Essays on the Econometric Analysis of Forward-Looking Behaviour in Health Care Markets

Daniel Kaliski

Nuffield College, University of Oxford

Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Economics

Hilary Term, 2018

Supervisors: Michael P. Keane (University of New South Wales) and Ian Crawford (Nuffield College, Oxford)

Words: Approx. 53 000

Acknowledgements

This dissertation would not have been possible without the supervision of Michael P. Keane and Ian Crawford. I have benefited immensely from their invaluable guidance. Olivia S. Mitchell provided additional supervision for my job market paper, which is the first chapter of the dissertation. I am grateful to all three for their support on the academic job market, in addition to the sacrifices of time and energy they made in reading my work, commenting on it, and teaching me how to be a better thinker, writer, and economist.

In addition to my advisors, the first chapter of this thesis benefited from comments and advice from Jason Abaluck, Johannes Abeler, Abi Adams, Abby Alpert, Juan-Pablo Atal, Maria Balgova, James Best, Steve Bond, Martin Browning, Minsu Chang, Norma B. Coe, Frank DiTraglia, Rob Garlick, Olga Gdula, Ian Jewitt, Hanming Fang, Jesus Fernandez-Villaverde, Isaac (Zac) Gross, Michal Hodor, Rury Holman, Johannes Jaspersen, Michal Kolesar, Ines Lee, Hamish Low, Dan Polsky, Simon Quinn, Victor Rios-Rull, Paul Sangrey, Molly Schnell, Daniela Scur, Andrew Shephard, Petra Todd, Hanna Wang, and Peter Zweifel, as well as seminar participants at the 2017 National Tax Association Meetings (and my discussant there, David Powell), the 2018 European Health Economics Association (EuHEA) meetings, the 2018 European Economics Association meetings, the 2018 European Winter Meetings of the Econometric Society, the Munich Risk and Insurance Centre, Oxford and the University of Pennsylvania. The second chapter benefited greatly from conversations with Steve Bond, who together with Abi Adams also provided comments on an earlier version of the third chapter.

I am grateful to both the Rhodes Trust and the University of Oxford's Department of Economics for their financial support during my D.Phil. degree.

Second to last, I gratefully acknowledge the support of my friends Natalya Din-Kariuki, Arushi Garg, Ask Foldspang Neve, Meg Scott, Hanna Wang, Ines Lee, Daniela Scur, James Best and Simukai Chigudu - some of whom were subjected to more drafts of papers and practice job talks than I deserved to receive comments on.

Lastly, I have to thank my parents for their unwavering support throughout my time at Oxford.

In the short term I can only repay them with *naches*,[1] and I hope the completion of this dissertation will count towards it.

[1] Yiddish: happiness from the achievements of one's offspring.

Thesis Abstract

This dissertation makes three contributions to the economics literature. Chapter 1 finds that expanding health insurance that covers treatment but not prevention can discourage prevention to a large extent, when the prior literature only recovered small effects (if at all). It also uses a simple economic model to reconcile the results with estimates found in other studies. Chapter 2 shows that a regression-discontinuity design can be used to determine whether the long-run effect of a discontinuous change differs from the short-run effect traditionally estimated by these designs when the assignment variable is deterministically linked to time. Chapter 3 examines subjective medical expenses forecasts, and finds that they depart from the forecasts that would be made by a hypothetical agent that incorporated all of the relevant information in the decision-making environment.

Introduction

The three chapters of this thesis are investigations into what can be learned from data regarding three issues of interest to economists: the strength of *ex ante* moral hazard (Chapter 1), the difference between short-run and long-run effects (Chapter 2), and the formation of individuals' expectations regarding the future (Chapter 3). A key challenge that is addressed directly in each chapter is that data can typically support multiple competing hypotheses. Even with "quasi-experimental" variation from a natural experiment, the data do not "speak for themselves" - the inferences that can be drawn from the variation at hand will depend on the defensibility of auxiliary hypotheses rooted either in theory or supporting evidence from other contexts. In consequence, it has long been recognised in empirical economics that researchers have to make assumptions that rule out competing hypotheses in order to reject a null hypothesis in favour of a single alternative. The three chapters of this thesis evaluate the above-mentioned questions in the context of individuals' use of health care by using different combinations of assumptions and data. In particular, I examine what can be learned regarding forward-looking behaviour in the demand for health services without estimating a fully specified structural model which makes explicit assumptions regarding preferences, beliefs, and action spaces, and instead making more limited assumptions regarding the processes of interest from either economic theory or prior evidence (and hence drawing more conservative conclusions). The three chapters represent different approaches to dealing with the limitations of quasi-experimental analyses that eschew estimation of a fully specified structural economic model.

