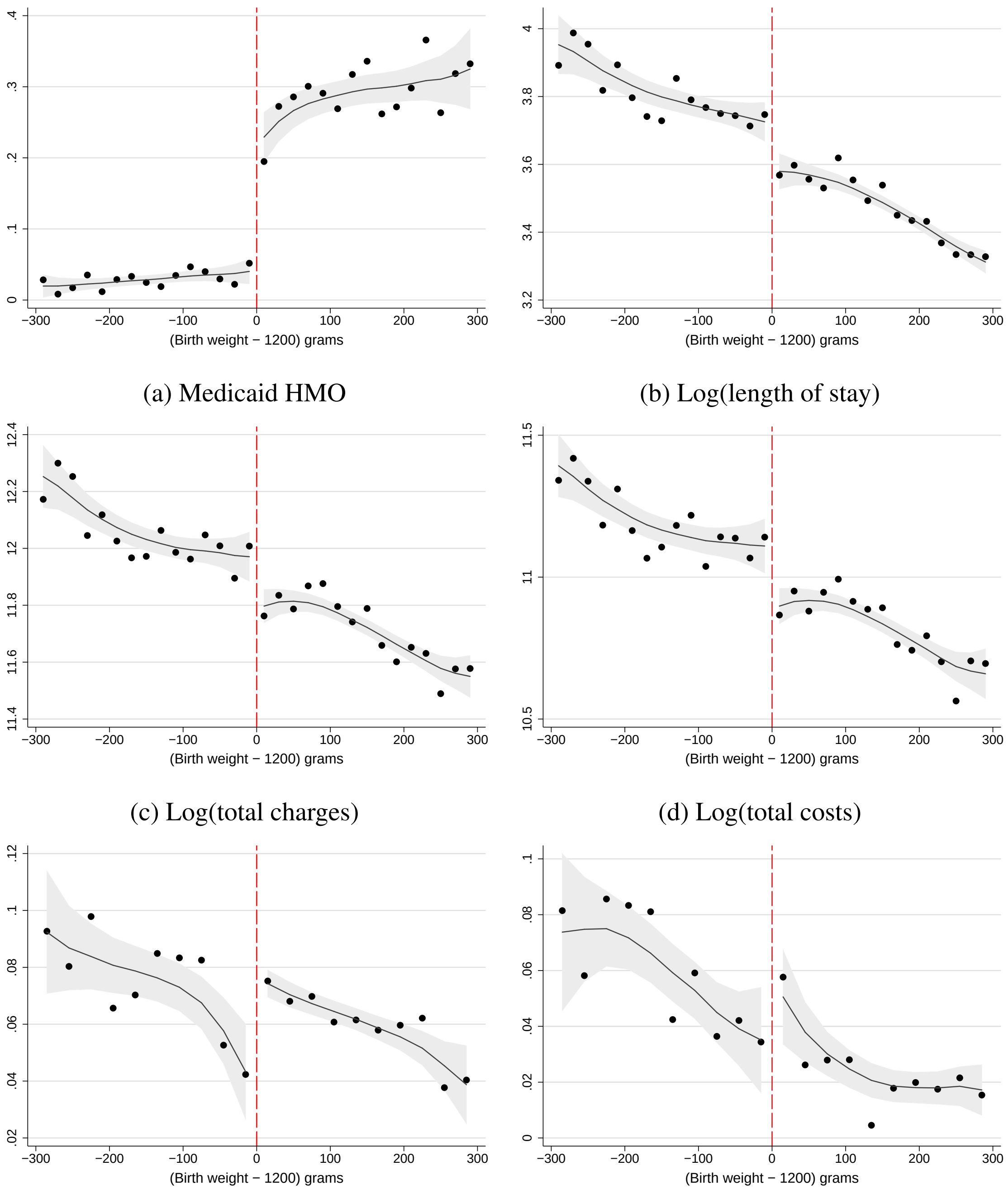


(a) Frequency for each gram

(b) Mean frequency for each 20-gram bin

Figure 1.3: Frequency of the running variable

Notes: Panel (a) plots the frequency of birth weight at each gram. Panel (b) plots mean frequency for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.



(a) Medicaid HMO

(b) Log(length of stay)

(c) Log(total charges)

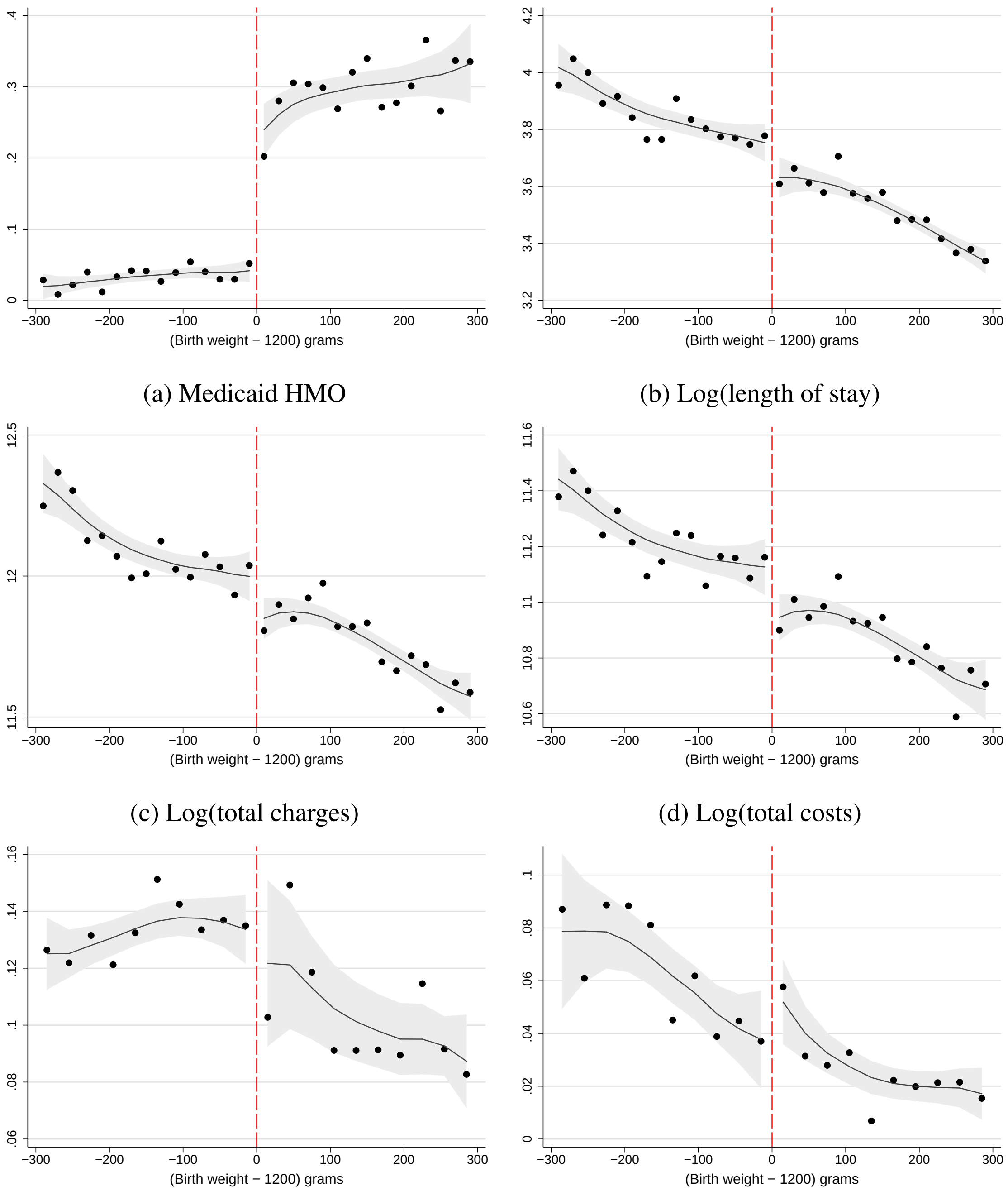
(d) Log(total costs)

(e) Transfer

(f) Mortality during hospitalization

Figure 1.4: Effects of birth weight≥1,200 grams on discharge outcomes at birth, New York City

Notes: Panels (a)-(d) plot mean values of each outcome variable for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold. For panels (e) and (f) I use a bigger 30-gram bin for better visibility since transfer and death are both rare events and thus noisy. I test whether using wider bins over-smooths the data following Lee and Lemieux (2010) but find no evidence of it. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.



(a) Medicaid HMO

(b) Log(length of stay)

(c) Log(total charges)

(d) Log(total costs)

(e) Readmission

(f) 1 year mortality during hospitalization

Figure 1.5: Effects of birth weight≥1,200 grams on cumulative outcomes, New York City

Notes: Each outcome aggregates the value at the individual level including the value at transferred hospitals (if transferred). Panels (a)-(d) plot mean values of each outcome variable for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold. For panels (e) and (f) I use a bigger 30-gram bin for better visibility since readmission and death are both rare events and thus noisy. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.

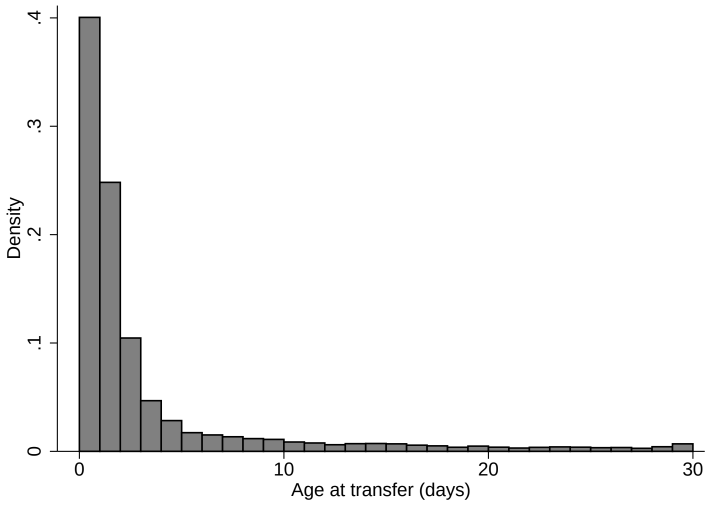


Figure 1.6: Age at transfer in the first month

Notes: 90% of neonatal transfers occur within the first month following birth. In particular, 70% of transfers occur within the first three days after birth.

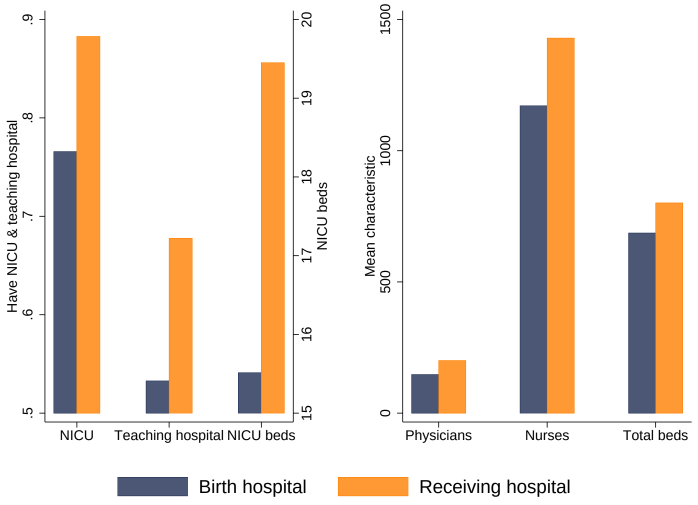
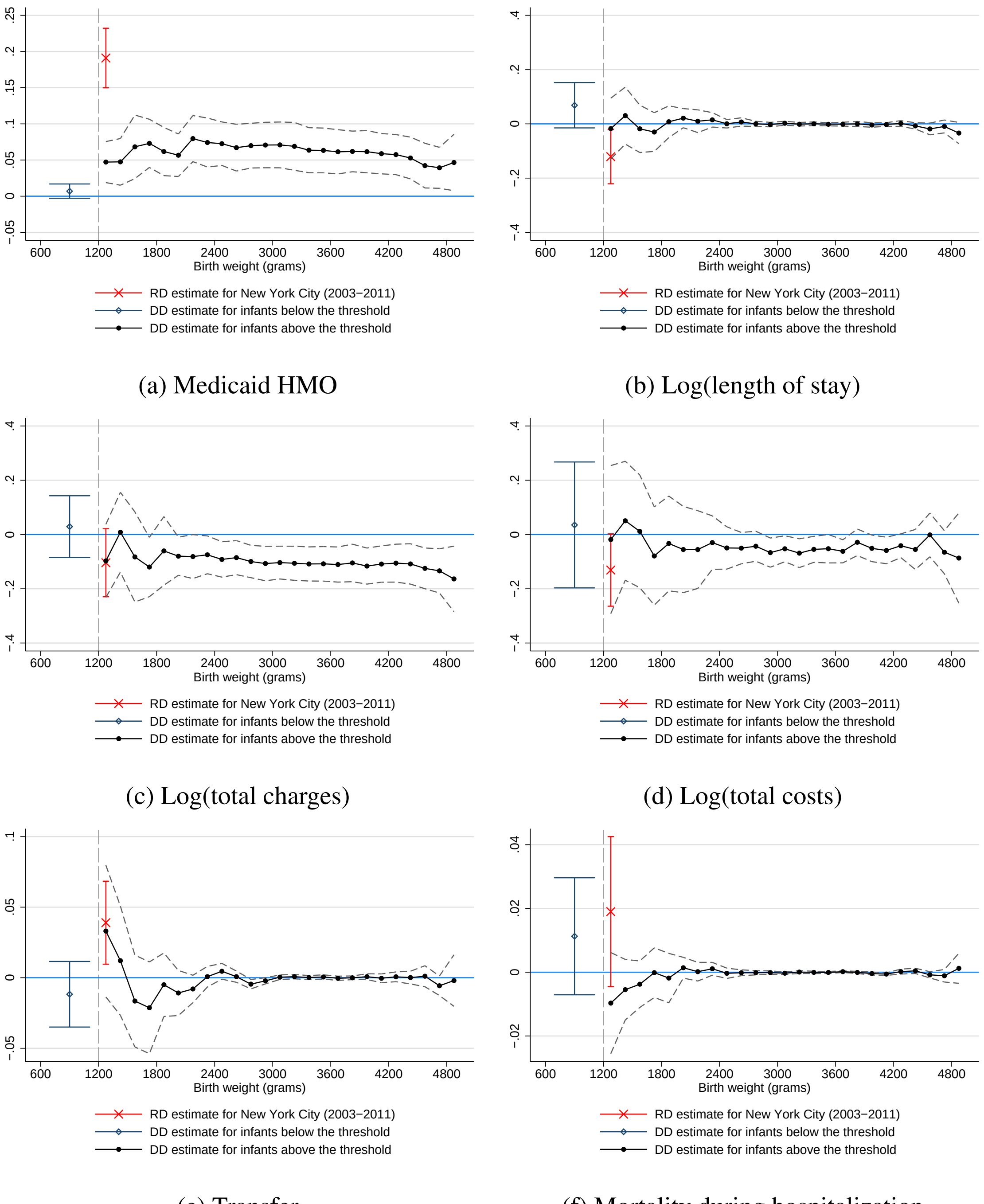


Figure 1.7: Characteristics of birth hospitals and receiving hospitals

Notes: Navy bars summarize mean characteristics of birth hospitals. Orange bars describe mean characteristics of hospitals that receive transfers.



(a) Medicaid HMO

(b) Log(length of stay)

(c) Log(total charges)

(d) Log(total costs)

(e) Transfer

(f) Mortality during hospitalization

Figure 1.8: Difference-in-difference estimates by birth weight

Notes: I estimate a difference-in-difference model by birth weight groups. Below the 1,200-gram threshold, I aggregate infants between 600 and 1,200 grams for precision and plot the difference-in-difference estimate in a navy bar. Above the 1,200-gram threshold, I plot the difference-in-difference estimates by birth weight groups in 150-gram increments (black). The estimates (solid lines) are plotted with 95% confidence intervals (dotted lines). The corresponding RD estimate for the New York City sample is shown in red (x) along with its 95% confidence intervals.

1.13 Tables

Table 1.1: Summary statistics, infants in New York State from 2003-2011

	(1)	(2)	(3)
		Near the 1,200-gram threshold	
	Full sample	Birth weight∈[900,1,200)	Birth weight∈[1,200,1,500]
Birth weight (grams)	3,273	1,050	1,357
Medicaid	0.427	0.544	0.508
Non-HMO	0.380	0.945	0.519
HMO	0.620	0.055	0.481
Total charges (USD)	$9,609	$204,796	$145,434
Total costs (USD)	$3,500	$75,758	$52,670
Length of stay (days)	3.710	46.370	33.016
Died during hospitalization	0.003	0.049	0.024
Subsequent visits	0.039	0.167	0.129
Transfers	0.010	0.127	0.107
NICU utilization	0.100	0.741	0.746
Observations	2001577	9076	11021

Notes: Total charges are list prices. Total costs are total charges multiplied by hospital-year-specific cost-to-charge ratios. Total charges and total costs are in 2011 values adjusted by CPI-U.

Table 1.2: Balance of covariates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Female	White	Black	Hispanic	Asian	Median income	Scheduled	Weekend
Panel A. Patient characteristics								
Birth weight≥1,200 g	-0.013	-0.023	0.026	0.015	-0.012	0.032	0.034	-0.007
	(0.018)	(0.021)	(0.018)	(0.016)	(0.009)	(0.059)	(0.028)	(0.018)
Observations	12701	7177	9636	7177	9636	4617	3357	10061
Mean below cutoff	0.497	0.355	0.316	0.145	0.054	2.353	0.713	0.261
Mean above cutoff	0.493	0.374	0.309	0.135	0.047	2.420	0.731	0.260
Bandwidth (grams)	250	150	200	150	200	150	150	200
	NICU	Teaching hospital	NICU beds	Physicians	Nurses	Total admissions	Total beds	Births
Panel B. Hospital characteristics								
Birth weight≥1,200 g	-0.002	0.009	-0.438	-7.346	-14.516	-560.159	-11.567	-97.628
	(0.009)	(0.014)	(0.586)	(11.330)	(35.162)	(814.747)	(18.180)	(98.570)
Observations	6278	10057	4184	7477	7477	7477	7477	7477
Mean below cutoff	0.955	0.715	20.7	180.1	1287.3	35357.2	753.1	4044.2
Mean above cutoff	0.945	0.698	20.4	187.4	1279.3	34946.2	732.3	3994.7
Bandwidth (grams)	150	200	100	150	150	150	150	150

Notes: Panel A shows the RD estimates for patient characteristics. Panel B shows the RD estimates for hospital characteristics. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.3: Effects of birth weight≥1,200 grams on discharge outcomes, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Discharge outcomes at birth hospitals						
Birth weight≥1,200 g	0.228***	-0.124**	-0.109*	-0.140**	0.024*	0.019
	(0.018)	(0.051)	(0.064)	(0.069)	(0.013)	(0.016)
Observations	5490	4065	4049	3096	5490	2735
Mean below cutoff	0.033	51.7	$244,943	$93,838	0.070	0.038
Mean above cutoff	0.277	42.0	$208,055	$77,391	0.065	0.037
Bandwidth (grams)	200	150	150	150	200	100
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Readmission	Mortality
Panel B. Aggregating at the individual level						
Birth weight≥1,200 g	0.236***	-0.089*	-0.072	-0.100	-0.000	0.015
	(0.018)	(0.049)	(0.062)	(0.067)	(0.021)	(0.016)
Observations	5490	4065	4047	3074	4065	2735
Mean below cutoff	0.039	53.2	$250,584	$95,366	0.140	0.040
Mean above cutoff	0.284	43.5	$215,080	$79,707	0.110	0.039
Bandwidth (grams)	200	150	150	150	200	100
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Readmission	Mortality
Panel C. Aggregating at the individual level, with hospital fixed effects						
Birth weight≥1,200 g	0.237***	-0.080*	-0.057	-0.090*	0.003	0.017
	(0.018)	(0.044)	(0.046)	(0.054)	(0.021)	(0.016)
Observations	5490	4065	4047	3074	4065	2735
Mean below cutoff	0.039	53.2	$250,584	$95,366	0.140	0.040
Mean above cutoff	0.284	43.5	$215,080	$79,707	0.110	0.039
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for each outcome from discharge records at birth hospitals. Panel B shows the RD estimates for outcome aggregated at the individual level. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Panel C additionally includes hospital fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.4: Effects of birth weight≥1,200 grams on discharge outcomes, rest of the state