The first chapter, which was used as my job market paper, "Does Insurance for Treatment Crowd Out Prevention? Evidence from Diabetics' Insulin Usage", provides evidence that, contrary to the consensus in the health economics literature, there is significant *ex ante* moral hazard in health behaviors - insuring individuals against the financial consequences of ill health makes them less likely to take costly measures to avoid falling ill. The main results are from a regression-discontinuity design and a difference-in-discontinuities design that exploit the exogeneity of eligibility for Medicare coverage at age 65 for U. S. residents, and the expansion of that programme to include prescription drug coverage in 2006 (Medicare Part D). In addition to the empirical results, I also use a theoretical

framework derived from a simple model of the intertemporal allocation of preventive medicine to reconcile the results with the existing literature, and develop a partial identification strategy that exploits the availability of repeated observations on the same individuals before and after they qualify for the Medicare programme to relax the assumptions necessary for point identification of the main results.

In the second chapter, "Dynamic Adjustment Following Discontinuous Regime Changes", I address whether the short-run effects estimated in regression-discontinuity designs with age as the assignment variable have long-run effects beyond the level shifts at the cutoff that they recover. To do so, I derive a general method for testing for whether the short-run and long-run effects of a discontinuous change differ in contexts where the discontinuous change is linked to the passage of time, as in the case of individuals crossing an age-eligibility threshold. A sufficient condition for the long-run and short-run effects to differ is that there is a persistent change to the gradient of the expected outcome with respect to the assignment variable after the discontinuous change happens. I am able to show that this holds when the outcome of interest exhibits positive serial correlation, which is true of many economic variables of interest such as labour supply or consumption. Prior work has shown that any change in this gradient at the cutoff can be consistently estimated in a regression-discontinuity design by testing whether the slope of the local polynomial around the cutoff differs on its left- and right-hand sides, which is straightforward (though subject to larger biases for a given bandwidth than is the RDD estimator of the level shift at the cutoff). I then apply this method to the empirical setting in Chapter 1. I am unable to find evidence that the negative effect of insurance for treatment on insulin usage observed at age 65 in Chapter 1 cumulates or dissipates over the longer run. This is despite the fact that there are strong reasons to believe that the short-run and long-run effects differ based on simulation evidence, auxiliary evidence from the data that suggests there is positive serial correlation in insulin usage decisions, and economic theory (via the Grossman model of investment in health). The method derived in Chapter 2 may yet prove useful in other applications where the assignment variable is deterministically linked to time, or has some structure that admits positive serial correlation across units of observation.

In the third chapter, "Resolution of Uncertainty, New Information, and Accuracy in Subjective

Medical Expenditures Forecasts", I examine what can be learned from directly collecting data on individuals' stated beliefs regarding their future out-of-pocket medical expenditures. I use data from the Health and Retirement Study's 2010, 2012 and 2014 survey waves, in which individuals are asked to give the probabilities that their next twelve months' out-of-pocket medical expenditures will exceed each of two to three threshold amounts. The responses to these survey questions allow some questions regarding the accuracy of individuals' beliefs to be answered without distributional assumptions - in particular, I find that individuals do not incorporate the sequential resolution of uncertainty into their forecasts, and display ignorance of at least one facet of the United States' insurance environment (the entitlement of divorced or widowed spouses to remain on their former spouse's employer's health insurance plan). They nonetheless require some assumption on the subjective distribution of out-of-pocket medical expenditures in order to make dollar comparisons between predicted and realised expenditures. By assuming that they follow a lognormal distribution, as other authors have found to be the case for medical expenditures, I am able to further infer that individuals underestimate the extent to which both the mean and variance of future medical expenditures change when they gain access to health insurance or are diagnosed with a persistent medical condition such as cancer or diabetes.

Two themes are common to all three chapters. First, either economic theory or parametric assumptions supported by prior evidence are necessary to evaluate the relationship of the empirical results to quantities of interest. Second, repeated observations of the same individuals over time allow for analyses that would not be possible with repeated cross-sections alone. Both of these have a long tradition in empirical economics, with the first one going back at least to the work of the Cowles Commission. In quasi-experimental empirical work of the kind found in this thesis it has been less common to make explicit reference to economic models than is done in Chapters 1 and 2. It has been relatively more common for them to use panel data; this thesis departs from the usual use of panel data to achieve identification on its own, instead using it to provide auxiliary tests of supporting assumptions (Chapter 2) or to relax assumptions behind point identification and allow for partial identification (Chapter 1).

The first theme arises in the first chapter via its use of an economic model to reconcile the em-

pirical results with other studies and relate the measured responses to hypothetical compensated lifetime responses. In the second chapter extensive reference is made to the Grossman model of human capital accumulation to motivate why we should expect to reject the null hypothesis that there is no difference between the short-run and long-run effects of providing insurance for treatment on investment on health. In the third chapter, even directly elicited data on subjective expectations of future medical expenses cannot yield information on the central moments of individuals' subjective medical expenses distributions without some assumption regarding the underlying distribution from which out-of-pocket medical expenses are drawn.

The second theme - that panel data allows for inferences that are difficult or impossible in pooled cross-sections - arises most simply in the third chapter: without observing how the same individuals revise their forecasts of medical expenses before and after an unforeseen event, we should have to make strong assumptions regarding how these forecasts react to new information. In the first and second chapters panel data allows for auxiliary information that supplements the main analyses, either by allowing for partial identification instead of point identification (Chapter 1) or by allowing for a test of positive serial correlation (Chapter 2), which is an important case in which long-run and short-run effects of a policy change differ.

Contents

1 Does Insurance for Treatment Crowd Out Prevention? Evidence from Diabetics' Insulin Usage **12**

1.1 Introduction 13

1.2 Medical and Institutional Background 17

1.3 Empirical Strategy 23

1.3.1 The (Fuzzy) Regression-Discontinuity Estimator 23

1.3.2 Difference-in-Discontinuities 26

1.3.3 The Quasi-Difference-in-Differences Design 27

1.4 Data 28

1.5 Results 32

1.5.1 Crowding Out of Prevention by Insurance for Treatment: 1998-2004 34

1.5.2 Mechanisms: Crowding Out via *Ex Ante* Moral Hazard 38

1.5.3 The Effect of Prescription Drug Coverage on Ex Ante Moral Hazard and Aggregate Outcomes 40

1.5.4 Mechanisms: Prescription Drug Coverage Under Part D Post-2006 45

1.5.5 Aggregate Effects of Prescription Drug Coverage: 1998-2014 . . . 47

1.5.6 Summary of Findings 53

1.6 Theoretical Framework: A Model of the Intertemporal Allocation of Prevention 54

1.7 Conclusion 64

2 Dynamic Adjustment Following Discontinuous Regime Changes 106

2.1 Introduction 107

2.2 Nonparametric Identification of Differences Between Short-Term and Long-Term Effects 108

2.3 Monte Carlo Evidence on the Performance of the Difference-in-Slopes Method 115

2.4 Application: The Insulin Usage of Diabetics Before and After Qualifying for Medicare Coverage at Age 65 123

2.5 Conclusion 127

3 Resolution of Uncertainty, New Information, and Accuracy in Subjective Medical Expenditures Forecasts 145

3.1 Introduction 146

3.2 Review of the Literature 148

3.3 Data 150

3.4 Empirical Strategy 158

3.4.1 Event-Study Design for Anticipated and Unanticipated Shocks . 158

3.4.2 Assessing the Accuracy of Forecasts 161

3.4.3 Translating Subjective Medical Expenses Forecasts into Changes in Subjective Means and Variances 163

3.5 Results 166

3.5.1 Incoherent Subjective Forecasts of Out-of-Pocket Medical Expenditures 167

3.5.2 Revisions to Forecasts Due to Anticipated Receipt of Health Insurance 169

3.5.3 Revisions to Forecasts Due to Unanticipated Persistent Changes 172

3.5.4 Implied Changes to Means and Variances of Subjective Medical Expenses Distributions 176
3.6 Conclusion 178

Chapter 1

Does Insurance for Treatment Crowd Out Prevention? Evidence from Diabetics' Insulin Usage