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Discharge outcomes at birth hospitals						
Birth weight≥1,200 g	0.147***	0.021	0.039	0.051	0.011	0.021
	(0.018)	(0.065)	(0.073)	(0.074)	(0.019)	(0.014)
Observations	4571	3414	3407	3191	4571	2263
Mean below cutoff	0.032	49.1	$204,180	$75,151	0.149	0.030
Mean above cutoff	0.194	40.5	$167,210	$60,307	0.140	0.029
Bandwidth (grams)	200	150	150	150	200	100
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Readmission	Mortality
Panel B. Aggregating at the individual level						
Birth weight≥1,200 g	0.151***	0.041	0.057	0.072	-0.002	0.018
	(0.018)	(0.062)	(0.070)	(0.071)	(0.021)	(0.015)
Observations	4571	3414	3407	3174	3415	2263
Mean below cutoff	0.036	51.6	$212,942	$78,495	0.113	0.034
Mean above cutoff	0.204	42.4	$173,966	$63,014	0.093	0.030
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for each outcome from discharge records at birth hospitals. Panel B shows the RD estimates for outcome aggregated at the individual level. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.5: Heterogeneity by driving time to the nearest hospital with a NICU, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Below the median driving time						
Birth weight≥1,200 g	0.272***	-0.136*	-0.141	-0.109	0.041**	0.023
	(0.030)	(0.079)	(0.108)	(0.119)	(0.020)	(0.020)
Observations	2321	1713	1700	1230	2321	1158
Mean below cutoff	0.043	53.4	$287,628	$107,557	0.069	0.031
Mean above cutoff	0.324	43.3	$246,442	$87,755	0.079	0.028
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Above the median driving time						
Birth weight≥1,200 g	0.218***	-0.083	-0.077	-0.128	0.019	0.023
	(0.025)	(0.072)	(0.080)	(0.087)	(0.018)	(0.024)
Observations	2648	1962	1959	1486	2648	1312
Mean below cutoff	0.026	51.1	$200,293	$84,847	0.066	0.044
Mean above cutoff	0.258	41.5	$167,791	$70,934	0.055	0.040
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for hospitals whose driving time to the nearest hospital with a NICU is below the median, while panel B shows the RD estimates whose driving time is above the median. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.6: Heterogeneity by NICU crowdedness, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Below the median NICU occupancy						
Birth weight≥1,200 g	0.246***	-0.055	-0.043	-0.029	0.007	0.030
	(0.033)	(0.081)	(0.100)	(0.103)	(0.025)	(0.029)
Observations	1442	1063	1058	808	1442	724
Mean below cutoff	0.019	52.3	$268,717	$104,320	0.063	0.035
Mean above cutoff	0.255	43.4	$244,479	$89,590	0.052	0.032
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Above the median NICU occupancy						
Birth weight≥1,200 g	0.236***	-0.191**	-0.226**	-0.234**	0.037**	0.028
	(0.028)	(0.075)	(0.092)	(0.099)	(0.018)	(0.028)
Observations	2040	1513	1509	1121	2040	1010
Mean below cutoff	0.019	52.8	$275,354	$103,933	0.050	0.046
Mean above cutoff	0.285	42.5	$223,628	$84,551	0.053	0.038
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for months when the NICU occupancy is below the median for a given hospital in a given year. Panel B shows the RD estimates for relatively more crowded months when the NICU occupancy is above the median for a given hospital-year. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.7: Heterogeneity by crowdedness at the nearest hospital with a NICU, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Below the median NICU occupancy at the nearest hospital						
Birth weight≥1,200 g	0.232***	-0.180**	-0.227**	-0.300**	0.044**	0.029
	(0.030)	(0.084)	(0.104)	(0.119)	(0.021)	(0.027)
Observations	1846	1379	1373	995	1846	938
Mean below cutoff	0.023	52.0	$275,300	$111,706	0.062	0.046
Mean above cutoff	0.271	43.3	$237,916	$90,772	0.062	0.032
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Above the median NICU occupancy at the nearest hospital						
Birth weight≥1,200 g	0.286***	-0.151*	-0.099	-0.074	-0.019	0.022
	(0.038)	(0.079)	(0.107)	(0.114)	(0.024)	(0.037)
Observations	1284	932	928	668	1284	624
Mean below cutoff	0.024	54.0	$280,850	$108,544	0.074	0.034
Mean above cutoff	0.295	41.9	$225,560	$87,478	0.054	0.049
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for months when the NICU occupancy at the nearest hospital with a NICU is below the median, while panel B shows the RD estimates for months when the NICU occupancy at the nearest hospital with a NICU is above the median. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.8: Heterogeneity by predicted list prices, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Below the median predicted list prices						
Birth weight≥1,200 g	0.231^{***}	-0.025	0.050	-0.095	0.016	-0.002
	(0.029)	(0.063)	(0.085)	(0.084)	(0.017)	(0.017)
Observations	2226	1619	1619	1233	2226	1078
Mean below cutoff	0.034	47.9	$218,768	$86,188	0.054	0.019
Mean above cutoff	0.274	37.9	$174,036	$67,170	0.050	0.010
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Above the median predicted list prices						
Birth weight≥1,200 g	0.227^{***}	-0.167^{**}	-0.174^{**}	-0.111	0.035^{*}	0.038^{*}
	(0.023)	(0.070)	(0.086)	(0.092)	(0.019)	(0.022)
Observations	3202	2409	2393	1831	3202	1632
Mean below cutoff	0.031	54.1	$261,268	$98,819	0.076	0.048
Mean above cutoff	0.282	45.8	$237,610	$86,562	0.071	0.050
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for infants whose predicted list charges are below the median, while panel B shows the RD estimates for infants whose predicted list charges are above the median. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.9: Effects of birth weight≥1,200 grams on discharge outcomes, robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Hospitals in New Jersey and Maryland						
Birth weight≥1,200 g	0.030	0.018	0.032	0.071	0.001	0.008
	(0.023)	(0.081)	(0.088)	(0.095)	(0.021)	(0.015)
Observations	4755	3548	3542	3144	4755	2372
Mean below cutoff	0.206	43.0	$215,660	$58,100	0.151	0.031
Mean above cutoff	0.197	35.8	$177,229	$45,062	0.124	0.023
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Infants born after April 2012						
Birth weight≥1,200 g	0.109	0.022	0.018	0.209	-0.029	0.026
	(0.069)	(0.180)	(0.214)	(0.230)	(0.042)	(0.048)
Observations	900	669	669	554	900	438
Mean below cutoff	0.437	53.5	$427,550	$134,895	0.101	0.056
Mean above cutoff	0.503	42.3	$330,527	$109,927	0.055	0.042
Bandwidth (grams)	200	150	150	150	200	100

Notes: Panel A shows the RD estimates for each outcome at birth hospitals in New Jersey and Maryland from 2003-2011. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and state dummy for New Jersey. Panel B shows the RD estimates for each outcome at birth hospitals for infants admitted after April 2012. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.10: Difference-in-difference estimates

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Without county-specific time trends						
MMC mandate	0.111***	-0.009**	-0.075**	-0.095***	-0.000	-0.000
	(0.022)	(0.004)	(0.037)	(0.022)	(0.001)	(0.000)
Observations	4173544	4169319	4168406	2311157	3448242	4173535
Mean	0.170	3.8	$7,132	$3,446	0.011	0.004
Panel B. With county-specific time trends						
MMC mandate	0.065***	-0.000	-0.106***	-0.062**	-0.001	0.000
	(0.015)	(0.003)	(0.031)	(0.025)	(0.001)	(0.000)
Observations	4173544	4169319	4168406	2311157	3448242	4173535
Mean	0.170	3.8	$7,132	$3,446	0.011	0.004

Notes: Panel A presents a difference-in-difference estimate for each outcome without including the county-specific trends. Panel B shows the estimates including the county-specific trends. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 1.11: Mean complier characteristics

	(1)	(2)	(3)	(4)
	RD window [1050 g,1350 g]		DD	Full sample
	Complier mean	Overall mean	Complier mean	Overall mean
Panel A. Hospital characteristics				
Total beds	756.8	750.4	634.7	581.5
NICU beds	20.5	20.3	14.4	13.5
Number of physicians	188.7	184.6	148.5	127.9
Number of nurses	1286.7	1295.9	1093.0	845.0
Total admissions	35181.4	35480.7	31757.9	25590.2
Total births	3872.6	4028.5	3716.6	3145.5
NICU	0.92	0.94	0.81	0.72
Teaching hospital	0.70	0.70	0.55	0.49
Indigent care	0.72	0.71	0.77	0.63
Lives covered, capitated (*1995 values*)	7008.3	7177.0	5782.7	7413.5
Share covered by Medicaid, infants	0.57	0.47	0.59	0.40
Share covered by Medicaid, all patients	0.37	0.31	0.35	0.26
Share covered by HMO, infants	0.18	0.21	0.17	0.24
Share covered by HMO, all patients	0.21	0.22	0.21	0.20
Panel B. Patient characteristics				
Birth weight (grams)	1305.5	1204.5	3263.3	3287.4
Fraction low birth weight (<2,500 grams)	1.00	1.00	0.07	0.08
Female	0.50	0.48	0.48	0.48
White	0.21	0.36	0.26	0.51
Black	0.37	0.31	0.20	0.17
Hispanic	0.22	0.15	0.29	0.15
Asian	0.07	0.05	0.11	0.06
Median income, quartile 1	0.56	0.36	0.53	0.30
Median income, quartile 2	0.18	0.20	0.20	0.22
Median income, quartile 3	0.17	0.19	0.12	0.21
Median income, quartile 4	0.09	0.25	0.15	0.28
Admission scheduled	0.61	0.64	0.86	0.78
Admission on the weekend	0.24	0.25	0.23	0.22
Observations		8848		4173544

Notes: Column 1 presents mean characteristics of compliers within the RD estimation window. Column 2 shows the overall mean characteristics within the RD estimation window. Column 3 describes the complier mean for the DD specification. Column 4 shows the full sample mean. Median income is measured at the patient zip code level. I follow Angrist and Pischke (2009) and Almond and Doyle (2011) to compute complier characteristics. Refer to Appendix Section B for further details.

Chapter 2. Late-Career Job Loss and Retirement Behavior of Couples

2.1 Introduction

Evidence from different sources shows that couples coordinate the timing of retirement (Hurd 1990a; Blau 1998; Gustman and Steinmeier 2000; Michaud 2003; Gustman and Steinmeier 2004). A number of studies examine the mechanisms underlying the phenomenon of "joint retirement," with a focus on understanding how individuals near retirement age would behave in response to changes in the social security system. Existing literature finds that complementarities in tastes for leisure between spouses are important in explaining joint retirement.[36] Leisure complementarities exist when spouses enjoy retirement more when their partners are retired as well. However, financial capability for supporting retirement is another crucial channel that affects the retirement behavior of couples by changing households' budget constraints. In addition, examining the role of financial considerations would also provide insight into identifying the group of individuals who rely heavily on the social security system.

While the literature on couples' retirement accounts for the impact of complex financial incentives that stem from the social security system and pensions, it does not specifically examine the impact of unexpected shocks to earnings of individuals. This paper bridges this gap in the literature by assessing the effect of displacement - job loss due to business closings or layoffs - on retirement behavior of both spouses. There are two possible responses of spouses if their partners' job loss increases the probability of their retirement. To restore lost income, the spouse of a displaced worker may stay in the labor force longer (i.e., delay retirement) compared to the spouse of a non-displaced worker. On the contrary, the spouse might leave the labor force and retire early for the

[36] Given evidence of joint retirement, many studies have carefully modeled the environment in which couples jointly make employment decisions (Gustman and Steinmeier 2000; Gustman and Steinmeier 2004; Maestas 2001; Michaud 2003; Michaud and Vermeulen 2004; Blau and Gilleskie 2006; Van der Klaauw and Wolpin 2008; Casanova 2010). These studies focus on leisure complementarities between spouses as the key mechanism underlying joint retirement. Studies with reduced-form approaches also reach a similar conclusion showing that financial incentives alone cannot fully explain joint retirement decisions (Coile 2004; Banks et al. 2010).

sake of leisure complementarities. I test which of the two hypotheses dominates using the Health and Retirement Study (HRS). The retirement decision of the spouses of displaced workers can be different depending on the displaced workers' retirement status. For instance, the negative earnings shock to the household will be relatively modest if the displaced worker finds employment instead of choosing to retire. Therefore, I focus on how spouses respond only when their partners retire following job displacement.

This paper is related to the literature on the impact of job displacement. Displacement leads to long-term earnings losses, lower wealth holdings, lower employment rates, higher mortality, and decreased health insurance coverage (Ruhm 1991; Olson 1992; Jacobson et al. 1993; Chan and Stevens 1999; Chan and Stevens 2001; Munnell et al. 2006; Sullivan and Von Wachter 2009; Stevens and Moulton 2013). Moreover, Chan and Stevens (2004) find that job displacement increases the probability of retiring. Recent studies by Coile and Levine (2007, 2011) focus on the impact of recessions on older workers' retirement decisions and find that the probability of retirement increases in economic downturns. I extend this literature by considering the joint response of both spouses.

2.2 Research Design

2.2.1 Data

I use nine waves of the Health and Retirement Study (HRS) from 1994 to 2010. The HRS is a longitudinal panel data set that surveys a representative sample of individuals over the age of 50 and their spouses every two years. The HRS is ideal for studying retirement in a household context for at least three reasons. First, it surveys the relevant age group, 50 and older. Second, it tracks both spouses over time, a feature that is missing in most administrative data sets. Third, it contains detailed information such as employment, income and assets, Social Security and pension plans, and health status.

2.2.2 Sample Construction

I construct a sample that consists of full time workers who were between 50 and 70 years of age and were married when they first appeared in the survey. In total, there are 22,002 person-year observations. In terms of couples, there are 11,001 couple-year observations from 2,165 unique heterosexual couples, where both spouses meet the sample restriction. Using this sample of older working couples, I define the treatment (displaced) and the control (non-displaced) groups.

2.2.3 Definition of Job Loss

For the treatment group, I exploit extensive information on employment to identify individuals who have lost their jobs. Respondents are asked whether they are working at the same job as in the previous wave (i.e., two years ago). If they are no longer at the job, they are asked why they left the previous employer. I define displaced workers as those who stopped working because of business closings or layoffs. This is consistent with the definition commonly used in the literature on job loss (for example, see Chan and Stevens (1999, 2001)). These reasons are less likely to be correlated with worker characteristics that might affect retirement decisions.

Other reasons for job loss in the survey are poor health/disabled, family care, better job, quit, and retired. I exclude those who lost a job due to poor health/disabled as it is not the focus of the current study although it is an involuntary reason for job loss.[37] If an individual left employment voluntarily (i.e., due to family care, better job, or quit), I classify him as non-displaced. Hence, workers who stopped working for voluntary reasons form the control group in addition to employed workers.[38] Based on these rules, I construct a time-varying binary indicator for individuals who

[37] While retirement likely implies joint leisure for spouses whose partners retire after losing a job due to layoffs or plant closings, it might imply looking after the sick partners for spouses whose partners retire following a job loss due to poor health or disability. Thus, the spousal response can be different depending on the specific reason for job loss. That said, including job loss due to poor healthy/disabled as a source of job displacement in addition to layoffs and plant closings does not change the results.

[38] When I drop couples where either spouse stopped working due to voluntary reasons, I lose 2,739 couple-year observations. Excluding these couples does not affect the results; these results are available upon request. I am unable

reported job loss because of business closings or layoffs since the last interview. In the sample, the incidence of job loss due to business closings or layoffs during the sample period is around 5% (1,143 out of 22,002).

Figure 2.1 presents the average displacement rates from 1994 to 2010. The solid line indicates my preferred measure of job loss that includes both business closings and layoffs. To address that job loss may be endogenous, I use job loss due to business closings only as another measure of displacement. The dashed line shows the trend in average displacement rates using this measure. Both schedules evolve in a similar pattern over time. Since I restrict my sample to those working full time in 1994, job loss rates are the lowest in 1994 relative to the years that follow. There are discernible increases in job loss in 2002 and 2010, consistent with the recessions in early 2000s and in 2008-2009.

2.2.4 Definition of Retirement

I define retirement using self-reported current employment status. A person is retired if he is not working, not looking for a job, not temporarily laid off, not disabled, and not a homemaker. I allow for temporary retirement (i.e., a person can retire in one period and be working next period) because retirement is frequently not a complete exit from the labor force for many individuals. I check whether the results are robust to another measure of retirement, which indicates whether the individual considers herself completely retired. This measure additionally captures individuals' intentions to rejoin the labor force even when they are currently retired. The results are not sensitive to this different measure of retirement.

to conduct the analysis using those who lost a job due to voluntary reasons as the sole control group (i.e., excluding those who are employed), due to insufficient observations. Dropping couples where either spouse is working leaves 69 couples.

2.2.5 Summary Statistics

One of the most important determinants of retirement is financial incentives. The HRS contains information on pension coverage, income, and assets. In addition, I use information on self-reported health, age, health insurance, and other time-varying factors in the main estimation with individual fixed effects. Table 2.1 shows means of selected variables in 1994 separately by whether an individual is ever displaced (columns (1) and (2)) as well as by whether the individual's spouse is ever displaced (columns (4) and (5)). For instance, the first row in panel A summarizes average age for husbands who have ever been displaced in column (1); average age for husbands who have never been displaced in column (2); the difference between these two groups in column (3); average age for husbands whose wives have ever been displaced in column (4); average age for husbands whose wives have never been displaced in column (5); and the difference between the last two groups in column (6). Panel B reports the same set of summary statistics for wives. Estimates are all weighted using survey weights to account for the oversampling of blacks, Hispanics, and Florida residents. All monetary measures are in 2006 real values adjusted by CPI-U.

Columns (1)-(3) show that husbands who have ever been displaced were slightly older, less likely to have health insurance, and less likely to participate in pension programs than never-displaced workers. The annual earnings gap between the two groups is sizable, approximately $14,600 in 2006 dollars. Moreover, never-displaced workers had more financial wealth and more money in their Individual Retirement Account than ever-displaced workers. Ever-displaced wives display similar disadvantages compared to never-displaced wives. Columns (4)-(6) present a similar pattern in summary statistics by spousal displacement status.

2.2.6 Empirical Strategy

I first estimate the effect of job displacement on an individual's own retirement. Table 2.1 suggests that job displacement is not randomly assigned. To control for these differences in observable characteristics, I control for individual fixed effects (α_i) and time-varying characteristics (X_{it}) in the estimations described below.

$$R_{it} = \beta D_{it} + \delta' X_{it} + \alpha_i + \mu_t + e_{it} \tag{6}$$

R_{it} is an indicator for retirement status of person i in year t.[39] The impact of job loss on the individual's probability of retirement is captured by β, the coefficient on the indicator for job loss, D_{it}. D_{it} takes the value one if the individual i reports a job loss in year t. X_{it} includes time-varying covariates: age, age squared,[40] self-reported health, an indicator for having health insurance, lagged earnings, lagged financial and IRA wealth, and lagged indicators for defined benefit plan and defined contribution plan. I use lagged variables to take into account that these measures could have been affected by their current job status. For example, those who reported a job loss between the surveys might have depleted their wealth. The pension measures are also lagged because they tend to be tied to the individuals' current employment status. The results are not sensitive to excluding time-varying covariates. α_i is an individual fixed effect, which accounts for time-invariant characteristics of individuals. μ_t is a year effect that captures the general time pattern of retirement in the economy. I estimate the equation as a linear probability model separately for husbands and wives.[41] Standard errors in all specifications are clustered at the individual level to take into account serial correlation of retirement.

Then, I estimate the following equation:

$$R_{it} = \beta_1 D_{it} + \beta_2 SR_{it} + \beta_3 SD_{it} + \beta_4 SR_{it} * SD_{it} + \delta' X_{it} + \alpha_i + \mu_t + e_{it} \tag{7}$$

[39] Year t indicates the year of the interview. Respondents are asked in year t whether they have lost a job or have retired since the last wave in year $t-2$. Thus, the retirement and job loss variables in year t include those that happened in the past two years.

[40] Using dummies for each age instead of a quadratic function of age does not change the results.

[41] Fixed effects estimators of nonlinear models (e.g., logit or probit) can be severely biased due to the incidental parameters problem. See, for example, Lancaster (2000).

SR_{it} is a binary indicator if the spouse is retired in year t, and SD_{it} is an indicator that takes one if the spouse is displaced in year t. Thus, $SR_{it} * SD_{it}$ would take one if the spouse was displaced and retired in year t. β_4 captures the impact on an individual's probability of retirement when his partner retires following a job displacement.

To identify the main coefficient of interest β_4 in equation (2), both individual displacement and spousal displacement should be exogenous to individual retirement decisions. However, job loss may be endogenous as some unobservable characteristics of displaced workers can be correlated with a tendency to retire. For example, those who value leisure highly might shirk at work and also prefer to retire early. It is impossible to test this assumption, but I tackle this issue in two ways. First, I repeat the estimation using business closings as the only measure of individual and spousal displacement, as this measure is plausibly more exogenous. Second, displacement due to economy-wide shocks (e.g., recessions) is likely to be exogenous to worker characteristics. I examine the subgroup of workers who reported job loss in the 2002 and 2010 surveys, which cover the periods that correspond to recessions in early 2000s and 2008-2009.

Another challenge in estimating equation (2) is that spousal retirement is endogenous to their partners' own retirement. For example, husbands who retire following a job loss might do so only because they know that their wives have substantial earnings ability and thus will naturally stay in the labor force longer. To deal with this issue, I exploit the earliest age at which a person can claim Social Security benefits, which is 62 in the US. Social Security benefits that are available at age 62 are reduced when claimed before the full retirement age, and are increased by delayed retirement credits when received after the full retirement age, up to age 70. Social security benefits are designed to replace part of the employment earnings; the replacement rate ranged from 26% to 56% in 2013 depending on the worker's prior earnings level. The literature documents that financial incentives from Social Security benefits play an important role in determining the timing of retirement (Hurd 1990b; Anderson et al. 1999).

Specifically, I use an indicator for spousal age 62 and older, $A_{it} = I(\text{spousal age} \geq 62)$, as an instrument for spousal retirement, SR_{it}. Analogously, I use $A_{it} * SD_{it}$ as an instrument for the interaction term, $SR_{it} * SD_{it}$. This instrument is valid if the husband's incentive to retire at age 62 affects his wife's retirement decisions only because it increases the likelihood of his own retirement. Thus, the identification assumption is that the early entitlement age is an exogenous institutional feature that is not correlated with couples' characteristics. If a husband retires after reaching the age of 62 solely due to incentives created by the social security system, his retirement is exogenous to his wife's retirement decision. While it is not possible to test this exclusion restriction, I examine the distribution of age at retirement in my sample. Figure 2.2 shows disproportionately high number of people retiring at age 62 (the early entitlement age) and another modest spike at age 65 (the normal retirement age). This suggests that a substantial proportion of people are induced to time their retirement at the early entitlement age determined by the social security system. A number of papers use social security incentives as an instrument for retirement (Bound and Waidmann 2007; Banks et al. 2010; Rohwedder and Willis 2010; Mazzonna and Peracchi 2012).