Abstract

I provide new evidence that health insurance can discourage investment in health. I find that, in the United States before 2006, 13-30% of female diabetics who used insulin to manage their condition stopped using insulin once they turned 65 and became eligible for health insurance via Medicare. I reconcile these results with those from other studies by developing a model of the trade-off between prevention and treatment. The model explains the large effect sizes in this chapter via two mechanisms. First, individuals substitute prevention efforts away from periods when the price of treatment is low and toward periods when the price of treatment is high. Second, this effect is stronger for preventive measures that have larger effects on health. The model also shows that the long-term crowding out of prevention is at least as large as the shift in the timing of prevention estimated in this chapter. The introduction of more generous subsidies for insulin under Medicare Part D in 2006 eliminated this effect, saving up to $487 million per annum in forgone health care costs.

1.1 Introduction

If treating an illness is cheap, do the people at risk of developing it take fewer precautions to avoid it? The financial burden is only one component of the utility loss due to illness (Einav and Finkelstein, 2018, Kenkel, 2000), so we cannot reason from first principles about whether it is important enough to greatly affect behaviour. To answer this question, we need evidence from a setting with two properties. First, there is a change in the price of treatment without a corresponding change in the price of investment in health. Second, this reduction in the risk of incurring a large medical bill when ill must be large enough to play a significant role in health investment decisions.

This chapter studies the decision to invest in health in a setting with both properties. Americans with diabetes gain access to coverage for treatment at age 65 via Medicare, but in most cases did not gain coverage for insulin before 2006. Insulin usage is also strongly linked to the probability of suffering complications of diabetes (Diabetes Control and Complications Trial Research Group, 1993), which in turn are expensive to treat when they arise (Bommer et al., 2017, Zhuo et al., 2014, Salas et al., 2009, Gilmer et al., 1997).

My first main finding is that insulin was used less often when American diabetics qualified for Medicare coverage before 2006. Before 2006, diabetic women are between one and a half and twice as likely as diabetic men to have no form of health insurance before they qualify for Medicare. I find a reduction of 7.9 percentage points in the proportion of diabetic women who report using insulin to manage their diabetes when they qualify for Medicare in this period, from a baseline of 26 percentage points. I find no evidence of offsetting increases in other preventive behaviors, such as diet, exercise, or use of oral medication. After 2006, the Medicare program was expanded to include a prescription drug benefit (Part D) which included private plans with more generous coverage for insulin, which was only covered in special cases - or with prohibitive coinsurance rates of up to 50% - before 2006. The second main finding of this chapter is that this expansion of the program cancelled out the *ex ante* moral hazard effect of providing coverage for treatment. I am unable to reject the hypothesis that this offset was twice as large as the original negative effect, so that qualifying for Medicare coverage from 2006 onward had a net *positive* effect on the likelihood

of insulin usage.

Expanding Medicare to include prescription drug coverage is likely to have saved up to $487 million per annum in health care costs among female diabetics. This is partly due to a flatter trend in the incidence of heart disease in this group, for whom I find the strongest evidence for *ex ante* moral hazard in insulin usage. This is in line with research which shows that heart disease is the most common complication of diabetes, and that diabetic women's risk for cardiovascular complications is much greater relative to non-diabetic women than diabetic men's risk is relative to non-diabetic men (Juutilainen et al., 2004). My calculations indicate that these cost savings are up to one-third as large as those that would result from a similarly effective tobacco control program. Since the latter is widely believed to be the most effective method of improving population health in the developed world, the findings in this chapter imply that insulin is in the first rank of effectiveness among public health initiatives. At the same time, the return on investment for insulin subsidies may be significantly lower in the United States than in the past relative to elsewhere in the developed world due to the rapid increase in the price of insulin in that country since 2006.

My results cast doubt on the current consensus that expanded coverage unambiguously improves population health (Sommers et al., 2017). While this is likely to be true in the aggregate, jointly providing coverage for treatment and prevention can mask the moral hazard effects of coverage for treatment, which can crowd out investments in health at the margin. It may still be that universal health care regimes are better at incentivising investments in health by being more likely to pay for them, since this is one of the main methods by which they hold down overall costs. The main policy implication of this chapter's results is that policymakers have nonetheless underestimated the extent to which these incentives are necessary even where they are already provided.