2.3 Results

Based on the empirical strategy presented above, I estimate the effect of job displacement and retirement on couples both by OLS and 2SLS. Thereafter, I present the results of the event study approach and the heterogeneity analysis.

2.3.1 Main Results

Table 2.2 reports the results of OLS estimations separately for husbands (columns (1)-(4)) and wives (columns (5)-(8)). Columns (1) and (5) show the effect of individuals' job loss on their own retirement probability. Job loss significantly increases the probability of retirement for both husbands and wives.

Column (2) shows that wives' job displacement does not have a direct impact on their husbands' retirement. Column (6) shows that the effect of husbands' job displacement on wives' retirement is negative but not significant. The impact of spousal retirement is positive and significant, indicating a tendency of joint retirement as shown in columns (3) and (7). This suggests that couples value spending leisure together, and thus coordinate the timing of retirement. Moreover, notice that the coefficient is larger in magnitude for husbands than for wives. This is consistent with the literature that finds that husbands are more responsive to wives' retirement incentives than wives are (see, for example, Coile (2004)).

Columns (4) and (8) show how individuals' retirement decisions are affected by spousal job displacement and retirement status. The coefficient on the interaction term, spouse displaced and retired, is negative but insignificant for husbands. However, it is negative and significant for wives. That is, wives are 12 percentage points less likely to retire when their husbands retire following a job loss, and the magnitude is large enough to offset the increased probability of retirement due to their own job loss. This effect is more than half the mean (20.7%) and statistically significant at the 0.01 level. This is evidence that wives delay retirement when their husbands retire after losing a job. I interpret this as indicating that an unexpected job loss inducing displaced workers to leave the labor force involuntarily, and their spouses delaying retirement for financial support in response to that.

I repeat the estimation of equation (2) using instrumental variables to address the endogeneity of spousal retirement. Table 2.3 summarizes the first-stage regressions. Since there are two endogenous variables (SR_{it} and $SR_{it} * SD_{it}$), there are two first-stage regressions:

$$SR_{it} = \beta_1 D_{it} + \beta_2 A_{it} + \beta_3 SD_{it} + \beta_4 A_{it} * SD_{it} + \delta' X_{it} + \alpha_i + \mu_t + e_{it} \tag{8}$$

$$SR_{it} * SD_{it} = \beta_1 D_{it} + \beta_2 A_{it} + \beta_3 SD_{it} + \beta_4 A_{it} * SD_{it} + \delta' X_{it} + \alpha_i + \mu_t + e_{it} \tag{9}$$

Columns (1) and (2) show the regression results of equations (3) and (4) for husbands. Similarly, column (3) and (4) show the estimates of equations (3) and (4) for wives. First-stage F-statistics reported in the last row confirm that age 62 is a strong predictor of retirement.

Table 2.4 shows the results of the 2SLS regressions. The coefficients on the instrumented spousal retirement are positive and significant for both husbands and wives. Moreover, wives are less likely to retire in response to their displaced husbands' retirement, and the magnitude is larger than that from the OLS regression in column (8) of Table 2.2. The difference may reflect different populations captured by 2SLS and OLS. For instance, those who retire in accordance with Social Security (i.e., compliers) may be more credit-constrained, which can explain the larger 2SLS estimates. These results suggest that wives' retirement delay is not driven by their husbands' endogenous retirement.

In additional analyses, I use a more exogenous measure of displacement, which considers business closings as the only valid reason for job displacement (i.e., excluding layoffs). Wives tend to delay retirement in response to their husbands' job loss and retirement (OLS coefficient: -0.064 (0.061); 2SLS coefficient: -0.183 (0.120)). Although the estimates are not as precise due to low incidence of business closings (1.6%), this suggests that the results are not solely driven by those who were laid off. I also repeat the main estimations for the subgroup of workers who experienced a job loss during recessions in 2000-2001 or in 2008-2009, since displacement due to economy-wide shocks is likely to be exogenous to worker characteristics. Wives are 29 percentage points less likely to retire (OLS coefficient: -0.289 (0.371)), which is comparable to the 2SLS estimate from my preferred specification (from column (2) of Table 2.4: -0.271 (0.089)).

Moreover, I examine another measure of retirement, which indicates whether individuals consider themselves completely retired. OLS regressions suggest that both husbands and wives tend to delay retirement (husbands: -0.129 (0.058); wives: -0.142 (0.039)). 2SLS regressions indicate that the estimate for husbands is of the same magnitude but insignificant (-0.164 (0.123)) while the magnitude of the estimate for wives is larger and significant (-0.333 (0.113)).

2.3.2 Event Study Approach

To investigate how persistent the impacts of individuals' job loss and retirement are on their spouses' retirement, I employ an event study approach. An event study also serves as a way to test causal interpretations of the main results. For example, showing that a pre-trend in the retirement probability is not correlated with a future job displacement, while the likelihood of retirement changes sharply at the time of displacement, would support the notion that job displacement triggered the response in retirement decisions. This is a common approach taken in the literature in the analysis of job displacement (Jacobson et al., 1993). Specifically, I estimate the following equation:

$$R_{it} = \beta_1 D_{it} + \beta_2 A_{it} + \beta_3 SD_{it} + \sum_{-10}^{10} \beta_{4k} (SRSD)_{it}^{k} + \delta\, X_{it} + \alpha_i + \mu_t + e_{it} \qquad (10)$$

The dummy variables, $(SRSD)_{it}^{k}$ (k=-10, -8, ..., 0, ..., 8, 10) indicate k years before and after spousal retirement following job displacement (hereafter, the event). Since I observe up to 16 years before and after the event, $(SRSD)_{it}^{-10}$ takes one for 10 to 16 years before the event and zero otherwise. $(SRSD)_{it}^{10}$ is analogously defined for 10 to 16 years after the event. The omitted time period is 2 years before the event. Hence, the coefficients β_{4k} measure the change in the probability of retirement not only at the time of the event but also k years before and after, relative to the time period just before the event. I estimate this both by OLS and by 2SLS, instrumenting the spousal retirement variables with an indicator for spousal age greater than or equal to 62.

Figure 2.3 shows the dynamics of individuals' retirement probability after their spouses retire following a job loss. Panels (a) and (b) show the results of OLS regressions for husbands and wives, respectively. For panels (c) and (d), I use an indicator for spousal displacement after reaching 62 as an instrument for spousal retirement following displacement. Consistent with Tables 2.2 and 2.3, husbands do not immediately respond to their wives' retirement following a job loss, whereas wives do. Interestingly, husbands do seem to delay retirement four years later, which suggests that the earnings shock following their displaced wives' retirement might not take effect until a few years later.

Panel (d) is particularly informative - it presents a sharp drop in wives' probability of retirement when their husbands report a job loss and retire, and this increases steadily over time. The decline in the probability of retirement does not persist and is not statistically different from zero four years after spousal retirement following a job loss. This provides additional evidence that wives may be staying in the labor force to compensate for their husbands' involuntary retirement induced by job displacement.

An unexpected shock to one individual's wage income can have a significant impact on the retirement behavior of both spouses. Interestingly, however, only wives delay retirement in response to their husbands' job loss and retirement. This poses multiple hypotheses about the possible underlying mechanisms. Is a husband's job loss more critical to household finance than a wife's job loss, as husbands are more likely to be the primary earners? Do wives have a better chance of staying in the labor force, as they are generally younger? Or do wives who are younger than husbands need to spend a few more years working before being able to claim Social Security benefits? I investigate these hypotheses in the following section.

2.3.3 Heterogeneity Analysis

I examine heterogeneous effects by subgroups to disentangle potential mechanisms of the main effects. I repeat the main 2SLS estimations for different subgroups; Table 2.5 reports the coefficients on the interaction between spousal displacement and spousal retirement.

I consider individual age below or above 62 to test whether the effects vary depending on whether the spouse of the displaced worker is eligible for Social Security benefits. As shown in columns (2) and (3) of panel B, the delay of retirement is mostly driven by wives younger than 62 and thus not yet eligible for Social Security benefits. This is consistent with the hypothesis that a financial mechanism is at play. I also examine the effects by the level of total financial and IRA wealth to test the hypothesis that people who have secure retirement wealth would not necessarily respond to their partners' job loss and retirement. In columns (4)-(6) I divide the sample into three groups based on the sum of total financial and IRA wealth.[42] I find that the wives' response is the largest when couples were in the lowest quartile of the wealth distribution. This suggests that wives in the relatively poor household are most likely to delay retirement in response to their husbands' retirement following a job loss.

Finally, I examine whether the effects are heterogeneous depending on who assumes the role of the primary earner in the household. I define the primary earner as the spouse who contributes more than half of the total household earnings in the period preceding the job loss. Column (7) shows a significant drop in wives' probability of retiring when their husbands are the primary earners. If husbands who were primary earners lost a job and subsequently retired, it could cause a substantial loss in household income. This could have led their wives (i.e., the secondary earners) to delay retirement to compensate for their husbands' job loss. This might explain why I do not find any effect on the husbands' retirement decision when their wives lose a job and retire. As wives are likely to earn less than husbands in general (64% of the case in my sample), losing wives' income might not be as substantial a loss to the household.

2.4 Discussion and Conclusion

While the existing literature focuses on complementarities in tastes for leisure between spouses in examining retirement behavior couples, I examine the role of an unexpected earnings shock: a late-career job loss. I find that wives delay retirement when their husbands retire following a job loss, instead of seeking other employment. The decline in wives' probability of retiring persists for at least a couple of years following their husbands' job loss. Wives' job loss, however, does

[42] Financial wealth is calculated as the sum of dollar values of stock, bonds, and savings. The HRS also asks whether the couple has any money or assets held in an Individual Retirement Account (i.e., in an IRA or KEOGH account), and how much is in their accounts. The 25th percentile wealth was $10,000 and the 75th percentile wealth was $175,000 in 2006 dollars. The mean wealth was $184,700 in 2006 dollars.

not have a statistically significant impact on their husbands' retirement decision, though the point estimates are negative.

This evidence suggests that uncertainty in household income has a significant impact on how couples time their retirement. In addition, it shows how married men are privately insured by their wives, mitigating the impact of an unexpected earnings shock. However, selection into retirement is not random, which makes it difficult to interpret the effect of husbands' retirement on that of wives as causal. To address this issue, I use the early entitlement age for Social Security benefits, which is 62 in the US, as an instrument for retirement. I find that two stage least squares (2SLS) regressions yield the same conclusions.

Heterogeneity analysis reveals that the results appear stronger for subgroups that are relatively more credit-constrained. The drop in the probability of retirement is pronounced when couples are in the lowest quartile of the wealth distribution. In addition, I find that wives tend to delay retirement more when their husbands were the primary earners in the household. This is consistent with a story where the secondary earner tends to delay retirement to compensate for the primary earner's job loss. Moreover, I find that wives younger than 62 are much more likely to delay retirement in response to their partners' retirement following a job loss relative to wives older than 62, implying that Social Security benefits can help relax the household budget constraint and allows wives to join their husbands in retirement.

For future research, it would be useful to investigate how ever-married or single individuals cope with an unexpected job loss. For instance, single individuals might need to seek a different source of insurance following an unexpected earnings shock, due to a lack of within-household insurance. In addition, exploring policy tools that can help correct this potential discrepancy stemming from different family structures or marital status would be an important agenda for future research. For example, tagging current marital status besides just age when designing the social security system would be worth considering.

2.5 Figures

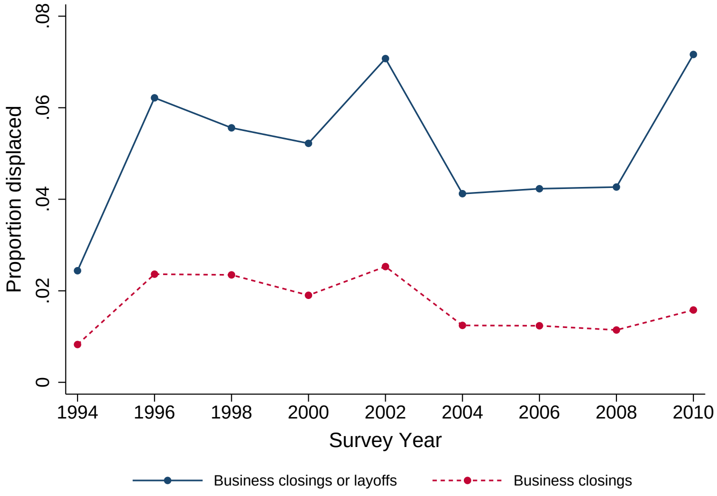


Figure 2.1: Incidence of Displacement over the Sample Period

Source: Author’s tabulations from the 1994-2010 Health and Retirement Study.

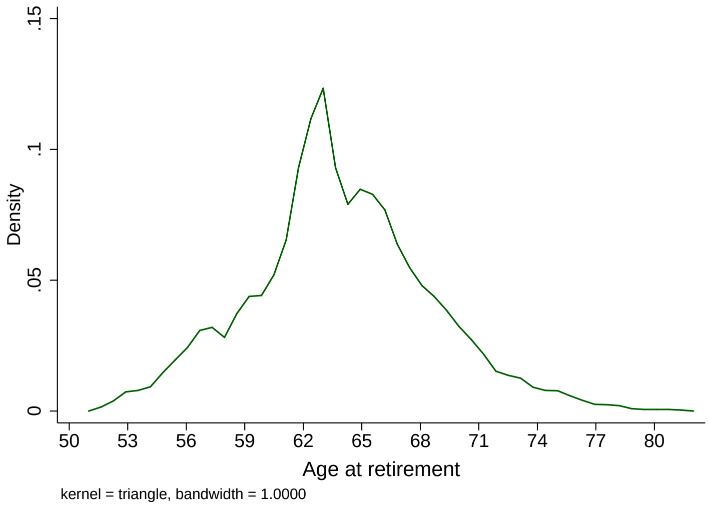
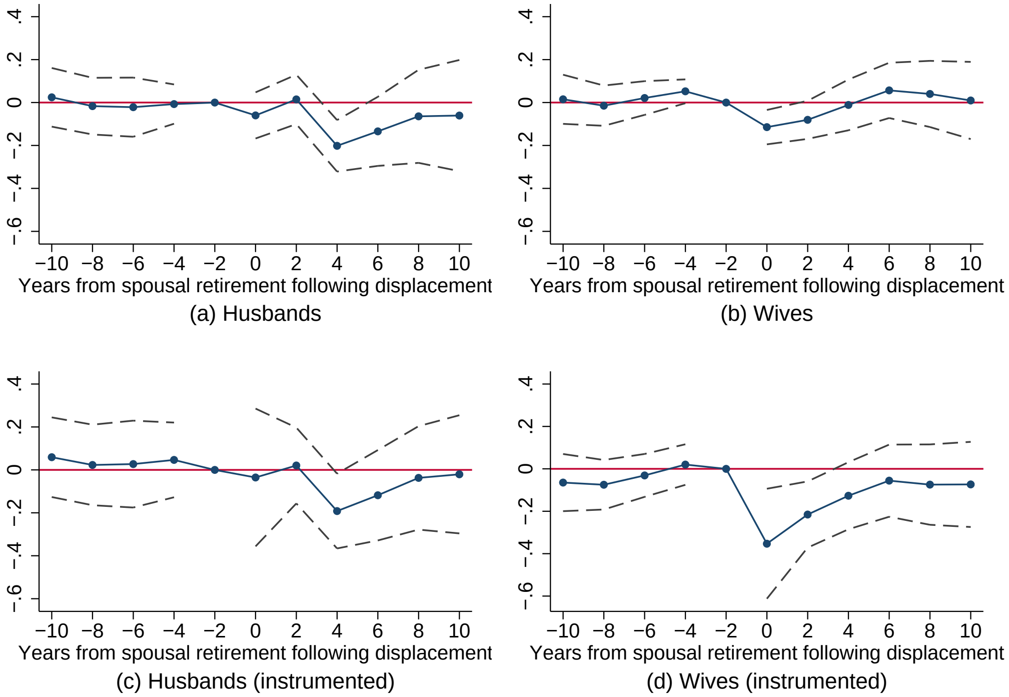


Figure 2.2: Retirement Hazard Rate

Source: Author's tabulations from the 1994-2010 Health and Retirement Study.
Notes: The figure reports the density of age at retirement in the sample. I define retirement using self-reported current employment status. Specifically, a person is retired if he is not working, not looking for a job, not temporarily laid off, not disabled, and not a homemaker. I use the triangular kernel and a bandwidth of 1.



(a) Husbands

(b) Wives

(c) Husbands (instrumented)

(d) Wives (instrumented)

Figure 2.3: Probability of Retirement by Years from the Displaced Spousal Retirement

Source: Author's tabulations from the 1994-2010 Health and Retirement Study.

Notes: Each figure plots point estimates from a regression of retirement status on a set of dummies that indicate years from spousal retirement after displacement. Individual fixed effects, year effects, and time-varying controls (age, age squared, self-reported health, an indicator for having health insurance, lagged earnings, lagged financial and IRA wealth, lagged indicators for defined benefit plan, and defined contribution plan) are also included in the regressions. The dashed lines plot 95-percent confidence intervals computed based on standard errors clustered at the individual level. The omitted time period is 2 years before the spousal post-displacement retirement. Panels (a) and (b) show results from OLS regressions. Panels (c) and (d) present 2SLS regressions using an indicator for spousal age greater than or equal to 62 to instrument for spousal retirement variables.

2.6 Tables

Table 2.1: Summary statistics

	(1)	(2)	(3)	(4)	(5)	(6)
	Ever	Never	Difference	Spouse ever	Spouse never	Difference
	displaced	displaced	(1)-(2)	displaced	displaced	(4)-(5)
Panel A. Husbands						
Age	61.79	61.43	0.36**	61.84	61.43	0.41***
			(0.15)			(0.16)
Education	13.45	13.39	0.06	13.31	13.43	-0.12**
			(0.04)			(0.05)
White	0.92	0.91	0.01**	0.92	0.91	0.01*
			(0.00)			(0.00)
Have health insurance	0.80	0.82	-0.02**	0.78	0.83	-0.04***
			(0.01)			(0.01)
Have pension	0.30	0.39	-0.09***	0.35	0.37	-0.02
			(0.01)			(0.01)
Earnings	$35,200	$49,800	-$14,600***	$42,400	$47,200	-$4,800*
			(1,500)			(2,500)
Financial wealth	$82,600	$108,700	-$26,100***	$78,600	$108,500	-$29,900***
			(7,400)			(7,300)
IRA wealth	$75,600	$98,500	-$22,900***	$71,800	$98,500	-$26,700***
			(7,900)			(7,800)
Observations	2742	8259	11001	2446	8555	11001
Panel B. Wives						
Age	59.49	58.96	0.53***	59.53	58.92	0.61***
			(0.15)			(0.14)
Education	13.23	13.43	-0.20***	13.38	13.39	-0.01
			(0.04)			(0.04)
White	0.93	0.91	0.01***	0.92	0.91	0.00
			(0.00)			(0.00)
Have health insurance	0.83	0.87	-0.04***	0.86	0.86	-0.00
			(0.01)			(0.01)
Have pension	0.31	0.44	-0.13***	0.42	0.41	0.01
			(0.01)			(0.01)
Earnings	$23,800	$31,800	-$8,000***	$27,300	$31,100	-$3,800***
			(1,100)			(1,000)
Financial wealth	$78,600	$108,500	-$29,900***	$82,600	$108,700	-$26,100***
			(7,300)			(7,400)
IRA wealth	$71,800	$98,500	-$26,700***	$75,600	$98,500	-$22,900***
			(7,800)			(7,900)
Observations	2446	8555	11001	2742	8259	11001

Source: Author's calculations from the 1994-2010 Health and Retirement Study.
Notes: All estimates are weighted with survey weights. All monetary variables are inflation-adjusted using 2006 CPI-U. Financial wealth is defined as the sum of dollar values of stock, bonds, and savings. Note that financial wealth and IRA wealth are measured at the household level. Standard errors are in parentheses.
* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 2.2: Effects of job loss on each spouse's retirement decision: OLS estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Husbands				Wives			
Displaced	0.143***	0.142***	0.148***	0.148***	0.069***	0.069***	0.068***	0.071***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)	(0.021)	(0.020)
Spouse displaced		0.006		-0.002		-0.021		-0.006
		(0.019)		(0.021)		(0.016)		(0.017)
Spouse retired			0.203***	0.206***			0.175***	0.191***
			(0.019)	(0.020)			(0.018)	(0.018)
Spouse displaced and retired				-0.029				-0.122***
				(0.047)				(0.038)
Observations	11001	11001	11001	11001	11001	11001	11001	11001
Mean retirement rate	0.265	0.265	0.265	0.265	0.207	0.207	0.207	0.207

Source: Author's estimations from the 1994-2010 Health and Retirement Study.

Notes: Standard errors in parentheses are clustered at the individual level. * Significant at 10%, ** significant at 5%, *** significant at 1%. All estimates are weighted with survey weights. Each regression contains individual fixed effects, year effects, and time-varying covariates (age, age squared, self-reported health, an indicator for having health insurance, lagged earnings, lagged financial and IRA wealth, lagged indicators for defined benefit plan, and defined contribution plan).