I focus on diabetics in particular for three main reasons. First, diabetes is one of the fastest-growing noncommunicable diseases in the world. Recent forecasts have estimated that the global diabetic population will have more than doubled between 2000 and 2030, from 171 million to 366 million people, even if obesity rates remain constant (Wild et al., 2004). In the United States, the proportion of the population with diabetes has been estimated at between 12 and 15 percent (Menke et al., 2015).

Second, diabetics typically have high medical expenses that are closely linked to how well they are able to manage their condition. There are individual actions, such as injecting insulin, that are closely tied to their eventual health outcomes, which is not true of most individuals, or even most individuals with chronic conditions. This allows me to circumvent the usual problems encountered by studies of moral hazard in health behaviors, where any single practice typically has a limited marginal contribution to the costs of care.

This also means that incentives for better or worse self-management among this group matter for the eventual costs of their health care, which typically have a high social cost. In the United States, most of their medical expenses paid for by the Medicare program after age 65. The global burden of diabetes has recently been estimated at $1.31 trillion U.S. dollars, or 1.8% of global gross domestic product (GDP) (Bommer et al., 2017), and 65% of those costs were estimated to result from the direct costs of maintaining the health of diabetics or treating complications due to their condition.

The third reason is that prior to the passage of the Patient Protection and Affordable Care Act (ACA) in 2010, diabetics were routinely ineligible for privately purchased insurance in the United States due to their pre-existing condition. As a result, treatment for medical complications arising from insufficient control of their condition was not just uninsured but *uninsurable* if they didn't have access to insurance via their or their spouse's employer or the low-income health insurance program Medicaid. Hence their risk of incurring large medical expenses at age 65 changed from "background risk" to insurable (and insured) risk. A large literature spanning both macroeconomics and microeconomics focuses on the different implications for behavior of uninsurable and insurable risk (Aiyagari, 1994, Carroll et al., 2003, Curcuru et al., 2010, Eeckhoudt et al., 1996, Guerrieri and Lorenzoni, 2017). Differences in diabetics' behavior when faced with background risk and insured risk shed light on the relative importance of the two for choice under uncertainty. They afford us an answer to the question "what would happen if we converted uninsurable background risks to risks against which agents had insurance?".

In addition, this is the first study to use the distinction between the Marshall, Hicks, and Frisch elasticities of a decision with respect to a price change in order to analyze *ex ante* moral hazard

in prevention.[1] The model from which I derive these elasticities has three functions. First, it allows me to reconcile my results with those from other studies. Second, it sheds light on the distributional effects of crowding out prevention - those for whom prevention matters most are the same individuals for whom it is crowded out the most. Third, it makes quantitative predictions regarding other responses such as the income elasticity of investments in health.

The Hicks elasticity corresponds to the pure substitution effect due to the change in relative prices resulting from an unexpected change in the price of health care relative to other spending. The Marshallian elasticity corresponds to the pure substitution effect of the Hicks elasticity plus a countervailing income effect - since individuals are richer, they buy more prevention. The Frisch elasticity is the elasticity of intertemporal substitution: it is the effect on differences in the usage of preventive care across periods of differences in the price of treatment across periods, holding the marginal utility of lifetime wealth constant. These distinctions allow me to explain differences in results between the literature on experimental results (which contain income and wealth effects, and hence recover the Marshallian elasticity) and the literature that uses Medicare eligibility as part of a regression-discontinuity design (which recovers Frisch elasticities, since the estimated responses are long-anticipated reactions to eligibility, and hence the result of intertemporal substitution). In addition, it is able to explain why studies that use Medicare RDDs, such as Card et al. (2008) and Dave and Kaestner (2009) do not recover large effects: the Frisch elasticity is small for behaviors such as the marginal half-hour of exercise that are not close substitutes with treatment *ex post*, i.e. do not have a close connection with individuals' need for treatment *ex post*. I leave the estimation of Hicks elasticities of prevention with respect to the price of treatment to future research.