Table 2.3: Effects of job loss on each spouse's retirement decision: First-stage estimates

	(1)	(2)	(3)	(4)
	Husbands		Wives	
Dependent variable:	Spouse retired	Spouse displaced & retired	Spouse retired	Spouse displaced & retired
Displaced	-0.026	-0.004	-0.002	0.010
	(0.016)	(0.005)	(0.020)	(0.010)
Spouse displaced	0.039*	0.118***	0.080***	0.166***
	(0.020)	(0.024)	(0.025)	(0.031)
Spousal age≥62	0.110***	-0.009***	0.093***	-0.011***
	(0.018)	(0.003)	(0.017)	(0.004)
Spouse displaced and spousal age≥62	0.127**	0.388***	0.136***	0.382***
	(0.051)	(0.054)	(0.041)	(0.047)
Observations	11001	11001	11001	11001
Mean of dependent variable	0.207	0.011	0.265	0.019
F-statistic on the excluded instruments	30.8	41.4	28.3	55.6

Source: Author's estimations from the 1994-2010 Health and Retirement Study.

Notes: Standard errors in parentheses are clustered at the individual level. * Significant at 10%, ** significant at 5%, *** significant at 1%. All estimates are weighted with survey weights. Each regression includes individual fixed effects, year effects, and time-varying covariates (age, age squared, self-reported health, an indicator for having health insurance, lagged earnings, lagged financial and IRA wealth, lagged indicators for defined benefit plan, and defined contribution plan).

Table 2.4: Effects of job loss on each spouse's retirement decision: 2SLS estimates

	(1) Husbands	(2) Wives
2SLS using spousal age		
Displaced	0.151^{***}	0.074^{***}
	(0.020)	(0.018)
Spouse displaced	-0.005	0.016
	(0.028)	(0.027)
Spouse retired	0.312^{**}	0.394^{***}
	(0.134)	(0.126)
Spouse displaced and retired	-0.050	-0.271^{***}
	(0.108)	(0.089)
Observations	11001	11001
Mean retirement rate	0.265	0.207

Source: Author's estimations from the 1994-2010 Health and Retirement Study.
Notes: Standard errors in parentheses are clustered at the individual level. * Significant at 10%, ** significant at 5%, *** significant at 1%. All estimates are weighted with survey weights. Each regression includes individual fixed effects, year effects, and time-varying covariates (age, age squared, self-reported health, an indicator for having health insurance, lagged earnings, lagged financial and IRA wealth, lagged indicators for defined benefit plan, and defined contribution plan). I use an indicator for spousal age greater than or equal to 62 as an instrument for spousal retirement and the interaction between the indicator and spousal displacement as an instrument for spousal retirement interacted with spousal displacement.

Table 2.5: Heterogeneous effects on retirement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Full sample	Age		Wealth (percentile)			Primary earner	
		<62	≥62	<25th	25-75th	≥75th	Husband	Wife
Panel A. Husbands								
Spouse displaced and retired	-0.050	-0.259	0.036	0.045	-0.174	0.080	-0.159	0.078
	(0.108)	(0.400)	(0.233)	(0.229)	(0.147)	(0.308)	(0.173)	(0.380)
Observations	11001	4682	5794	2355	5335	2601	4045	2075
Mean retirement rate	0.265	0.066	0.428	0.232	0.246	0.333	0.173	0.385
Panel B. Wives								
Spouse displaced and retired	-0.271***	-0.266**	0.583	-0.408***	-0.141	-0.365	-0.377**	-0.093
	(0.089)	(0.106)	(1.259)	(0.148)	(0.133)	(0.257)	(0.190)	(0.210)
Observations	11001	6657	3868	2355	5335	2601	4045	2075
Mean retirement rate	0.207	0.064	0.441	0.153	0.194	0.284	0.206	0.163

Source: Author's estimations from the 1994-2010 Health and Retirement Study.

Notes: Standard errors in parentheses are clustered at the individual level. * Significant at 10%, ** significant at 5%, *** significant at 1%. All estimates are weighted with survey weights. Each regression contains individual fixed effects, year effects, and time-varying covariates (age, age squared, self-reported health, an indicator for having health insurance, lagged earnings, lagged financial and IRA wealth, lagged indicators for defined benefit plan, and defined contribution plan). Spousal retirement variables are instrumented with spousal age greater than or equal to 62. Wealth is defined as the sum of total financial and IRA wealth. The 25th percentile wealth was $10,000 and the 75th percentile wealth was $175,000 in 2006 dollars. The primary earner of the household indicates the spouse who contributes more than half of the total household earnings in the period preceding the job loss.

Chapter 3. Retention Heterogeneity in New York City Schools

(with Douglas Almond and Amy Ellen Schwartz)

3.1 Introduction

US school districts increasingly rely on standardized tests to evaluate teachers and students. Performance on "high stakes" tests can be a key determinant of whether students are retained or "held back" in their grade. Well-identified studies have found retention can be beneficial for short-term subsequent academic performance but possibly detrimental to longer-term outcomes that might be of greater importance (Jacob and Lefgren, 2004, 2009). Reliance on such tests is controversial in the US. For example, New York State is grappling with a sharply increased opt-out rate in spring 2015 by students who declined to sit for the statewide proficiency exam (*New York Times*, 2015).

We depart from previous literature by considering heterogeneity in how performance on standardized tests maps into consequences for students. Despite benchmarking from a common test and cutoff score, substantial scope for discretion exists in how exam results are utilized. Failing the exam can merely "start a conversation" about retention, where more often than not the student is promoted to the next grade. The lack of deterministic link between exam performance and retention opens the door to other factors shaping the retention decision. At present, we have little sense of how non-test factors shape retention among students who scored the same.

We focus on New York City public schools where roughly 5,500 students are retained each year. We analyze longitudinal data on 250,000 New York City public school students scoring near the failure threshold. Passing the annual proficiency exam essentially guarantees promotion to grades 4-9, while roughly 13% of those students failing the exam are retained. Compliers in our application are those who are retained because they failed the proficiency exam. Because there is a large population of never takers (promoted despite exam failure), the compliant sub-population may differ from not only the overall New York City student population (obvious), but also from the sub-population located near the threshold (less obvious). We analyze retention and average complier characteristics (Angrist and Pischke, 2009) using regression discontinuity methods.

We document pronounced heterogeneity in compliance along observable characteristics of the student. Moreover, this heterogeneity departs in important ways from what we had expected *a priori*.[43] In particular, we expected compliance to be highest among the youngest students, who were closest to the age-at-school entry cutoff. These students narrowly missed beginning kindergarten a year later and are on average less developed academically, socially, and physically than peers (particularly in early grades). Using administrative data on birth month, however, we do not find that retained students tend to be young for their grade. Nor do we find older students are more likely to be promoted after failure. Instead, we find race and gender to be important. Hispanic students are 60% more likely and Black students 120% more likely to be retained due to exam failure (relative to White students). Female students are 25% more likely to be retained in their grade due to exam failure than boys.[44] Poverty (free or reduced-price lunch eligibility[45]) and poor performance on previous exams also increase the likelihood of retention. Like age for grade, biometric measures of student height and weight do not seem to play a large role beyond the exam score. Again, we had expected smaller-stature students might face a higher retention risk when they fail because they might "fit in" physically in their repeated grade. We also show these biometric and demographic characteristics are smooth at the threshold. Thus it is not the case that, for example, Black students have discontinuously worse characteristics just below the threshold for passing. Nor do we find any evidence of heaping near the threshold.

We discuss two classes of "explanations" for the retention heterogeneity we uncover: student-level differences and school-level differences. Regarding the former, it is not the case that the predictive power of the baseline test score is different for girls or minorities than for the rest of

[43] See Tomchin and Impara (1992) for a description of factors affecting the probability of retention.

[44] Significant at the 1% level: see Section 3.5 and footnote 46.

[45] Students are eligible for free lunch if their parents or guardians make less than 130% of the poverty line and reduced lunch if their parents/guardians make less than 185% of the poverty line.

the student population (located near the failure threshold). Thus, we do not see evidence that, for example, girls are more likely to be retained when they fail because failure is a stronger predictor of future (poor) performance. On average, girls perform better in subsequent periods than boys with identical baseline scores. Other factors equal, this would suggest that the compliance rate among girls should be lower than for boys. Higher compliance of girls' retention with exam failure is puzzling. The unexplained gender gap is widest among Whites: failing increases a girl's retention rate to 5.9%, but when a White boy fails, only 0.9% are retained. Indeed, we cannot reject that exam failure has *zero* impact on retention for non-Hispanic White boys.

Turning to school-level characteristics, these are "balanced" by sex so disproportionate retention of girls who fail cannot be attributed to differential exposure to school characteristics. Race and ethnicity, in contrast, do vary with school-level characteristics. Among these school-level factors, "high retention" schools have more minority students on average. Furthermore, predominantly Black schools tend to be high compliance schools, i.e. schools where retention rates jump more below the failure threshold. While school-level factors thus appear important to racial heterogeneity in retention, so too do within-school factors. Blacks are substantially more likely to be retained than Whites (for identical baseline scores) at predominantly non-Black schools.

The existing literature has overlooked compliance heterogeneity: we know of no published work on the subject.[46] In addition to student composition of schools, we also consider faculty (Dee, 2005). The final retention decision is made by the school principal. We find a striking pattern whereby girls are substantially more likely to be retained due to exam failure at schools with a female principal. That said, because other (unobserved) characteristics of the school presumably vary by principal's characteristics (*cf.* student gender), we characterize this pattern as descriptive. Furthermore, because girls perform better on average than boys, the *unconditional* retention rates remain lower for girls than boys: girls score better on average and fewer girls fail (overall). This and the fact that relatively few students are retained in a given school each year may have obscured

[46] Two recent working papers using Florida records are discussed in Section 3.2.1.

higher retention rates among girls who just fail.

3.2 Background

3.2.1 Literature Review

Previous papers have used regression discontinuity approaches to consider impacts of retention on subsequent outcomes, beginning with Jacob and Lefgren (2004). Among third graders in Chicago public schools, Jacob and Lefgren (2004) found positive effects of retention and more mixed impacts among sixth graders. Jacob and Lefgren (2009) found that retention increased subsequent high school dropout rates. These findings are noteworthy as longer-term endpoints (like high school completion) might be more important endpoints for parents, students, and policy makers than shorter-term achievement. Because compliance rates are an order of magnitude higher in Chicago than in New York,[47] there is a different scope for heterogeneity in compliance in Chicago's context compared to New York.

Mariano and Martorell (2013) follow Jacob and Lefgren (2004, 2009) and exploit test score cutoffs used in assignment to summer school and retention in New York City. Specifically, they consider 2004-2008 data on fifth graders failing proficiency exam in 2004-2006. They find modest positive effects of summer school on English achievement. They estimate cohort-over-cohort test score differences ("external drift") and subtract it from the RD estimates of retention (see 3.5.5 section). They find large and positive effects of grade retention on both Math and English. As in Jacob and Lefgren (2004, 2009), heterogeneity in compliance is not considered.

Student characteristics, however, might conceivably play a role by shaping interactions between teachers and students. Dee (2005) uses National Education Longitudinal Study of 1988 (NELS:88) to examine the role of demographic similarity between teachers and students on teachers' perceptions of students. Dee (2005) makes within-student comparisons of teachers' perceptions, taking

[47] 41% of sixth-graders who failed to meet the promotion cutoff were retained in Chicago from 1993-1994 to 1998-1999 (Jacob and Lefgren, 2004).

advantage of the structure of NELS:88 data, which surveyed teachers in two different academic subjects, on their perceptions of individual students. Dee (2005) finds that teachers are more likely to have negative perceptions towards students who do not share the same race/ethnicity and gender. His findings suggest that demographic characteristics of students such as gender and race/ethnicity may potentially matter for retention decisions as well, as they are partly based on teachers' evaluations of students.

Labelle and Figlio (2013) and Schwerdt et al. (2015) consider Florida's test-based promotion policy and evaluate various future outcomes. Labelle and Figlio (2013) stands out as most similar to our approach (we discovered their conference draft after conducting our analysis). Labelle and Figlio (2013) examine whether Florida's grade retention policy that mandated promotion to the fourth grade conditional on meeting a minimum standard in third grade reading was being implemented differently depending on maternal education (using matched educational data and birth records). They employ a regression discontinuity design, taking advantage of the score cutoff for determining retention, finding that students whose mothers have less than a high school degree are 20 percent more likely to be retained than students whose mothers have a bachelor's degree or more. Factors besides parental education, including eligibility for free school lunch and other dimensions of student performance, shape heterogeneity in compliance as well. They also estimate the effect of retention on future test scores instrumenting for grade retention with scoring below the promotion cutoff. They find that retention leads to short-term gains in test scores but that the gains fade out over time, consistent with Jacob and Lefgren (2009). They find no evidence, however, that differential retention by maternal education has differential impacts on students' future test scores.

Labelle and Figlio (2013) additionally show that students are more likely to be retained if they are Black (9 percent increase), male (13 percent increase), have a foreign born mother (13 percent increase), and qualify for free or reduced-price lunch (9 percent increase). Within subgroups categorized by student race, free or reduced-price lunch eligibility, and school characteristics, they still find a similar (but imprecise) pattern in retention probabilities by maternal education. Heterogeneity by student gender is not discussed. They attribute differential retention by maternal education to systematic differences in parental behavior in response to retention risk, although they cannot directly test this hypothesis.

Schwerdt et al. (2015) emphasize the impacts of test failure on retention and future outcomes. They attempt to address the endogeneity of the subsequent exam to retention and consider subsequent reading, Math test scores, and high school graduation. Short-term gains in both Math and reading fade over time. They also find no clear impact on graduation and little evidence of systematic heterogeneity by student and school characteristics. They also look at complier characteristics (Angrist and Pischke, 2009) and find that a complier is more likely to score level 1 in Math. Their conclusion is that early grade retention might be favorable (e.g. short-term gains and no detrimental effects), although long-term benefits are uncertain.

At present, there is no published work using regression discontinuity methods to consider heterogeneity in compliance with exam failure/passing. The magnitude of heterogeneity we find in New York City is substantially larger than that found in recent analyses of Florida students and manifests along additional dimensions, e.g. gender of student and principal.

3.2.2 Promotion Policy

In New York, students in grades 3-8 take the State Math and English Language Arts (ELA) tests each spring. The “scale score” is the number of correct answers converted into a vertically comparable score (comparable across grades). Scale scores are categorized into four performance levels separately for Math and ELA: level 1 - not meeting State learning standards, level 2 - partially meeting State learning standards, level 3 - meeting State learning standards, and level 4 - exceeding State learning standards.

Scoring level 2 (“partially meeting” standards) in both tests essentially guarantees promotion, whereas students who score level 1 (“not meeting” standards) in either subject are at risk of being retained. The failure threshold for each subject varies by year and grade. Retention procedures are less formalized in New York than Florida, with New York having few explicit exemptions. That said, English Language Learners and students with disabilities who receive special education services are exempt from New York's stated promotion criteria.[48] In our sample, 13% of students who failed to meet the promotion cutoff were retained. Thus, there is substantial scope for heterogeneity in compliance, driven predominantly by the "never takers".

3.3 Data

We analyze administrative data from the New York City public school system for the 2007-2008 to 2011-2012 academic years. Student-level panel data on New York State English Language Arts (ELA) and Mathematics scale scores are merged to demographic characteristics, including race, gender, free or reduced-price lunch eligibility, and age in months. Additionally, we observe students' weight, height, and BMI, measures further described in Almond et al. (2016). Unique student identifiers allow us to track students over time as long as they stay in the New York City public school system. When the student's grade level in year $t + 1$ is the same as that in year t, we code the student as retained. 1,507,700 student records for grades 3-8 are available 2007-2012, and approximately 2% are retained. The retention rates in grades 3-8 have increased over time from 1% in 2007-2008 to 3% in 2010-2011. Over our analysis period, roughly 4% of students are ever retained.

Table 3.1 reports mean student characteristics for the whole sample (column 1), those who passed both tests but scored within 10 units of the cutoff (column 2), and those who failed to meet the promotion cutoff in either test and within 10 scale score units (column 3). Relative to the overall sample, students in this "retention window" are more likely to be Black or Hispanic, and less likely to be Asian or White. The proportion of female students is *lower*, and the proportion of students who are eligible for free or reduced-price lunch higher near the cutoff. 13% of students below the

[48] Empirically, however, we find that these groups of students were also affected by the policy and thus do not exclude them in our analysis. That said, our results are not sensitive to excluding them.

failure threshold were retained while 0.7% of those “just above” the threshold were retained.

3.4 Estimation

To assess heterogeneity in how standardized test scores are utilized, we exploit the jump in retention rates at the failure threshold in a regression discontinuity framework. We estimate the following equation both “pooled” and separately by student characteristics:

$$Y_{igs,t+1} = \alpha_0 + \alpha_1 \cdot 1[X_{igst} < 0] + \alpha_2 \cdot X_{igst} + \alpha_3 \cdot 1[X_{igst} < 0] \cdot X_{igst} + \eta_{gst} + \epsilon_{igst} \tag{11}$$

where i is individual, g is grade, s is subject, and t is year. Y is an indicator for whether the student is retained or not. X_{igst} is minimum of the Math and English test scores, re-centered to zero at their respective failure thresholds. We use this measure as the main running variable, since students are at risk of grade retention when they score level 1 in *either* Math or English test.

We fit a linear relationship between the scale score and the probability of retention, allowing for different slopes above and below the cutoff (consistent with our figures). We include year×grade×subject fixed effects, η_{gst}, to control for year-, grade-, and subject-specific cutoffs. We estimate equation (1) by OLS and report robust standard errors.[49] We focus on the roughly 250,000 student observations within 10 scale score (approximately one third of a standard deviation for both Math and English) of the cutoff. In the tables, we report the RD estimate α_1, which measures the size of the discontinuity at the failure threshold.

3.4.1 Discontinuities in Baseline Covariates?

Figure 3.1 shows histograms of the running variable both in the full sample (panel (a)) and within 10 scale score from the failure cutoff (panel (b)). We do not observe any heaping around the

[49] We do not cluster our standard errors at the running variable level since we found out that clustered standard errors from separate regressions are inconsistent with clustered standard errors from pooled regressions. In addition, Kolesár and Rothe (2016) argue that the convention of clustering standard errors on the running variable performs poorly in a regression discontinuity framework with a discrete running variable.

failure cutoff (normalized to 0).[50] As there is no evidence of manipulation around the cutoff, we expect students to have similar characteristics above and below the cutoff. We summarize covariates by predicting the probability of retention using student gender, race/ethnicity, age in months, BMI, height, weight, free or reduced-price lunch eligibility, special education participation, and previous Math and English scale scores. Figure 3.2 compares this predicted probability of retention around the cutoff. There is no evidence of a discontinuity at the cutoff in the full sample (panel (a)) nor separately for females (panel (b)) or for Black students (panel (c)). The corresponding regression estimates of the discontinuities are precisely estimated zeros.

3.5 Results

Figure 3.3 summarizes the mean probability of retention for students near the cutoff. Consistent with stated school policy, the probability of retention drops discontinuously at the cutoff. Moreover, the linear specification seems to fit the data well (Gelman and Imbens, 2014). Table 3.2 reports the RD estimates from estimating equation (1) "pooled" and separately by subgroup. Overall, failing to meet the promotion cutoff increases the probability of retention by 5 percentage points (column 1).

We are particularly interested in documenting whether exam failure has different retention consequences by baseline characteristics. Panel A of Table 3.2 shows that Black students are 3.4 percentage points more likely to be retained than White students, more than double the White retention probability as induced by failure (2.9%). Hispanic students are around 2 percentage points more likely to be retained than non-Hispanic Whites, a 60% increase. Asians are, if anything, are less likely to be retained than non-Hispanic White students when they fail to meet the cutoff.

Girls are 1.2 percentage points (or 27%) more likely to be retained than boys failing the exam

[50] Dee et al. (2011) document evidence of manipulation of Regent's exam scores among New York City *high school* students, finding "roughly 3 to 5 percent of the exam scores that qualified for a high school diploma actually had performance below the state requirement". Key for us, they do not find any evidence of manipulation on the statewide Math and English exams given to students in grades 3-8. Likewise, we detect no evidence of manipulation among the proficiency exams taken prior to high school (i.e. grades 3-8).

(panel B of Table 3.2). The gender difference is statistically significant at the 0.01 level (p-value = 0.005). This is intriguing since the overall retention rate in grades 3-8 is higher for boys (2.1%) than for girls (1.7%). But when we examine the retention rates in a narrow window near the failure threshold, we find the opposite: girls are more likely to be retained than boys. Additionally, we find that low performance on previous year's Math test increases the probability of retention (panel C in Table 3.2).[51] Finally, those who are eligible for subsidized lunch are 1.3 percentage points more likely to be retained than those ineligible (panel D in Table 3.2).

Figure 3.4 presents these findings graphically. Panel (a) shows a large disparity in retention between Black and White students below the cutoff. Likewise, panel (b) shows the mean probability of retention is higher for Hispanics than for Whites, although the gap is smaller. Panel (c) shows that mean retention probabilities are similar between Asians and Whites. Panel (d) shows Blacks and Hispanics are roughly twice as likely to be retained than Asians and Whites when they fail. Panel (e) shows that girls are more likely to be retained than boys conditional on scoring identically below the cutoff. Below the threshold, the girl mean is above the boy mean at each scale score, but means are indistinguishable above the threshold. Additionally, we examine whether the probability of retention differs by age for grade, height for grade, and weight status category. We might expect students who are younger or smaller than their peers in the same grade are more likely to be retained, since they would potentially fit in better in their repeated grade (socially, physically, and academically). Parents might also be less likely to object to the retention decision if their child was a "close call" with respect to age at school entry cutoff. We also test whether students who are "too big to fail" are in fact less likely to be retained. However, we find surprisingly little heterogeneity along these dimensions (Table 3.3).

Given the stark heterogeneity by ethnicity and gender, we consider interactions between these

[51] Additionally, we examine whether the demographic heterogeneity we find disappears once we condition on previous test scores. We estimate separate regressions by student characteristics for each decile of previous test scores. We find that ethnicity and gender heterogeneity in retention probability generally persist across the distributions. This suggests that there are other factors driving the differential retention probabilities that are independent of previous academic performance.

dimensions. The gender gap is especially large among Whites (5.9 percentage points versus 0.9 percentage points). For all ethnicity groups, we find that girls are more likely to be retained than boys.

3.5.1 Racial Composition of Schools

Here we explore the role of school-level differences in explaining heterogeneity. Retention policies and practices are shaped by principals and teachers, and thus may differ by school. Given pronounced residential sorting within New York City, Black students might disproportionately attend schools that more strictly adhere to a test-based promotion policy than schools White students attend. We examine whether the probability of retention due to exam failure differs between schools with different Black shares, dividing schools into three equal-sized groups by their proportion of Black students: low (mean 5%), middle (mean 24%), and high (mean 60%). The overall mean retention rate in grades 3-8 is higher in high share schools (2.8%) than low share schools (0.8%). Furthermore, predominantly Black schools tend to be high compliance schools, i.e. where discontinuity in retention rates is larger at the failure threshold.[52] Thus school-level differences can "explain" (in a statistical sense) some of the individual differences in retention by race.

However, panel A of Table 3.4 shows that the Black-White gap in the probability of retention is much larger in schools with *low* share of Black students. At predominantly Black schools, we do not see a racial disparity in the effect of failing the exam. The difference in retention probability between Black and White students is only 0.7 percentage points, and it is not statistically significant. Panel B shows that these findings are not sensitive to including school fixed effects (nor would we expect them to be, as retention's predictors are and should be continuous at the cutoff). Thus, Black students are more likely to be retained *within* predominantly non-Black schools. To summarize, higher black retention rates are attributable to both school-level differences and dif-

[52] Exam failure increases retention probability by 3.6 percentage points in low share schools, by 4.4 percentage points in middle share schools, and by 6.5 percentage points in high share schools.

ferential responses to failure within predominantly non-Black schools. More generally, including school fixed effects indeed leaves our impact estimates essentially unchanged, including impact estimates by demographic subgroup.

3.5.2 Student-level Differences

Because students are not segregated by gender in New York City schools, gender heterogeneity in compliance cannot be driven by differences in school-level characteristics. We examine other observable student-level differences which may explain the gender gap. For instance, students who are more likely to be retained conditional on identical test scores might perform worse in other performance measures. This exercise is necessarily imperfect because we do not observe everything observed by teachers, principals, and parents. On the other hand, as researchers we *do* observe some key information unobserved by schools and parents: information on the *future* academic performance of students.

We compare average performance of girls and boys in baseline test scores, baseline attendance rate, and future test scores. (We depart from usual regression discontinuity analyses by *not* interpreting the jump scores at the failure threshold.[53]) As in previous studies, girls perform better than or as well as boys on average along these dimensions. Conditional on scoring identically on the baseline Math test in our retention window, girls also score better than boys on baseline English test, future Math test, and future English test. Moreover, they have similar slopes in the relationship between other test scores and baseline Math score as boys, implying that the predictive power of baseline test score is not different by gender (panel (a) of Figure 3.5).[54] Panel (b) shows that the slope of Black students above the cutoff is also similar to that of White students. It remains a

[53] The particular exam taken is determined by a student's year in school, so the exam taken changes discontinuously at the threshold due to retention. If one is willing to ignore that potential compositional effect, there is an apparent increase in short-run academic performance due to retention, as has been found in previous literature. See Section 3.5.5.

[54] The relationship above the failure cutoff is easier to interpret because it is not affected by endogenous retention.

puzzle that girls are about 25% more likely to be retained when they fail compared to boys, and that this gender gap is especially large for Whites.

Additionally, we examine whether heterogeneity in short-run benefits of retention can explain higher retention of girls and minorities conditional on test score. As retained and promoted students take different tests in subsequent years, it is fundamentally difficult to compare future test scores below and above the threshold. We attempt to address this issue by comparing same-grade test scores both in the baseline grade (i.e. test scores in the baseline year for the promoted versus test scores in the following year for the retained) and one grade above (i.e. test scores in the following year for the promoted versus test scores two years later for the retained). We find no obvious and robust heterogeneity in these future test scores, suggesting that it is unlikely that larger potential benefits on future performance for girls and minorities drive the differential retention decisions in the baseline.

3.5.3 Who Done It?

In this section, we focus on the role of principals. Teachers' perceptions of students can be based on their racial/ethnic and gender similarities to students (Dee, 2005). Unfortunately, we do not observe the classroom to which students are assigned within grade and school (or the demographics of teachers). But according to the New York City Department of Education website:

> *Principals will review these portfolios in August and make a holistic promotion decision for each student. Superintendents will continue to review promotion appeals for cases in which a parent disagrees with the principal's decision.*

As the final retention decision is made by principals and superintendents, we utilize data on school principal demographics, which come from a single 2008 cross-section of roughly 1,400 schools. This yields a subsample of 19,421 student records within 10 scale score of the cutoff. We consider whether the gender gap in retention varies by principal's gender.

Table 3.5 shows that the female-male difference in retention probability is pronounced in schools with female principals, while it essentially disappears in male principal schools. This is consistent with the findings from Hanna and Linden (2012) (admittedly in a radically different context): *"In fact, we observe the opposite, with discrimination against the low-caste children being driven by low-caste graders, and graders from the high-caste groups appearing not to discriminate at all even when controlling for the education and age of grader".* On gender, we do not know of an economics of education paper with a similar finding to ours. Bagues et al. (2015) argue that having women on faculty review committees in Italy and Spain, if anything, leads to fewer female faculty being promoted.

Additionally, we find that the Black-White gap in retention probability is large (11.3 percentage points versus 6.2 percentage points) in schools with White principals. The ethnicity gap disappears and is imprecisely estimated in Black principal schools, although this is partly due to the small number of White students in these schools. Because other (unobserved) characteristics of the school presumably vary by principal's observed characteristics, however, we characterize this pattern as descriptive.

3.5.4 Statistical Discrimination

Can the canonical theory of statistical discrimination (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977) explain the heterogeneity we find? Through this lens, principals make the retention decision based on the current test score, x, which is a noisy signal for the true level of academic success in the next grade, q. That is, $x = q + u$ where $u \sim N(0, \sigma^2_u)$. In addition, principals have formed expectations of academic success for different demographic groups from experience: $q_s \sim N(\bar{q_s}, \sigma^2_{q,s})$. Let $s = \{f, m\}$ denote the gender group. Then, the expected academic success of a student with test score x and gender s can be written as $\alpha_s x + (1 - \alpha_s)\bar{q_s}$, where $\alpha_s = \frac{\sigma^2_{q,s}}{\sigma^2_{q,s} + \sigma^2_{u,s}}$. (Since the signal s may be more informative for one group than another, we let σ^2_u to vary across groups and denote it as $\sigma^2_{u,s}$.) Intuitively, if the observed signal is noisy, α_s goes down and thus principals would put more weight on group mean and less weight weight on the observed signal. The female-male difference in the expected academic success conditional on scoring identically $x = k$ on the current test is:

$$E(q|x = k, s = f) - E(q|x = k, s = m) = (\alpha_f k + (1 - \alpha_f)\bar{q_f}) - (\alpha_m k + (1 - \alpha_m)\bar{q_m})$$

If we assume that there is no difference in group mean, $\bar{q_f} = \bar{q_m} = \bar{q}$, the above equation reduces to $(\alpha_f - \alpha_m)(k - \bar{q})$. Therefore, if the current test score is a noisier measure for boys than for girls (i.e. $\alpha_f > \alpha_m$), the female-male difference in the expected academic success is negative for below average students ($k < \bar{q}$).

Turning to our data, we assume that principals have formed expectations of group performance based on previous year's Math test score and observe current year's Math test score. In our full sample, previous year's Math test score is slightly higher on average ($\bar{q_f} = 681.7$ and $\bar{q_m} = 680.7$) and more precise ($\sigma^2_{q,f} = 33.5^2$ and $\sigma^2_{q,m} = 34.6^2$) for girls. In addition, the current Math test score is noisier for boys ($\sigma^2_{u,f} = 31.9^2$ and $\sigma^2_{u,m} = 32.8^2$). Using these estimates, $E(q|x = k, s = f) - E(q|x = k, s = m) = -0.002k + 1.998$. Evaluating this at the mean current Math test score below the cutoff $x = 637.4$, we find that the female-male difference is small and rather positive ($-0.002(637.4) + 1.998 = 0.7$). In this simple framework, the signal for boys is *not* noisy enough for our findings to be consistent with statistical discrimination.

3.5.5 BMI Impacts?

As in previous econometric studies of retention, considering the causal effects on subsequent academic performance is not straight-forward even with a valid instrument for retention. This is because the grade level of the exam students take in subsequent year is endogenous to retention decision. Therefore, it is difficult to distinguish the endogenous "exam taken" effect from the effect of retention on academic performance. We do not have a "silver bullet" solution to this problem.[55]

However, BMI testing does not vary by grade, and thus its evaluation is not compromised by endogenous retention. Moreover, as BMI percentiles vary by age in months and age itself is unaffected by retention, BMI percentiles are comparable for retained versus non-retained students. Furthermore, we observe BMI for all students, and have sufficient power to consider biometric impacts. Following a health economics literature on peer effects in BMI (Halliday and Kwak, 2009), timing of puberty and its responsiveness to social/environmental factors (Bharadwaj and Cullen, 2013), we test whether the higher probability of retention due to exam failure affects BMI in the following year. We instrument for retention with scoring below the failure threshold and estimate the effect of retention on next year BMI using 2SLS. Table 3.6 shows that retention due to exam failure does not have a statistically significant impact on next year BMI, although point estimates indicate that grade retention might lower BMI relative to promoted peers. We conclude the peer effect on BMI is not large in our compliant sub-population, although our 2SLS estimates are somewhat imprecise.

3.5.6 Complier Characteristics

In this section, we take a more systematic approach to describing heterogeneity in compliance to the retention policy. The LATE theorem states that if treatment effects are heterogeneous, an instrument captures the causal effect for the sub-population of compliers (in our application, those who are retained as a result of exam failure). While it is not possible to identify individual compliers, it is possible to describe the distribution of complier characteristics. We estimate compliers' mean observable characteristics following Angrist and Pischke (2009); Almond and Doyle (2011).[56]

[55] Mariano and Martorell (2013) address the endogeneity by estimating "external drift", which we do not pursue here.

[56] Curiously, seven years after Angrist and Pischke (2009) recounted a straight-forward approach to describe compliers, empirical economists seldom do. Recent methodological contributions in Angrist and Fernández-Val (2013),

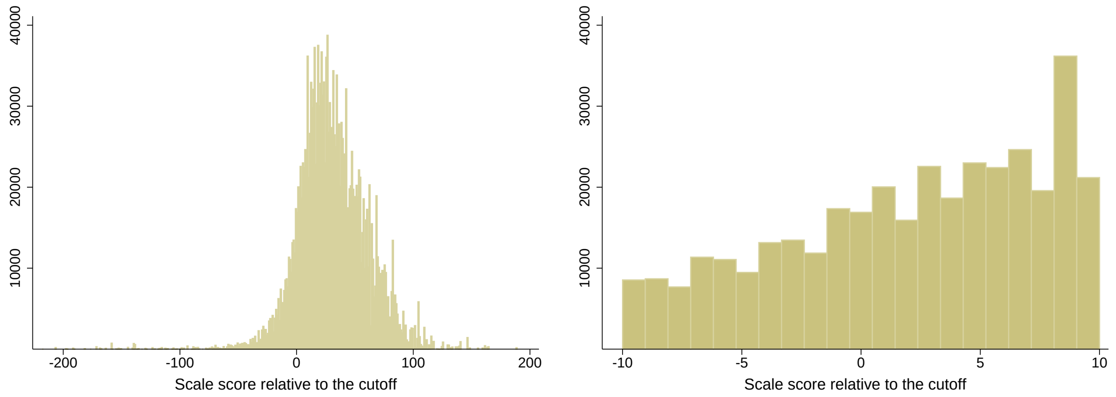
Table 3.7 shows that mean characteristics in fact vary substantially across different samples. Compliers are less likely to be Asian or White, while they are much more likely to be Black (49%) compared to those both in our retention window (38%) and in the full sample (31%). Insofar as race is concerned, compliance appears more selective than does scoring near the threshold. Turning to income, scoring near the threshold increases the share receiving a reduced-price lunch from 86 to 93%, while compliers are "only" 95% poor. Thus, performance on the test is more strongly related to income rather than how the test is used. Turning to gender, compliers are on average 48% female, versus 46% in our retention window (and 50% overall). The fraction obese is remarkably similar across these subgroups.

3.6 Discussion

The process by which retention decisions are made is often opaque despite utilization of standardized test scores and common thresholds. There is little systematic evidence on this "black box". We find both the magnitude and nature of this heterogeneity surprising. Why are younger students not more likely to be retained conditional on their exam score? In contrast, both race and gender help predict retention *conditional* on the baseline test score. Compliance with proficiency exams in New York City is thus selective. We find these descriptive patterns interesting *per se* and invite additional research on whether retention decisions are "fair". Are girls and minorities over-retained? The need for such work is underscored by previous research (from other contexts where students can be tracked for longer time periods) that there may be long-term impacts on marginally-retained students (Jacob and Lefgren, 2009). Such outcomes may be more important than the shorter-term benefits students show somewhat mechanically from repeating material they have seen in the previous year. Thus, it is not merely the case that the retention decision is perceived at the time as momentous by parents and students.

Dehejia et al. (2015), and Kowalski (2016) are notable exceptions.

3.7 Figures

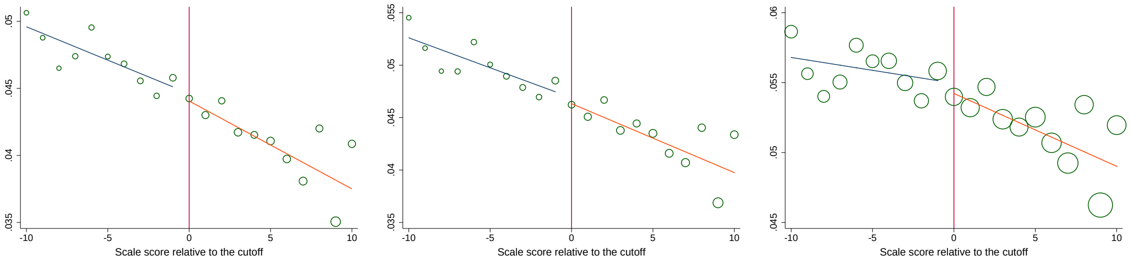


(a) All

(b) Retention Window

Figure 3.1: Distribution of the running variable

Notes: The running variable is minimum of the Math and English test scores re-centered to zero at their own failure thresholds.



(a) All

(b) Female only

(c) Black only

Figure 3.2: Predicted probability of retention

Notes: We estimate the predicted probability of retention using gender, race/ethnicity, age in months, BMI, height, weight, free or reduced-price lunch eligibility, special education participation, and previous Math and English scale scores. Each circle plots mean predicted probability of retention within each one scale score bin. The size of the circle depends on the number of observations in each bin. Lines are the fitted values from a regressions of the predicted probability on the exam failure dummy, allowing for different slopes above and below the cutoff.

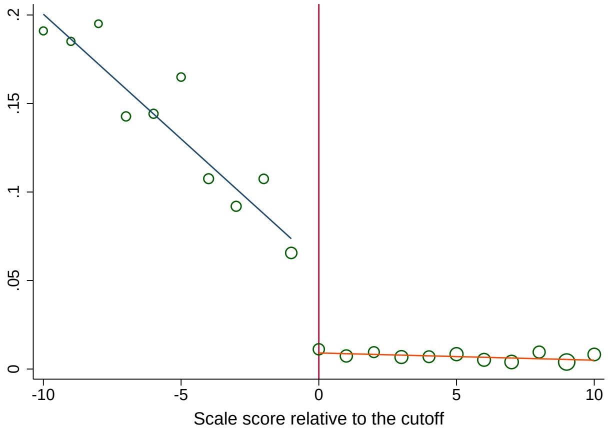
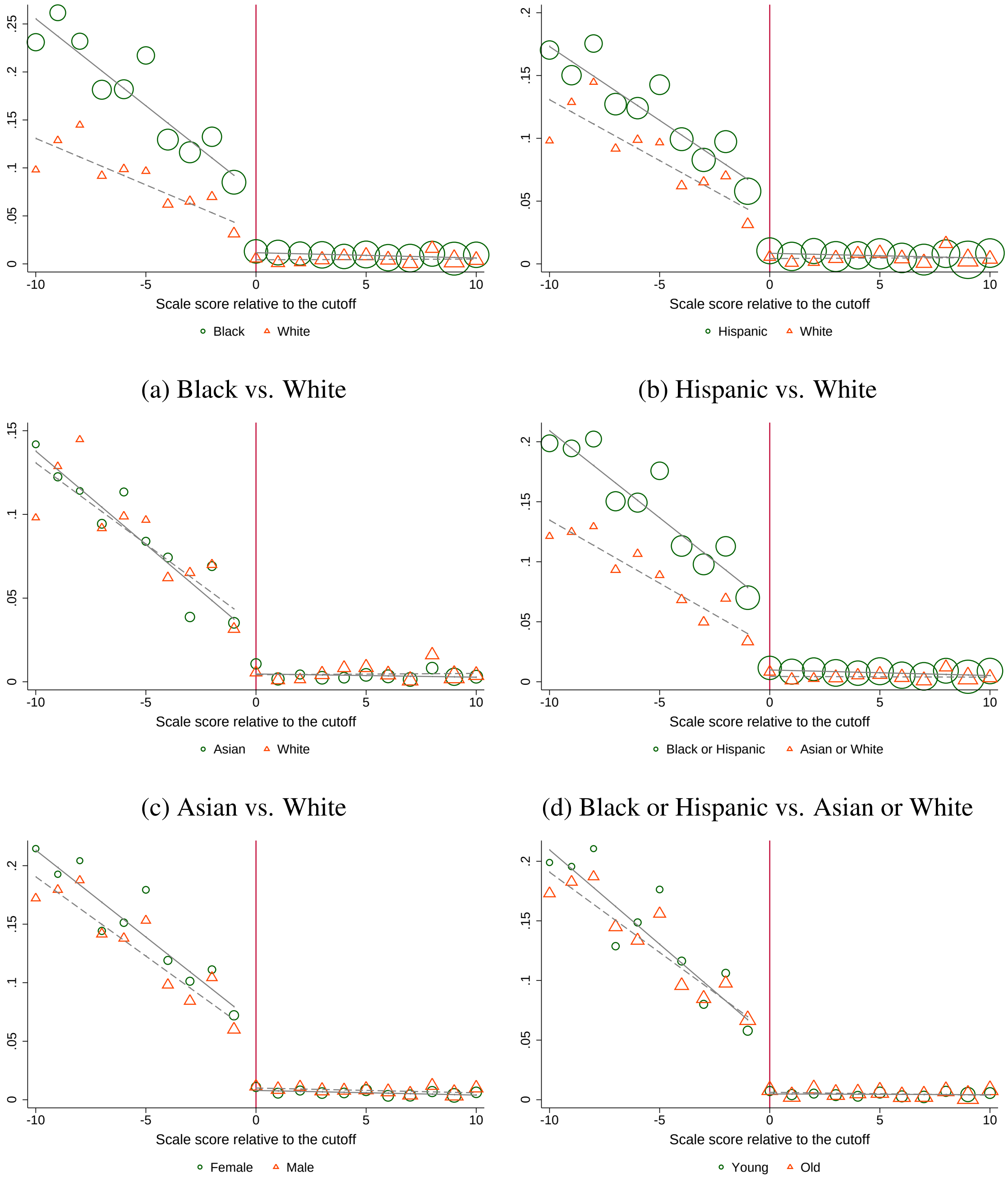


Figure 3.3: Probability of retention

Notes: Each circle plots mean probability of retention within each one scale score bin. The size of the circle depends on the number of observations in each bin. Lines are the fitted values from a regressions of a retention dummy on the exam failure dummy, allowing for different slopes above and below the cutoff.