The rest of this chapter is organised as follows. Section 2.2 presents the medical and institutional features that define diabetics' incentives and constraints in the U.S. health care system in the period 1998-2010. Section 1.3 outlines the identification arguments for the standard regression-discontinuity framework, the difference-in-discontinuities approach used to recover the effect of Part D, and the difference-in-differences regressions used to recover the aggregate effects on health and health care costs. Section 1.4 describes the data. Section 1.5 presents the empirical results

[1] See Keane (2011) for a review of the literature on estimating these quantities for the response of labor supply to changes in wages and/or taxes.

obtained by applying the methods described in Section 1.3 to the data described in Section 1.4. Section 1.6 develops a life-cycle model of prevention to explain the differences between this study's results and those in the rest of the literature, as well as the results' relationship to longer-term effects on prevention. Section 1.7 concludes.

1.2 Medical and Institutional Background

Diabetes is a disorder where the cells of the body do not respond to insulin (insulin resistance). Insulin's function is to regulate blood sugar levels. Since insulin decreases blood sugar levels, inability to absorb insulin results in both higher levels of blood sugar and higher volatility of blood sugar levels, both of which are corrosive to the blood vessels within the human body. As a result, diabetics are more likely to experience both disorders of the major blood vessels ("macrovascular" complications) such as heart attacks and strokes and disorders of the small blood vessels ("microvascular" complications) such as retinopathy (which results in blindness), neuropathy (nerve damage, which can cause ulcers and often necessitates limb amputation), and nephropathy (kidney failure). The latter category of disorders - microvascular complications - is observed at a much higher frequency among diabetics than individuals with other chronic conditions.

Injecting insulin is a form of preventive care for the vast majority of diabetics. For Type I diabetics, roughly 10% of the total diabetic population, poor control of their blood sugar levels will quickly result in life-threatening complications. For Type II diabetics insulin usage is a forward-looking behaviour where the short-term cost of purchasing insulin and blood glucose strips to monitor blood sugar levels is weighed against the longer-term costs of hospitalisation and medical complications. Type Is are typically diagnosed in childhood, while Type IIs develop the disease from middle age onwards. The rate of progression of the disease is highly individual-specific, and depends in part on adherence to prevention regimens that aim to reduce the level and volatility of blood glucose. Unlike Type Is, Type IIs are most likely to be recommended to use insulin only once their disease has progressed to the point where intermediate methods for controlling blood sugar

levels such as dieting or oral medication have become ineffective.[2] This will be significant for the empirical strategy in this chapter, since healthier diabetics are both more likely to have avoided needing insulin to manage their condition and more likely to survive to older ages.

Importantly for the results in this chapter, there are also notable physiological differences between male and female diabetics (Kautzky-Willer et al., 2016). Most women are less likely to suffer from cardiovascular disease than men. Diabetic women still have lower rates of heart disease than their male diabetic counterparts, but the relative risk of heart disease compared to their non-diabetic counterparts is higher for women than for men (Kautzky-Willer et al., 2016). As a result, we should expect that if diabetic women's preventive behavior changes and diabetic men's does not, the main differences in health outcomes between the genders should be cardiovascular. This is in fact what I find in this chapter (Section 1.5.5).

I first focus on the period 1998-2006 in this chapter for three reasons. First, a sizeable number of individuals enrolled in Medicare in this period did not have prescription drug coverage. In 1997, only 44 percent of Medicare beneficiaries had some form of prescription drug coverage (Soumerai and Ross-Degnan, 1999); by 2005, this had fallen to 35 percent (Soumerai et al., 2006). Piette et al. (2004) found in a 2002 survey of diabetics that 28% reported forgoing essential purchases such as food to pay for their medications, with 19% reporting nonadherence due to the high cost of their medications. Since there has never been a generic drug that can substitute for branded insulin, these figures are likely even higher for insulin than for diabetes medications such as metformin. In 2006, by contrast, prescription drug coverage was made available to Medicare Beneficiaries with the rollout of Medicare Part D, which also provided plans with more generous coinsurance rates for insulin than had previously been available, covering up to 100% of the cost of insulin purchases in some cases.

Without coverage for insulin, purchasing it independently could be prohibitively expensive, in part because there is no generic form of insulin. For example, Eli Lilly's fast-acting insulin, Humalog, cost $34.81 per vial (which would typically contain a month's worth of insulin) in 2001. This would amount to a yearly cost of $416.72. This is a modest estimate since many diabetics

[2]Unfortunately, I cannot distinguish between Type I and Type II diabetics in the data used for this chapter.