(a) Black vs. White

(b) Hispanic vs. White

(c) Asian vs. White

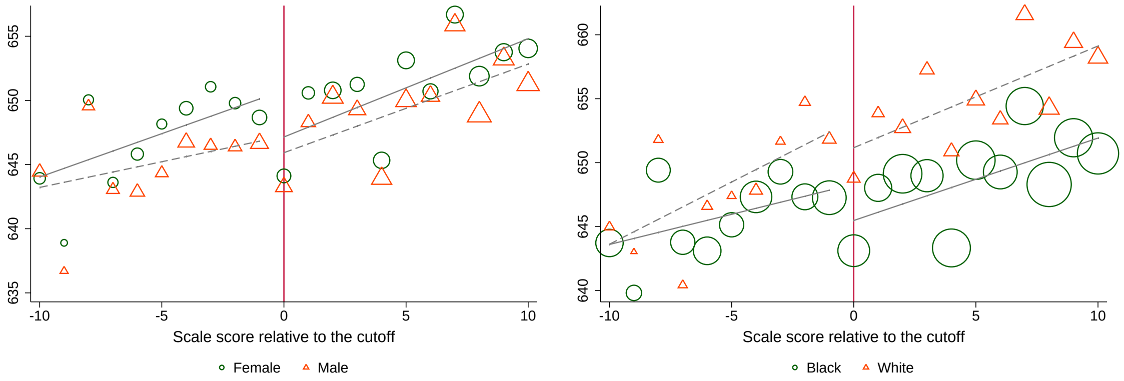
(d) Black or Hispanic vs. Asian or White

(e) Female vs. male

(f) Young vs. old

Figure 3.4: Heterogeneity in compliance

Notes: Each circle (or triangle) plots mean probability of retention within each one scale score bin. The size of the circle (or triangle) depends on the number of observations in each bin. Lines are the fitted values from a regression of a retention dummy on the exam failure dummy, allowing for different slopes above and below the cutoff. We divide each grade into three equal-sized groups based on age in months for grade. Panel (f) compares the youngest group with the oldest group.



(a) Female vs. male

(b) Black vs. White

Figure 3.5: Next year Math scale score

Notes: We use the re-centered baseline Math test score conditional on passing English as the running variable. Each circle (or triangle) plots mean Math test scores in the subsequent year within each one scale score bin. The size of the circle (or triangle) depends on the number of observations in each bin. Lines are the fitted values from a regressions of next year Math test score on the exam failure dummy, allowing for different slopes above and below the cutoff.

3.8 Tables

Table 3.1: Summary statistics

		Retention window	
	All	Above	Below
Asian	0.154	0.079	0.069
Black	0.307	0.376	0.388
Hispanic	0.394	0.469	0.485
White	0.142	0.073	0.054
Female	0.500	0.467	0.448
Free or reduced-price lunch	0.860	0.930	0.941
Age in months	133.1	134.5	135.7
Weight (lbs)	101.2	104.7	105.4
Height (inches)	58.2	58.4	58.5
Math level 1	0.049	0.000	0.376
Math level 2	0.245	0.712	0.482
Math level 3	0.461	0.261	0.132
Math level 4	0.245	0.027	0.010
English level 1	0.074	0.000	0.723
English level 2	0.370	0.957	0.256
English level 3	0.504	0.042	0.021
English level 4	0.053	0.001	0.000
Retention	0.018	0.007	0.128
N	1,507,700	77,543	168,047

Notes: Retention window indicates 10 scale score above and below the failure threshold.

Table 3.2: Effect of exam failure on the probability of retention

		A. Ethnicity				B. Gender	
	All	Asian	Black	Hispanic	White	Female	Male
Below cutoff	0.050	0.021	0.063	0.047	0.029	0.057	0.045
	(0.002)	(0.006)	(0.004)	(0.003)	(0.007)	(0.003)	(0.003)
Observations	245,590	18,636	93,331	116,477	16,383	113,207	132,383
Mean below cutoff	0.128	0.080	0.162	0.113	0.080	0.137	0.122
Mean above cutoff	0.007	0.004	0.009	0.006	0.005	0.006	0.008
		C. Previous Math test score				D. Subsidized lunch	
		Low	Middle	High		Eligible	Not eligible
Below cutoff		0.092	0.049	0.021		0.051	0.038
		(0.004)	(0.005)	(0.004)		(0.002)	(0.008)
Observations		55,800	53,868	55,312		223,661	15,956
Mean below cutoff		0.157	0.133	0.086		0.129	0.112
Mean above cutoff		0.010	0.007	0.005		0.007	0.009

Notes: Each column reports the estimated discontinuity in the probability of retention for different subsamples. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses. In panel C, we divide the sample into three equal-sized groups based on last year's Math scale score. Subsidized lunch in panel D indicates free or reduced-price lunch eligibility.

Table 3.3: Effect of exam failure on the probability of retention

	A. Age for grade			B. Height for grade			C. Weight status			
	Young	Middle	Old	Short	Middle	Tall	Underweight	Healthy	Overweight	Obese
Below cutoff	0.045	0.050	0.050	0.049	0.045	0.052	0.063	0.049	0.042	0.056
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.013)	(0.003)	(0.005)	(0.004)
Observations	60,054	61,948	105,338	54,018	73,424	89,368	7,512	112,641	40,757	84,680
Mean below cutoff	0.128	0.131	0.123	0.123	0.130	0.124	0.116	0.129	0.126	0.130
Mean above cutoff	0.005	0.004	0.005	0.005	0.004	0.004	0.004	0.005	0.004	0.012

Notes: Each column reports the estimated discontinuity in the probability of retention for different subsamples. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. We divide each grade into three equal-sized groups based on age in months for grade (panel A) and height for grade (panel B). Each student's body mass index (BMI) is classified to be underweight, healthy, overweight, and obese based on age- and sex-specific BMI cutoffs from Centers for Disease Control.

Table 3.4: Heterogeneity by school's proportion of Black students

	Low (mean=5%)				High (mean=60%)			
	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White
A. Without school fixed effects								
Below cutoff	0.027	0.055	0.039	0.017	0.023	0.066	0.064	0.059
	(0.008)	(0.017)	(0.005)	(0.008)	(0.017)	(0.005)	(0.007)	(0.024)
B. With school fixed effects								
Below cutoff	0.026	0.058	0.040	0.017	0.022	0.067	0.063	0.065
	(0.008)	(0.017)	(0.005)	(0.008)	(0.019)	(0.004)	(0.007)	(0.028)
Observations	11,475	4,133	40,165	10,612	2,988	68,828	26,948	1,797
Mean below cutoff	0.069	0.123	0.094	0.070	0.102	0.167	0.132	0.105
Mean above cutoff	0.003	0.008	0.004	0.004	0.005	0.010	0.010	0.009

Notes: We divide schools into three equal-sized groups by schools' proportion of Black students. The mean proportion of Black students is 5% in low share schools. It is 60% in high share schools. Each column reports the estimated discontinuity in the probability of retention for different race/ethnicity groups. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses.

Table 3.5: Gender heterogeneity in retention probability by principal's gender

	Female principal		Male principal	
Student:	Female	Male	Female	Male
Below cutoff	0.135	0.081	0.070	0.083
	(0.017)	(0.015)	(0.024)	(0.022)
Observations	6,481	7,188	2,689	3,063
Mean below cutoff	0.133	0.123	0.101	0.120
Mean above cutoff	0.005	0.006	0.005	0.011

Notes: We utilize data on principal gender from 2007-2008. First two columns compare the estimated discontinuity in the probability of retention by student gender in schools with a female principal. Last two columns examine heterogeneity by student gender in schools with a male principal. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses.

Table 3.6: Effect of retention on next year BMI

		By ethnicity				By gender	
	All	Asian	Black	Hispanic	White	Female	Male
Retention	-0.413	4.773	-0.143	-0.981	3.275	-1.019	0.321
	(0.958)	(7.809)	(1.283)	(1.435)	(6.366)	(1.273)	(1.450)
Observations	208,916	17,171	76,138	100,075	14,841	96,119	112,797
Mean below cutoff	21.9	20.1	21.9	22.2	21.3	22.0	21.7
Mean above cutoff	21.8	20.0	21.9	22.1	21.1	21.9	21.6

Notes: Each column reports the estimated effect of retention on next year BMI for different subsamples. We instrument for retention with scoring below the failure threshold and estimate the effect of retention on next year BMI using 2SLS. We assume linear relationship between next year BMI and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses.

Table 3.7: Mean characteristics

Characteristic	Complier $E(X\|D_1 = 1, D_0 = 0)$	Retention window (N=245,590)	All (N=1,507,700)
Asian	0.043	0.076	0.154
Black	0.492	0.380	0.307
Hispanic	0.428	0.474	0.394
White	0.033	0.067	0.142
Female	0.482	0.461	0.501
Free or reduced-price lunch	0.950	0.933	0.860
Age in months	135.2	134.9	133.1
Weight (lbs)	105.1	104.9	101.2
Height (inches)	58.5	58.5	58.2
Obese	0.338	0.345	0.332
N		245,590	1,507,700

Notes: First column summarizes mean characteristics of compliers following Angrist and Pischke (2009); Almond and Doyle (2011).

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Acemoglu, D. and Finkelstein, A. (2008). Input and technology choices in regulated industries: Evidence from the health care sector. *Journal of Political Economy*, 116(5):837–880.

Aigner, D. J. and Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, pages 175–187.

Aizer, A., Currie, J., and Moretti, E. (2007). Does managed care hurt health? Evidence from Medicaid mothers. *The Review of Economics and Statistics*, 89(3):385–399.

Almond, D. and Doyle, J. J. (2011). After midnight: A regression discontinuity design in length of postpartum hospital stays. *American Economic Journal: Economic Policy*, 3(3):1–34.

Almond, D., Doyle, J. J., Kowalski, A. E., and Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, 125(2):591–634.

Almond, D., Lee, A., and Schwartz, A. E. (2016). Impacts of classifying new york city students as overweight. *Proceedings of the National Academy of Sciences*, 113(13).

Anderson, P. M., Gustman, A. L., and Steinmeier, T. L. (1999). Trends in male labor force participation and retirement: Some evidence on the role of pensions and social security in the 1970s and 1980s. *Journal of Labor Economics*, 17:757–783.

Angrist, J. D. and Fernández-Val, I. (2013). Extrapolate-ing: External validity and overidentification in the late framework. In Acemoglu, D., Arellano, M., and Dekel, E., editors, *Advances in Economics and Econometrics, Tenth World Congress*, volume 3, chapter 11. Cambridge University Press.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arad, I., Gofin, R., Baras, M., Bar-Oz, B., Peleg, O., and Epstein, L. (1999). Neonatal outcome of inborn and transported very-low-birth-weight infants: Relevance of perinatal factors. *European Journal of Obstetrics, Gynecology, and Reproductive Biology*, 83(2):151–157.

Arora, P., Bajaj, M., Natarajan, G., Arora, N. P., Kalra, V. K., Zidan, M., and Shankaran, S. (2014). Impact of interhospital transport on the physiologic status of very low-birth-weight infants. *American journal of perinatology*, 31(03):237–244.

Arrow, K. (1973). The theory of discrimination. In Ashenfelter, O. and Rees, A., editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press.

Arrow, K., Auerbach, A., Bertko, J., Brownlee, S., Casalino, L. P., Cooper, J., Crosson, F. J., Enthoven, A., Falcone, E., Feldman, R. C., et al. (2009). Toward a 21st-century health care system: Recommendations for health care reform. *Annals of Internal Medicine*, 150(7):493–495.

Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2015). Does the gender composition of scientific committees matter? *IZA Discussion Paper Series*, (9199).

Banks, J., Blundell, R., and Rivas, M. C. (2010). The dynamics of retirement behavior in couples: Reduced-form evidence from england and the us. *Unpublished manuscript*.

Barreca, A. I., Guldi, M., Lindo, J. M., and Waddell, G. R. (2011). Saving babies? Revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, 126:2117–2123.

Bharadwaj, P. and Cullen, J. B. (2013). Coming of age: Timing of adolescence and gender identity formation. (Preliminary and incomplete. Do not cite without permission).

Bharadwaj, P., Løken, K. V., and Neilson, C. (2013). Early life health interventions and academic achievement. *American Economic Review*, 103(5):1862–1891.

Blau, D. M. (1998). Labor force dynamics of older married couples. *Journal of Labor Economics*, 16(3):595–629.

Blau, D. M. and Gilleskie, D. B. (2006). Health insurance and retirement of married couples. *Journal of Applied Econometrics*, 21(7):935–953.

Bound, J. and Waidmann, T. (2007). Estimating the health effects of retirement. *Michigan Retirement Research Center working paper 2007-168*.

Brown, J., Duggan, M., Kuziemko, I., and Woolston, W. (2014). How does risk-selection respond to risk-adjustment? New evidence from the Medicare Advantage Program. *American Economic Review*, 104(10):3335–3364.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Card, D. and Lee, D. S. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674.

Casanova, M. (2010). Happy together: A structural model of couples? joint retirement choices. *Unpublished Manuscript, Department of Economics, University of California, Los Angeles*.

Centers for Medicare & Medicaid Services (CMS) (2015a). Medicaid managed care enrollment and program characteristics, 2013.

Centers for Medicare & Medicaid Services (CMS) (2015b). Medicaid managed care trends and snapshots, 2000-2013.

Chan, S. and Stevens, A. H. (1999). Employment and retirement following a late-career job loss. *American Economic Review*, pages 211–216.

Chan, S. and Stevens, A. H. (2001). Job loss and employment patterns of older workers. *Journal of Labor Economics*, 19(2):484–521.

Chan, S. and Stevens, A. H. (2004). How does job loss affect the timing of retirement? *Contributions in Economic Analysis & Policy*, 3(1).

Clemens, J. and Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104(4):1320–1349.

Coile, C. (2004). Retirement incentives and couples' retirement decisions. *Topics in Economic Analysis & Policy*, 4(1).

Coile, C. C. and Levine, P. B. (2007). Labor market shocks and retirement: Do government programs matter? *Journal of Public Economics*, 91(10):1902–1919.

Coile, C. C. and Levine, P. B. (2011). Recessions, retirement, and social security. *American Economic Review*, 101(3):23–28.

Conover, C. J., Rankin, P. J., and Sloan, F. A. (2001). Effects of tennessee Medicaid managed care on obstetrical care and birth outcomes. *Journal of health politics, policy and law*, 26(6):1291–1324.

Currie, J. and Fahr, J. (2005). Medicaid managed care: Effects on children's Medicaid coverage and utilization. *Journal of Public Economics*, 89(1):85–108.

Cutler, D. M., McClellan, M., and Newhouse, J. P. (2000). How does managed care do it? *The RAND Journal of Economics*, 31(3):526–548.

Dafny, L. (2005). How do hospitals respond to price changes? *American Economic Review*, 95(5):1525–1547.

Dafny, L. and Gruber, J. (2005). Public insurance and child hospitalizations: Access and efficiency effects. *Journal of public Economics*, 89(1):109–129.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, pages 158–165.

Dee, T. S., Jacob, B. A., Rockoff, J. E., and McCrary, J. (2011). Rules and discretion in the evaluation of students and schools: The case of the new york regents examinations. manuscript, Columbia Business School.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2015). From local to global: External validity in a fertility natural experiment. NBER Working Paper 21459.

Duggan, M. (2004). Does contracting out increase the efficiency of government programs? Evidence from Medicaid HMOs. *Journal of Public Economics*, 88(12):2549–2572.

Duggan, M. and Hayford, T. (2013). Has the shift to managed care reduced Medicaid expenditures? Evidence from state and local-level mandates. *Journal of Policy Analysis and Management*, 32(3):505–535.

Duggan, M., Kearney, M. S., and Rennane, S. (2015). The Supplemental Security Income (SSI) Program. NBER Working Paper 21209.

Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of health economics*, 5(2):129–151.

Freedman, S. (2016). Capacity and utilization in health care: The effect of empty beds on neonatal intensive care admission. *American Economic Journal : Economic Policy*, 8(2):154–185.

Gaynor, M., Town, R. J., and Ho, K. (2015). The industrial organization of health care markets. *Journal of Economic Literature*, 53(2):235–284.

Gelman, A. and Imbens, G. (2014). Why high-order polynomials should not be used in regression discontinuity designs. Working Paper 20405, National Bureau of Economic Research.

Geruso, M. and Layton, T. (2015). Upcoding: Evidence from Medicare on squishy risk adjustment. NBER Working Paper 21222.

Gustman, A. L. and Steinmeier, T. L. (2000). Retirement in dual-career families: a structural model. *Journal of Labor Economics*, 18(3):503–545.

Gustman, A. L. and Steinmeier, T. L. (2004). Social security, pensions and retirement behavior within the family. *Journal of Applied Econometrics*, 19(6):723–737.

Hackbarth, G., Reischauer, R., and Mutti, A. (2008). Collective accountability for medical care—toward bundled Medicare payments. *New England Journal of Medicine*, 359(1):3–5.

Halliday, T. J. and Kwak, S. (2009). Weight gain in adolescents and their peers. *Economics & Human Biology*, 7(2):181 – 190.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68.

Harman, J. S., Hall, A. G., Lemak, C. H., and Duncan, R. P. (2014). Do provider service networks result in lower expenditures compared with HMOs or primary care case management in Florida's Medicaid program? *Health Services Research*, 49(3):858–877.

Harris, E. A. and Fessenden, F. (2015). 'opt out' becomes anti-test rallying cry in new york state. *The New York Times*.

Hawthorne, J. and Killen, M. (2006). Transferring babies between units: Issues for parents. *Infant*, 2(2):44–46.

Healthcare Cost and Utilization Project (HCUP) (2013). Hospital stays for newborns, 2011.

Herring, B. and Adams, E. K. (2011). Using HMOs to serve the Medicaid population: What are the effects on utilization and does the type of HMO matter? *Health Economics*, 20(4):446–460.

Ho, K. and Pakes, A. (2014). Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review*, 104(12):3841–3884.

Holahan, J. and Schirmer, M. (1999). Medicaid managed care payment methods and capitation rates: results of a national survey. Technical report, Urban Institute.

Howell, E. M., Dubay, L., Kenney, G., and Sommers, A. S. (2004). The impact of Medicaid managed care on pregnant women in Ohio: a cohort analysis. *Health services research*, 39(4p1):825–846.

Hurd, M. D. (1990a). The joint retirement decision of husbands and wives. In *Issues in the Economics of Aging*, pages 231–258. University of Chicago Press.

Hurd, M. D. (1990b). Research on the elderly: Economic status, retirement, and consumption and saving. *Journal of Economic Literature*, 28(2):565–637.

Iglehart, J. K. (2011). Desperately seeking savings: States shift more Medicaid enrollees to managed care. *Health Affairs*, 30(9):1627–1629.

Jacob, B. A. and Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics*, 86(1):226–244.

Jacob, B. A. and Lefgren, L. (2009). The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics*, 1(3):33–58.

Jacobson, L. S., LaLonde, R. J., and Sullivan, D. G. (1993). Earnings losses of displaced workers. *American Economic Review*, pages 685–709.

Kaestner, R., Dubay, L., and Kenney, G. (2005). Managed care and infant health: An evaluation of Medicaid in the US. *Social science & medicine*, 60(8):1815–1833.

Kaiser Family Foundation (KFF) (2015). Medicaid reforms to expand coverage, control costs and improve care: Results from a 50-state Medicaid budget survey for state fiscal years 2015 and 2016.

Kim, H. B. and Lee, S.-m. (2016). When public health intervention fails: Cost sharing, crowd-out, and selection in Korea's national cancer screening program.

Kolesár, M. and Rothe, C. (2016). Inference in regression discontinuity designs with a discrete running variable. manuscript, arXiv:1606.04086 [stat.AP].

Kowalski, A. E. (2016). Doing more when you're running late: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments. NBER Working Paper 22363.

Krieger, J. W., Connell, F. A., and LoGerfo, J. P. (1992). Medicaid prenatal care: A comparison of use and outcomes in fee-for-service and managed care. *American Journal of Public Health*, 82(2):185–190.

Kuziemko, I., Meckel, K., and Rossin-Slater, M. (2013). Do insurers risk-select against each other? Evidence from Medicaid and implications for health reform. NBER Working Paper 19198.

Labelle, C. L. and Figlio, D. N. (2013). The uneven implementation of universal school policies: Maternal education and florida's mandatory grade retention policy. Conference draft, Association for Education Finance and Policy (accessed 9/2015).

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Levinson, A. and Ullman, F. (1998). Medicaid managed care and infant health. *Journal of Health Economics*, 17(3):351–368.

Libersky, J., Dodd, A. H., and Verghese, S. (2013). National and state trends in enrollment and spending for dual eligibles under age 65 in Medicaid managed care. *Disability and Health Journal*, 6:87–94.

Maestas, N. (2001). Labor, love and leisure: complementarity and the timing of retirement by working couples. *Unpublished manuscript, University of California, Berkeley*.

Mariano, L. T. and Martorell, P. (2013). The academic effects of summer instruction and retention in new york city. *Educational Evaluation and Policy Analysis*, 35(1):96–117.

Marton, J., Yelowitz, A., and Talbert, J. C. (2014). A tale of two cities? The heterogeneous impact of Medicaid managed care. *Journal of health economics*, 36:47–68.

Mazzonna, F. and Peracchi, F. (2012). Ageing, cognitive abilities and retirement. *European Economic Review*, 56(4):691–710.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Michaud, P.-C. (2003). Joint labour supply dynamics of older couples. *IZA Discussion Paper no. 832*.

Michaud, P.-C. and Vermeulen, F. (2004). A collective retirement model: identification and estimation in the presence of externalities. *IZA Discussion Paper no. 1294*.

Mohamed, M. A. and Aly, H. (2010). Transport of premature infants is associated with increased risk for intraventricular haemorrhage. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 95(6):F403–F407.

Mori, R., Fujimura, M., Shiraishi, J., Evans, B., Corkett, M., Negishi, H., and Doyle, P. (2007). Duration of inter-facility neonatal transport and neonatal mortality: Systematic review and cohort study. *Pediatrics International*, 49(4):452–458.

Munnell, A. H., Sass, S., Soto, M., and Zhivan, N. (2006). Has the displacement of older workers increased? *Chestnut Hill, MA: Center for Retirement Research at Boston College*.

Nasr, A. and Langer, J. C. (2011). Influence of location of delivery on outcome in neonates with congenital diaphragmatic hernia. *Journal of pediatric surgery*, 46(5):814–816.