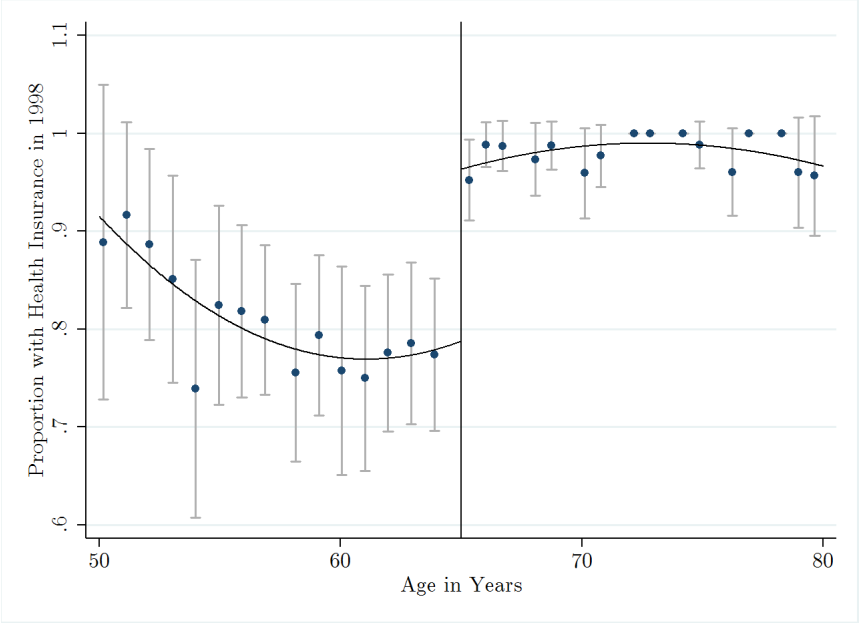
will require more than one vial's worth of insulin per month. Diabetics who use a "basal-bolus" regime, so called because it combines a baseline daily dose of insulin (the "basal" part) with regular injections before mealtimes (the "bolus" part), will require 9 vials every two months on average. This amounts to a yearly cost of $1879.74 in 1998 dollars. In 1998 200% of the federal poverty line (which would exclude the possibility of qualifying for Medicaid) outside of Alaska and Hawaii for a two-person household was $21 700. Therefore an uninsured married couple with one diabetic member could expect to spend 7% of total household income on insulin alone if they were at 200% of the federal poverty line in 1998. Moreover, even with coverage that included a prescription drug benefit, insulin's dual status as a non-generic drug and an "injectable" often led to small rates of reimbursement for insulin on health insurance plans (discussed in more detail below).

Second, in 1997 the United States passed the Balanced Budget Act (BBA), which contained several adjustments to the treatment of private health insurers that offered Medicare beneficiaries different packages of coverage as an alternative to traditional fee-for-service Medicare. Starting the sample period in 1998 therefore allows for relative stability in the Medicare program over the pre-2006 portion of the data. Third, the Health and Retirement Study (HRS) added several cohorts in 1998, greatly increasing the sample size, which is particularly useful when examining a subset of the full sample.

As a result, the sample period can be divided into two separate health care regimes. The first is the one which prevailed in 1998-2006, after the Balanced Budget Act of 1997 but before the 2006 implementation of the Medicare Modernization Act (MMA). The second is the one that prevailed in 2006-2010, after the Medicare program had been expanded to include prescription drug coverage under Medicare Part D, but before the passage of the Patient Protection and Affordable Care Act (ACA) in 2010. In Section 1.5.5, the aggregate results include data from the years 2010-2014. In that Section, I discuss why the passage of the ACA does not pose a significant threat to my ability to attribute the observed aggregate changes to the rollout of Part D.

Consider an American diabetic who is younger than 65 in 1998. Her health insurance options will depend upon the severity of her illness. In the worst case scenario, where her disease has already progressed to End-Stage Renal Disease (ESRD), also known as kidney failure, she will qualify for

Figure 1.1: Regression Discontinuity Plot - Insurance Status of Diabetics Before and After Age 65 in 1998



Notes: The figure represents binned data by age group with a second-order polynomial fit either side of the cutoff. The kernel used is the Uniform kernel. Medicaid recipients are excluded from the calculations. The dependent variable is an indicator variable equal to one if an individual reports having health insurance through any of the following sources: their employer, their spouse's employer, their union, veterans' agencies (Tricare), Medicaid, Medicare or a privately purchased plan.