Nasr, A. and Langer, J. C. (2012). Influence of location of delivery on outcome in neonates with gastroschisis. *Journal of pediatric surgery*, 47(11):2022–2025.

New York State Department of Health (NYSDOH) (2000). Medicaid coverage for newborns.

New York State Department of Health (NYSDOH) (2001). Automatic Medicaid enrollment for newborns (chapter 412 of the laws of 1999).

New York State Department of Health (NYSDOH) (2016). New York State Medicaid program transportation manual policy guidelines.

New York State Office of the State Comptroller (NYS Comptroller) (2014). Overpayments to managed care organizations and hospitals for low birth weight newborns.

Ohning, B. L. (2015). Transport of the critically ill newborn. *Medscape*.

Olson, C. (1992). The impact of permanent job loss on health insurance benefits. Working paper, Princeton University, Department of Economics, Industrial Relations Section.

Parker, J. D. and Schoendorf, K. C. (2000). Variation in hospital discharges for ambulatory care-sensitive conditions among children. *Pediatrics*, 106(Supplement 3):942–948.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*, 62(4):659–661.

Quinn, K. (2008). New directions in Medicaid payment for hospital care. *Health Affairs*, 27(1):269–280.

Rohwedder, S. and Willis, R. J. (2010). Mental retirement. *The journal of economic perspectives*, 24(1):119.

Ruhm, C. J. (1991). Are workers permanently scarred by job displacements? *American Economic Review*, 81(1):319–324.

Sacarny, A. (2014). Technological diffusion across hospitals: The case of a revenue-generating practice. Job market paper, Massachusetts Institute of Technology.

Schwerdt, G., West, M. R., and Winters, M. A. (2015). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. Working Paper 21509, National Bureau of Economic Research.

Shigeoka, H. and Fushimi, K. (2014). Supplier-induced demand for newborn treatment: Evidence from Japan. *Journal of health economics*, 35:162–178.

Sparer, M. (2008). *Medicaid Managed Care Reexamined*. United Hospital Fund.

Stevens, A. H. and Moulton, J. G. (2013). Effects of late-life job loss on wealth and labor supply. In Couch, K. A., Daly, M. C., and Zissimopoulos, J. M., editors, *Lifecycle Events and Their Consequences: Job Loss, Family Change, and Declines in Health*. Stanford University Press.

Sullivan, D. and Von Wachter, T. (2009). Job displacement and mortality: An analysis using administrative data. *The Quarterly Journal of Economics*, 124(3):1265–1306.

Tomchin, E. M. and Impara, J. C. (1992). Unraveling teachers' beliefs about grade retention. *American Educational Research Journal*, 29(1):199–223.

United Hospital Funds (UHF) (2000). Provider networks in Medicaid managed care plans. *Currents*, 5(4).

Van der Klaauw, W. and Wolpin, K. I. (2008). Social security and the retirement and savings behavior of low-income households. *Journal of Econometrics*, 145(1):21–42.

Van Parys, J. (2015). How do managed care plans reduce healthcare costs? Job market paper, Columbia University.

Appendix A. Hospital Payments Under MMC

A.1 State Payments to Health Plans

The state negotiates with each health plan to determine monthly capitation payments in New York State. Health plans submit data on enrollees and previous expenditures and propose new rates based on expected costs for each region they participate. The state reviews the data and offers a new set of rates that vary by age, sex, and region. These rates are applicable for a one-year period. The plans can receive a bonus up to 3 percent of the rate based on their performance on quality measures. In 2008, the state introduced a new payment system that accounts for health conditions of the enrollees by adjusting the capitation rates by Clinical Risk Groups. This new payment system was fully implemented in 2011 (Sparer, 2008).[57]

The New York State Medicaid program paid a monthly capitation rate of $138 on average for newborns younger than six months old in 1998 (Holahan and Schirmer, 1999), which is roughly $190 in 2011 values. For newborn services, however, plans receive lump-sum payments for costs related to newborn medical care in addition to monthly capitation payments. These lump-sum payments range from $2,277 to $6,651 per newborn weighing 1,200 grams or more (NYS Comptroller, 2014). Effective April 2012 following the expansion of the MMC mandate to infants with birth weight below 1,200 grams, plans receive lump-sum payments ranging from $68,355 to $105,108 per newborn for these low birth weight enrollees.

In return, health plans are responsible for providing health care services to their enrollees. Health plans offer a network of health care providers to their enrollees and reimburse the providers for their services. Health plans employ a number of payment methods to reimburse providers. I focus on reimbursement for inpatient services in this paper.

[57] It is unclear whether risk-adjusted payments can in fact reduce adverse selection and thus reduce government spending (Brown et al., 2014).

A.2 Plan Payments to Hospitals

For patients enrolled in MMC, hospitals are paid in several ways depending on contractual details between health plans and hospitals. However, plan-to-provider payment rates for MMC in New York State are classified as confidential and proprietary and thus not available. Although the exact payment methods and rates are unknown, most health plans in New York State reimburse providers through primary care capitation models (UHF, 2000). Inpatient payments associated with newborn medical care are often excluded in monthly capitation payments for primary care capitation models and are reimbursed on a fee-for-service basis using a Diagnosis-Related Group (DRG) method.[58] That is, each inpatient stay is classified into a DRG, and Medicaid pays a fixed rate to hospitals based on the DRG assigned to the patient (Quinn, 2008).

The New York State Department of Health (NYSDOH) provides inpatient payments base rates for enrollees in both the FFS system and the MMC system along with weights for each DRG.[59] The state Medicaid program uses the FFS rates for inpatient payments for patients enrolled in FFS. The MMC rates are intended to be used by health plans as base rates in negotiation with hospitals. As expected, these MMC rates are generally lower than the FFS rates that the state uses to pay hospitals directly. In 2009, for instance, the base discharge rate for FFS was $6,471.31 on average, while the base contract discharge rate for MMC was $5,284 on average.

[58] New York State implemented a severity-based methodology, All Patient Refined Diagnosis Related Groups (APR-DRGs) effective December 1, 2009. Prior to that, New York State utilized All Patient Diagnosis Related Groups (AP-DRG) for hospital payments.

[59] http://www.health.ny.gov/facilities/hospital/reimbursement/apr-drg/rates/ffs/index.htm

Appendix B. Computing Complier Characteristics

I follow the estimation proposed by Almond and Doyle (2011) to compute complier characteristics:

$$E(X|compliers) = \frac{p_C + p_A}{p_C}\left[E(X|D = 1, Z = 1) - \frac{p_A}{p_C + p_A}E(X|D = 1, Z = 0)\right]$$

where X indicates hospital/patient characteristics, D denotes the treatment, which is MMC participation in my context. Z denotes the instrument, which is exceeding the 1,200-gram threshold under the RD framework and the county-specific MMC mandate under the DD framework. p_A is the proportion of always takers, and p_N is the proportion of never takers. Assuming monotonicity (i.e., no defiers), I compute the proportion of compliers using the estimates, $p_C = 1 - p_A - p_N$.[60]

Given the independence of Z, I use the sample proportion of those enrolled in MMC even though their birth weight is below the threshold to estimate p_A in the RD framework. Similarly, for the DD framework, I use the sample proportion of those enrolled in MMC even though the MMC mandate is not implemented in their county. To estimate p_N for the RD framework, I use the sample proportion of those who are not enrolled in MMC even though their birth weight is above the threshold. For the DD framework, I use the sample proportion of those who are not enrolled in MMC even though the MMC mandate is implemented in their county.

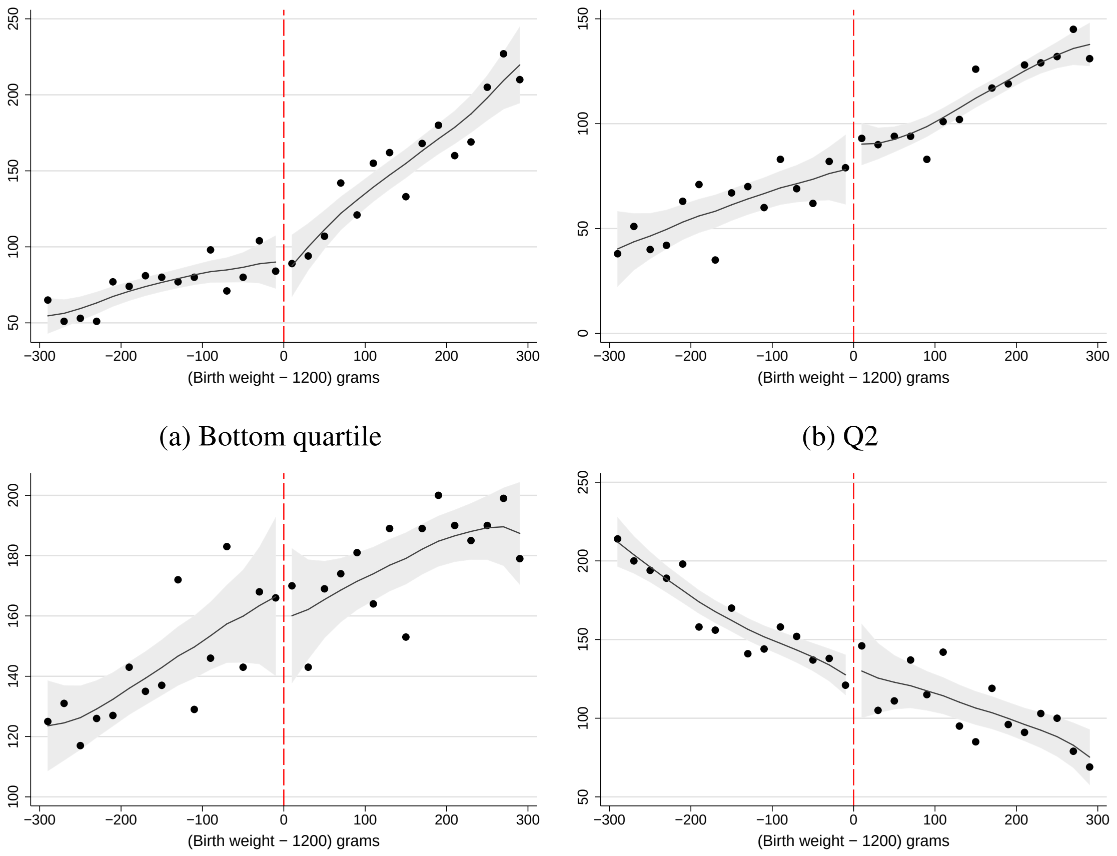
I use sample means for those who are affected by the instrument and participate in Medicaid HMO to estimate $E(X|D = 1, Z = 1)$ and sample means for those who are not affected by the instrument but participate in Medicaid HMO to estimate $E(X|D = 1, Z = 0)$. Tables below present each parameter for two instruments and show the estimates of $E(X|D = 1, Z = 1)$ and $E(X|D = 1, Z = 0)$ used in computing complier means in Table 1.11.

[60] The size of compliers can also be estimated from a simple regression of D on a binary Z.

	RD	DD
Z	Birth weight≥1,200 g	Years following the MMC mandate
p_A	0.04	0.05
p_N	0.74	0.73
$p_C = 1 - p_A - p_N$	0.22	0.22

	RD		DD	
	$E(X\|D = 1, Z = 1)$	$E(X\|D = 1, Z = 0)$	$E(X\|D = 1, Z = 1)$	$E(X\|D = 1, Z = 0)$
Panel A. Hospital characteristics				
Total beds	745.0	675.7	641.1	670.5
NICU beds	20.2	18.2	14.1	12.8
Number of physicians	177.1	108.7	148.8	150.2
Number of nurses	1256.5	1078.0	1039.1	790.5
Total admissions	34684.6	31752.9	30693.1	25785.4
Total births	3819.4	3505.5	3626.9	3213.5
NICU	0.92	0.88	0.80	0.76
Teaching hospital	0.68	0.59	0.56	0.57
Indigent care	0.71	0.65	0.69	0.31
Lives covered, capitated (*1995 values*)	6488.8	3423.4	5613.2	4831.9
Share covered by Medicaid, infants	0.57	0.54	0.59	0.57
Share covered by Medicaid, all patients	0.36	0.34	0.36	0.39
Share covered by HMO, infants	0.18	0.19	0.17	0.20
Share covered by HMO, all patients	0.21	0.21	0.20	0.17
Panel B. Patient characteristics				
Birth weight (grams)	1278.7	1120.6	3265.8	3277.2
Fraction low birth weight (<2,500 grams)	1.00	1.00	0.07	0.08
Female	0.49	0.43	0.48	0.49
White	0.21	0.22	0.27	0.28
Black	0.37	0.33	0.22	0.31
Hispanic	0.22	0.22	0.27	0.22
Asian	0.06	0.04	0.10	0.04
Median income, quartile 1	0.53	0.38	0.47	0.17
Median income, quartile 2	0.20	0.32	0.22	0.33
Median income, quartile 3	0.16	0.13	0.18	0.47
Median income, quartile 4	0.11	0.18	0.13	0.04
Admission scheduled	0.61	0.59	0.83	0.69
Admission on the weekend	0.25	0.33	0.23	0.25
Observations	8848	8848	4173544	4173544

Appendix C. Appendix Figures



(a) Bottom quartile

(b) Q2

(c) Q3

(d) Top quartile

Figure C.1: Mean frequency of the running variable by each 20-gram bin, by predicted list prices

Notes: Predicted list prices are computed from regressions of total charges on principal diagnosis and principal procedure fixed effects. I divide the sample by quartiles using the predicted list prices. Each panel plots mean frequency for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold for each quartile of predicted list prices. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.

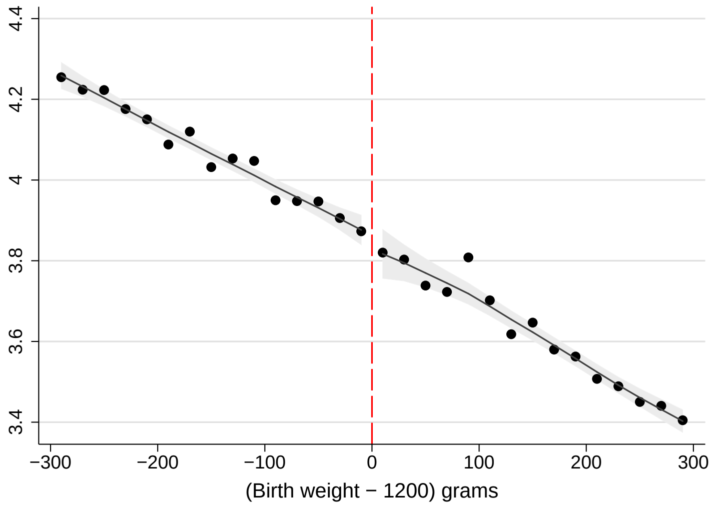
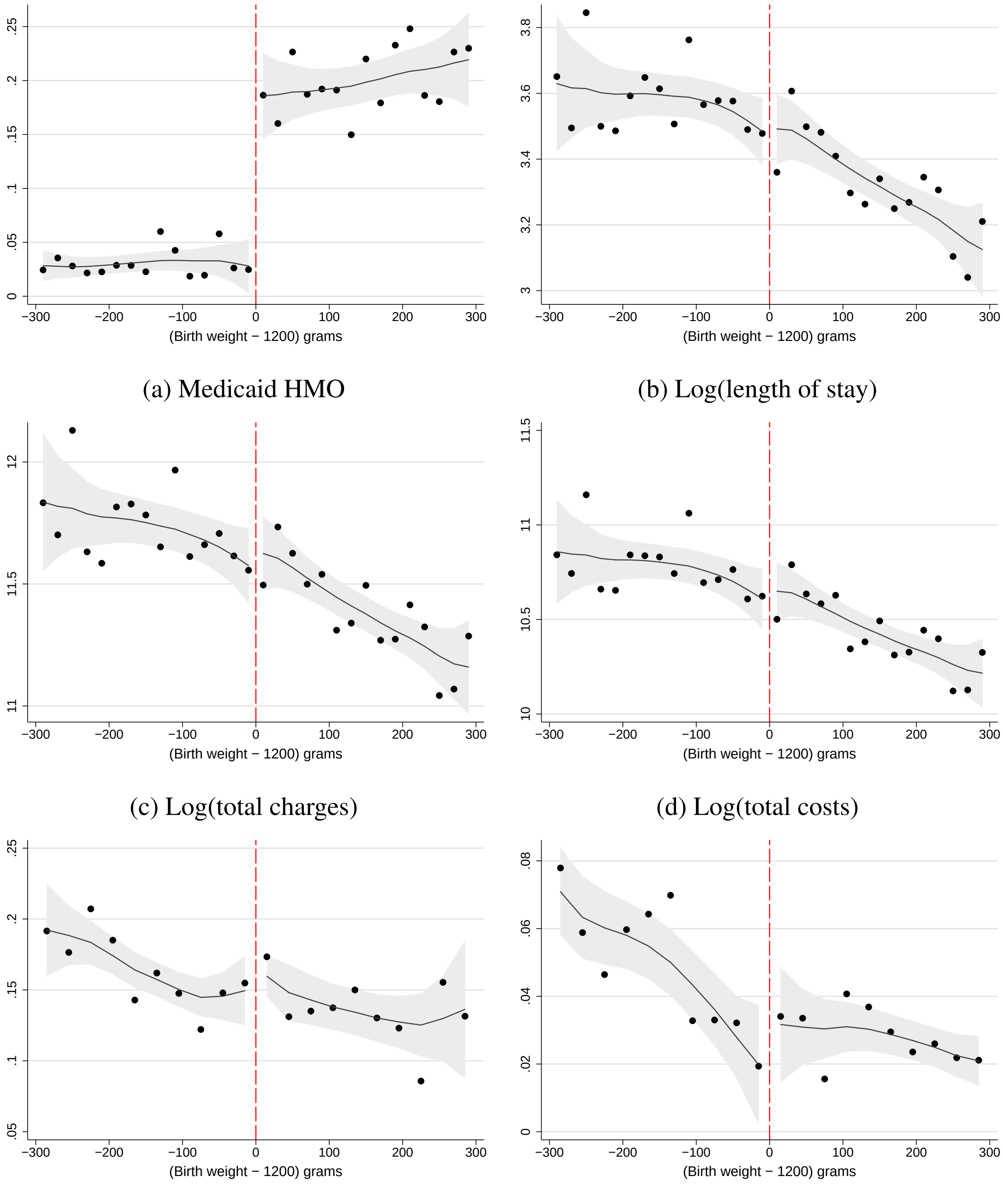


Figure C.2: Log(length of stay) for infants routinely discharged

Notes: The figure plots mean values of log(length of stay) for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold. The sample is restricted to those who are routinely discharged from birth hospitals. Each 20-gram bin contains roughly 250 discharge records. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.



(a) Medicaid HMO

(b) Log(length of stay)

(c) Log(total charges)

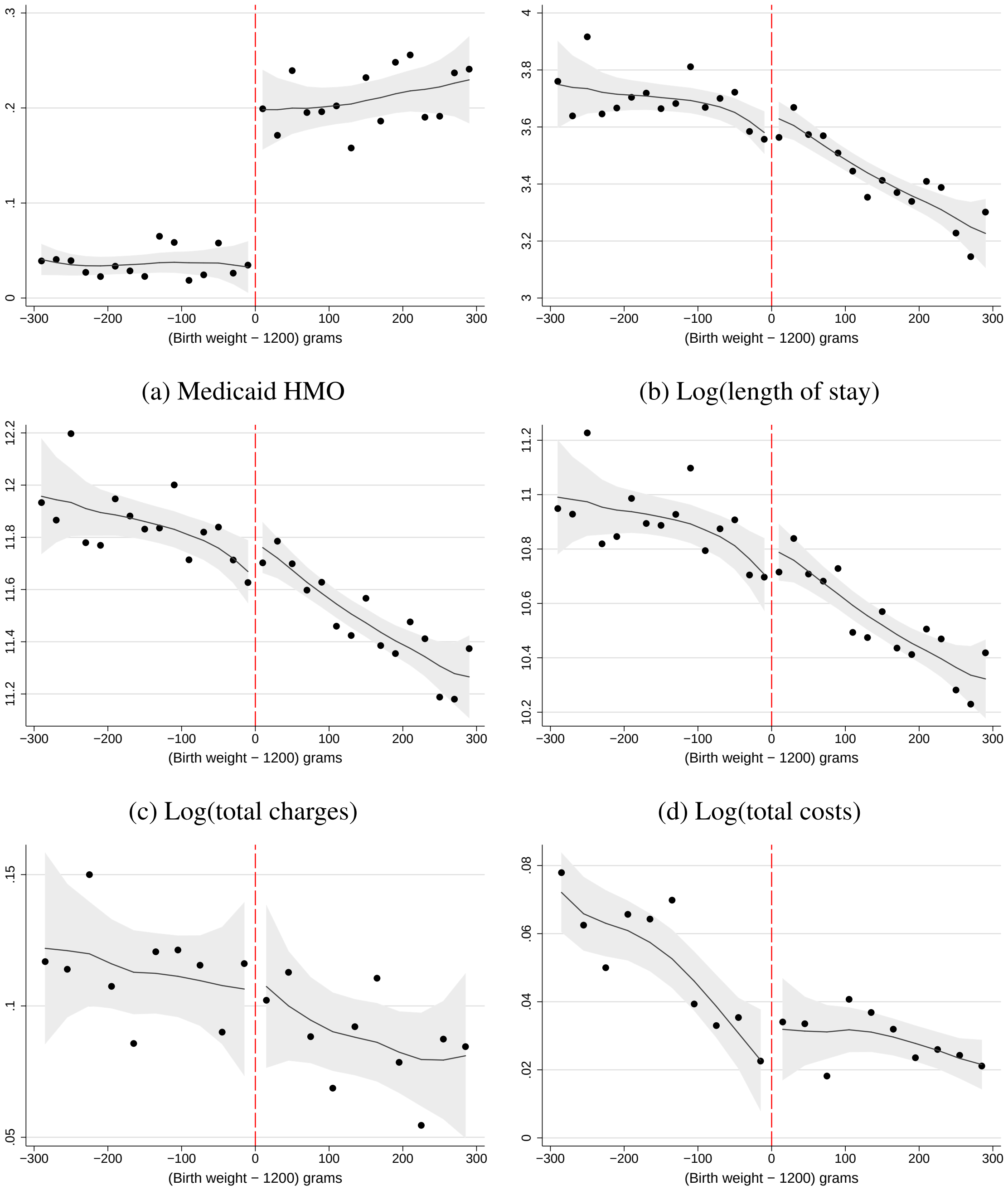
(d) Log(total costs)

(e) Transfer

(f) Mortality during hospitalization

Figure C.3: Effects of birth weight≥1,200 grams on discharge outcomes at birth, rest of the state

Notes: Panels (a)-(d) plot mean values of each outcome variable for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold. For panels (e) and (f) I use a bigger 30-gram bin for better visibility since transfer and death are both rare events and thus noisy. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.



(a) Medicaid HMO

(b) Log(length of stay)

(c) Log(total charges)

(d) Log(total costs)

(e) Readmission

(f) Mortality during hospitalization

Figure C.4: Effects of birth weight≥1,200 grams on cumulative outcomes, rest of the state

Notes: Each outcome aggregates the value at the individual level including the value at transferred hospitals (if transferred). Panels (a)-(d) plot mean values of each outcome variable for each 20-gram bin (dots) along with a local linear regression fitted lines (solid lines) and 95% confidence intervals below and above the threshold. For panels (e) and (f) I use a bigger 30-gram bin for better visibility since transfer and death are both rare events and thus noisy. I use the triangular kernel and a bandwidth of 150 grams for local linear regressions.

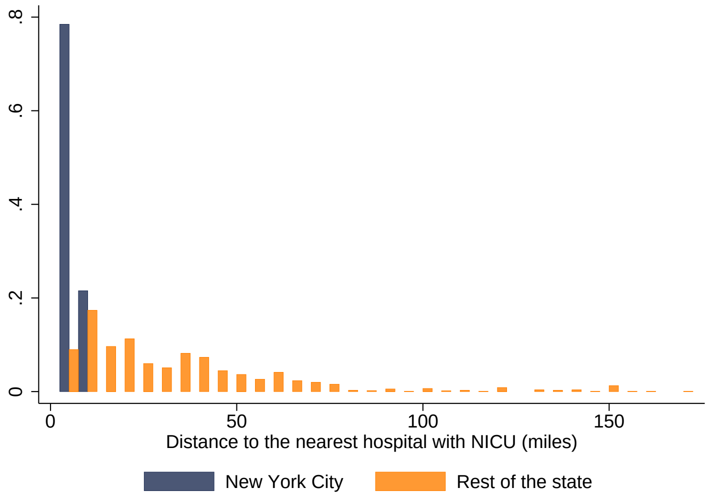


Figure C.5: Proximity to the nearest hospital with a NICU facility

Notes: Navy bars show the density of New York City hospitals by the distance to the nearest hospital with a NICU. Orange bars show the density of hospitals outside of New York City.

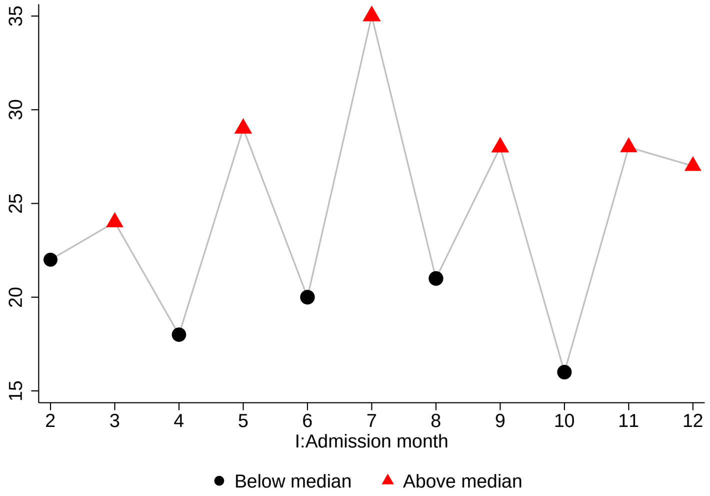
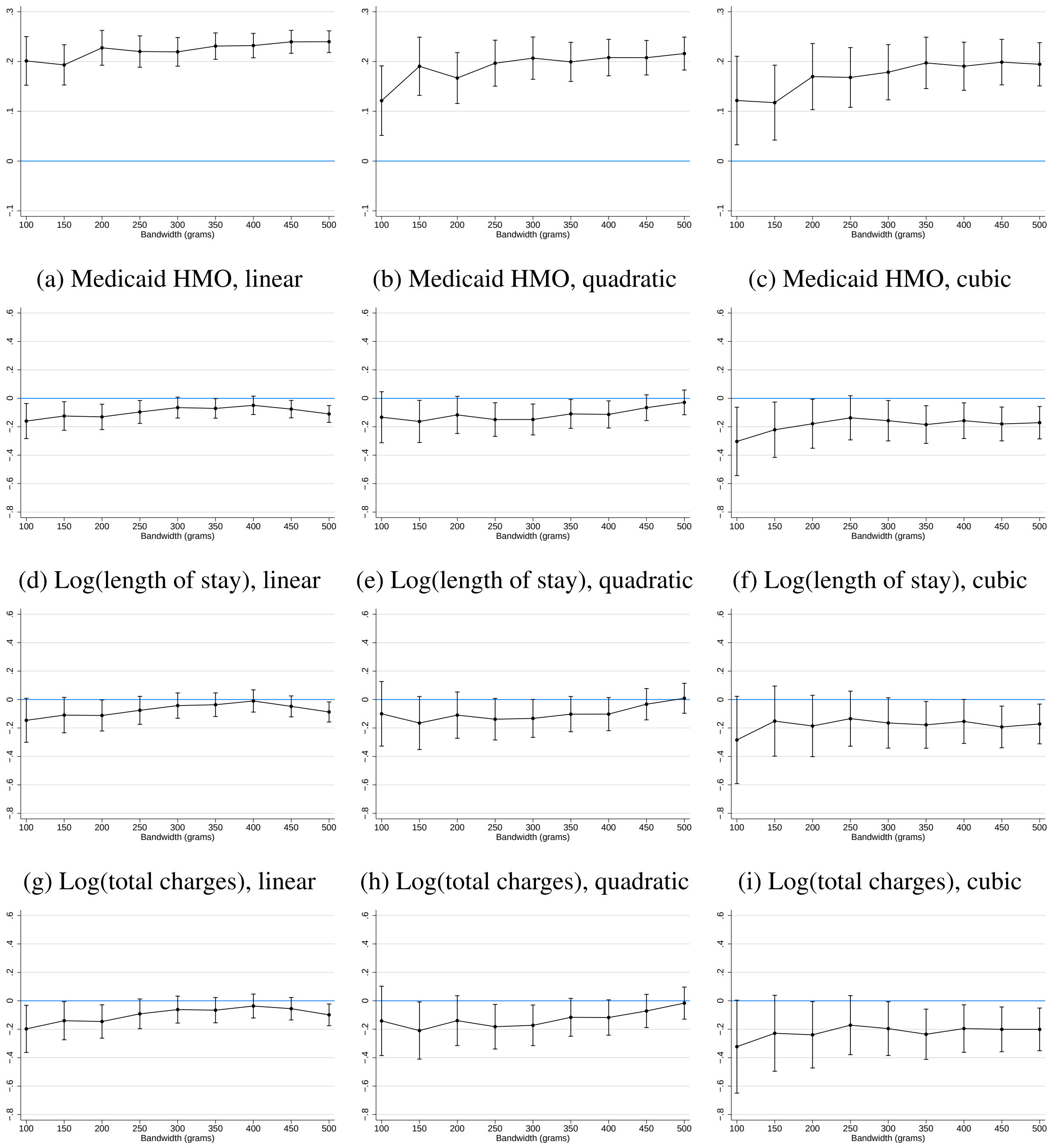


Figure C.6: An example hospital, 2005

Notes: This figures illustrates the monthly NICU occupancy for an example hospital in the year 2005. For instance, around 22 infants were admitted to NICU in January 2005 and stayed for at least 10 days. I use this value as an indication of the NICU occupancy for infants born in February. The figure shows that there is a large variation in the NICU occupancy across months.



(a) Medicaid HMO, linear (b) Medicaid HMO, quadratic (c) Medicaid HMO, cubic

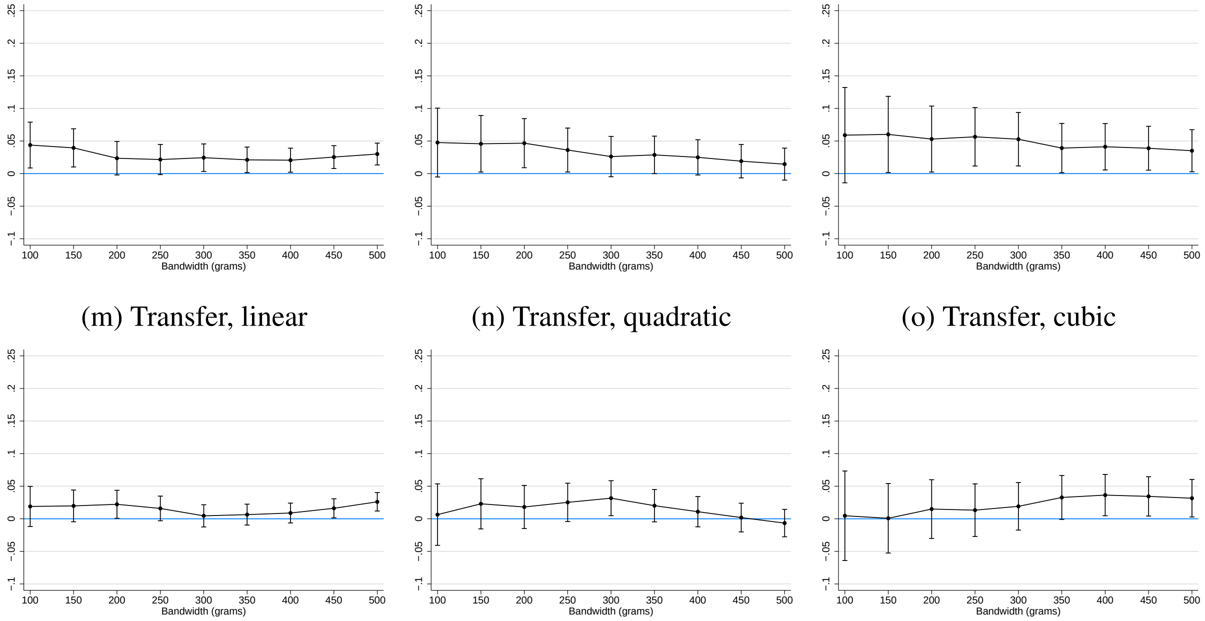
(d) Log(length of stay), linear (e) Log(length of stay), quadratic (f) Log(length of stay), cubic

(g) Log(total charges), linear (h) Log(total charges), quadratic (i) Log(total charges), cubic

(j) Log(total costs), linear (k) Log(total costs), quadratic (l) Log(total costs), cubic

Figure C.7: Sensitivity to bandwidth and polynomial, New York City

Notes: I repeat the estimation for each outcome for a different choice of bandwidth and polynomial. I use a range of bandwidths from 100 grams to 500 grams varying the degree of polynomials from degree 1 (linear), degree 2 (quadratic), to degree 3 (cubic).



(m) Transfer, linear

(n) Transfer, quadratic

(o) Transfer, cubic

(p) Mortality, linear

(q) Mortality, quadratic

(r) Mortality, cubic

Figure C.7: Sensitivity to bandwidth and polynomial, New York City (continued)

Notes: I repeat the estimation for each outcome for a different choice of bandwidth and polynomial. I use a range of bandwidths from 100 grams to 500 grams varying the degree of polynomials from degree 1 (linear), degree 2 (quadratic), to degree 3 (cubic).

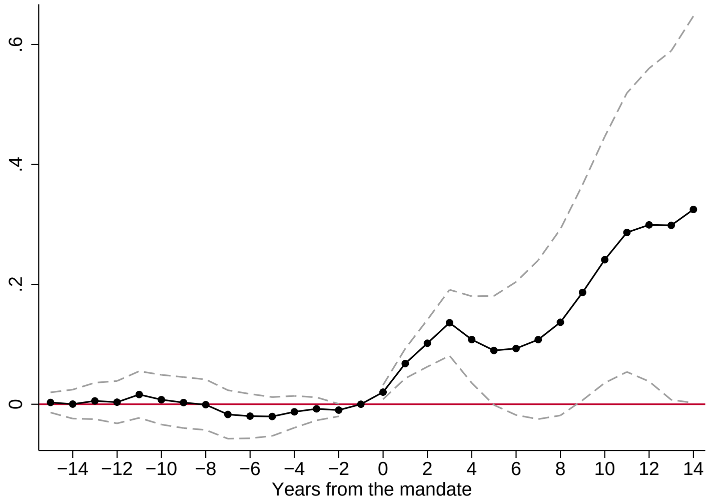


Figure C.8: Medicaid HMO participation by years from the MMC mandate

Notes: The above figure plots estimates from a regression of an indicator for Medicaid HMO participation on a set of dummies that indicate years from the MMC mandate for each county. County fixed effects, year fixed effects, and county-specific time trends are also included in the regression. The dashed lines plot 95% confidence intervals computed based on standard errors clustered at the county level.

Appendix D. Appendix Tables

Table D.1: Effects of birth weight≥1,200 grams on other health/quality outcomes, New York City

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Avoidable readmission	Level IV NICU stay	Any NICU stay	Chest X-ray	Ultrasound	Implant	Physical therapy	Respiratory therapy	Speech therapy
Above	0.002	-0.002	0.008	-0.028	-0.022	-0.006	0.041**	-0.007	0.006
	(0.013)	(0.019)	(0.017)	(0.025)	(0.021)	(0.011)	(0.019)	(0.020)	(0.019)
Observations	4065	4315	4315	3221	3221	3221	4315	3221	3221
Mean below cutoff	0.052	0.873	0.905	0.801	0.895	0.025	0.127	0.947	0.091
Mean above cutoff	0.043	0.869	0.905	0.754	0.885	0.024	0.118	0.893	0.080
Bandwidth (grams)	150	200	200	150	150	150	200	150	150

Notes: Column 1 shows the RD estimate for hospital readmission due to preventable conditions. Columns 2-9 show the RD estimates for utilization of various inpatient services at the individual level. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table D.2: Heterogeneity by NICU crowdedness, relative to the number of beds, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Below the median NICU occupancy relative to the number of beds						
Birth weight≥1,200 g	0.205***	-0.043	-0.048	0.024	0.018	0.020
	(0.035)	(0.078)	(0.101)	(0.098)	(0.026)	(0.030)
Observations	1266	947	942	732	1266	645
Mean below cutoff	0.017	53.0	$284,947	$107,507	0.058	0.036
Mean above cutoff	0.242	43.8	$253,561	$92,292	0.046	0.030
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Above the median NICU occupancy relative to the number of beds						
Birth weight≥1,200 g	0.244***	-0.222***	-0.249***	-0.221**	0.033*	0.028
	(0.030)	(0.076)	(0.092)	(0.098)	(0.019)	(0.029)
Observations	1744	1302	1298	982	1744	859
Mean below cutoff	0.016	53.1	$287,583	$106,648	0.040	0.040
Mean above cutoff	0.261	42.3	$230,545	$86,728	0.051	0.039
Bandwidth (grams)	200	150	150	150	200	100

Notes: I divide the monthly NICU occupancy measure by the number of NICU beds. Since the mean length of stay for infants who stay in NICU for at least 10 days is around one month, this measure roughly captures the daily occupancy rate in a given month. Panel A shows the RD estimates for months when this relative NICU occupancy rate is below the median for a given hospital in a given year. Panel B shows the RD estimates for months when the relative NICU occupancy rate is above the median for a given hospital-year. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table D.3: Heterogeneity by crowdedness at the typical destination, New York City

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Below the median NICU occupancy at the typical destination						
Birth weight≥1,200 g	0.228***	-0.194**	-0.236**	-0.219**	0.038	0.022
	(0.028)	(0.077)	(0.099)	(0.106)	(0.023)	(0.027)
Observations	1826	1349	1343	1015	1826	904
Mean below cutoff	0.019	52.4	$276,442	$106,429	0.067	0.027
Mean above cutoff	0.262	42.1	$228,440	$84,086	0.064	0.033
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Above the median NICU occupancy at the typical destination						
Birth weight≥1,200 g	0.261***	-0.133	-0.134	-0.279**	0.008	0.018
	(0.038)	(0.102)	(0.119)	(0.133)	(0.023)	(0.037)
Observations	1256	939	936	692	1256	647
Mean below cutoff	0.031	52.3	$264,805	$104,435	0.064	0.051
Mean above cutoff	0.320	43.6	$234,905	$90,673	0.062	0.039
Bandwidth (grams)	200	150	150	150	200	100

Notes: I define a typical destination hospital as the receiving hospital of the majority of any neonatal transfers from a given hospital. Panel A shows the RD estimates for months when the NICU occupancy at the typical destination hospital with a NICU is below the median, while panel B shows the RD estimates for months when the NICU occupancy at the typical destination hospital is above the median in a given hospital-year. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table D.4: Difference-in-difference estimates, other health/quality outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Any NICU stay	Chest X-ray	Ultrasound	Implant	Physical therapy	Respiratory therapy	Speech therapy
Panel A. Without county-specific time trends							
treat	0.004	-0.009*	-0.008*	-0.002***	-0.000	-0.012	-0.006
	(0.005)	(0.005)	(0.004)	(0.001)	(0.006)	(0.008)	(0.008)
Observations	1721856	1721856	1721856	1721856	1721856	1721856	1721856
Mean	0.129	0.077	0.061	0.003	0.012	0.074	0.007
Panel B. With county-specific time trends							
treat	0.005	-0.011*	-0.007**	-0.002	-0.000	-0.012***	-0.006*
	(0.007)	(0.006)	(0.003)	(0.001)	(0.003)	(0.004)	(0.003)
Observations	1721856	1721856	1721856	1721856	1721856	1721856	1721856
Mean	0.129	0.077	0.061	0.003	0.012	0.074	0.007

Notes: Panel A presents a difference-in-difference estimate for each outcome without including the county-specific trends. Panel B shows the estimates including the county-specific trends. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.

Table D.5: Heterogeneity by county-level median household income

	(1)	(2)	(3)	(4)	(5)	(6)
	Medicaid HMO	Log(LOS)	Log(total charges)	Log(total costs)	Transfer	Mortality
Panel A. Quartile 1						
Birth weight≥1,200 g	0.301***	-0.349***	-0.250**	-0.339***	0.078***	0.043
	(0.039)	(0.091)	(0.123)	(0.121)	(0.028)	(0.028)
Observations	1290	976	973	734	1290	688
Mean below cutoff	0.040	52.3	$252,267	$103,430	0.083	0.028
Mean above cutoff	0.324	42.6	$219,470	$81,610	0.099	0.032
Bandwidth (grams)	200	150	150	150	200	100
Panel B. Quartile 2						
Birth weight≥1,200 g	0.247***	-0.029	-0.010	-0.026	0.004	0.013
	(0.023)	(0.075)	(0.089)	(0.096)	(0.020)	(0.021)
Observations	3492	2599	2595	2107	3492	1721
Mean below cutoff	0.032	49.6	$194,713	$77,260	0.120	0.042
Mean above cutoff	0.296	40.1	$157,699	$62,790	0.114	0.036
Bandwidth (grams)	200	150	150	150	200	100
Panel C. Quartile 3						
Birth weight≥1,200 g	0.158***	0.014	-0.079	0.023	0.011	-0.006
	(0.030)	(0.091)	(0.099)	(0.108)	(0.023)	(0.025)
Observations	1497	1107	1101	939	1497	741
Mean below cutoff	0.019	53.9	$268,293	$98,228	0.064	0.026
Mean above cutoff	0.190	44.9	$214,479	$77,044	0.055	0.033
Bandwidth (grams)	200	150	150	150	200	100
Panel D. Quartile 4						
Birth weight≥1,200 g	0.118***	0.001	0.050	0.049	0.019	0.027*
	(0.019)	(0.071)	(0.081)	(0.083)	(0.020)	(0.016)
Observations	3782	2797	2787	2507	3782	1848
Mean below cutoff	0.037	49.5	$232,172	$79,622	0.115	0.033
Mean above cutoff	0.178	40.5	$196,712	$66,658	0.105	0.031
Bandwidth (grams)	200	150	150	150	200	100

Notes: Each panel shows the RD estimates for counties classified into each quartile of a county-level median income measure. I take an average of median household income levels across zip codes in each county to construct the county-level income measure. Here, I use county as a service area for a hospital since hospitals typically serve an area larger than a zip code. In addition to the indicator for birth weight≥1,200 g, each regression includes a linear spline of birth weight, admission year fixed effects, admission month fixed effects, and hospital county fixed effects. Robust standard errors are reported. The means of logged outcomes are reported in levels.

* Significant at 10%, ** significant at 5%, *** significant at 1%.