Medicare, which is normally only available to over-65s. If she has made a successful application for disability benefits (SSDI) and been collecting them for two years, she will also qualify for Medicare despite being under 65. If her income and assets are low enough, she can qualify for her state's Medicaid program, which will give her both coverage for hospitalisations and subsidies for insulin, blood glucose strips and other supplies necessary to manage her condition. If she is not eligible for Medicaid, and is well enough to work, she will be reliant on her employer, her spouse's employer, or (in some cases) her trade union to enrol her in a health insurance plan. Private insurance plans outside of employer-based plans will almost certainly deny her coverage on the basis that she has a pre-existing condition (employer-based plans could only do this for a year after an employee is hired; thereafter, they become part of the group-based insurance plan offered by their employer's

insurance provider). This shows up in the data as a much larger gain for female diabetics than male diabetics in access to coverage at age 65, which cannot be discerned in the aggregate change displayed in Figure 1.1, but can be observed in Table 1.1. Female diabetics who are not enrolled in Medicaid, the publicly provided health insurance program for low-income populations in the United States, are consistently between one and a half and twice as likely to be uninsured before they qualify for Medicare coverage at age 65, with the gap only narrowing significantly in the last survey year before the Affordable Care Act (ACA) is passed in 2010. This is likely due to the lower attachment of these cohorts of women to the labor market before age 65. It is for this reason that the remainder of this chapter focuses on changes in female diabetics' health behaviors.

Once she turns 65, she will become eligible for Medicare Parts A and B, which will cover her for treatment and doctor's appointments. (Part A is for "inpatient" services such as hospital stays, whereas Part B is for "outpatient" services such as doctor's visits, X-rays, outpatient surgeries and laboratory work). Traditional Medicare will not, however, cover her insulin unless she is one of the rare individuals who is recommended by her doctor to use an insulin pump, in which case 80% of the cost of the pump and its insulin will be paid for by Medicare Part B. In the overwhelming majority of cases she will be liable for the costs of preventing a serious medical episode, but not for the costs of treating her once it occurs. Diabetics could opt to receive their Medicare coverage through a privatised plan on the Medicare+Choice programme (renamed Medicare Advantage in 2003), which did typically provide coverage for prescription medications, but access to insulin would still face the

Table 1.1: Percentage of Diabetics Uninsured Ages 60-64 By Gender (Excl. Medicaid Recipients), 1998-2008

	1998	2000	2002	2004	2006	2008
Men	15.98	14.65	9.70	7.11	7.74	12.50
Women	25.87	26.36	15.32	18.60	16.33	14.29

Notes: Each cell in the top row displays the percentage of diabetic men who are uninsured, excluding participants in Medicaid, the public health insurance program for low-income low-asset United States citizens, for a given wave of the Health and Retirement Study (HRS). Each cell in the bottom row reports the same percentages for women.

following obstacles. First, a growing proportion of these plans in 1998-2006 - 26% in 2002 (Christian-Herman et al., 2004) - would restrict prescription drug benefits to generics, which excludes insulin since there was and is no generic form of insulin. Second, these plans were mostly available to urban Medicare beneficiaries, since firms offering Medicare Advantage plans have operated in almost no rural counties (McGuire et al., 2011), and more rural states such as Alabama tended (and still tend) to have the highest incidences of diabetes. Third, insulin is usually classed differently from most prescription medications as it is an "injectable" and if covered typically involves higher co-payments than other prescription medications (Joyce et al., 2007, Boland, 1998). For example, Boland (1998) gives the example of one of the largest Health Maintenance Organisations (HMO) in New York, Independent Health, increasing its coinsurance rate for injectables to 50% in 1997, before the beginning of the sample period. Moreover, though there was an active effort in the United States Congress' Balanced Budget Act of 1997 to encourage take-up of Medicare+Choice plans, initial enrollment was low and declined as many insurers exited the Medicare+Choice market (McGuire et al., 2011). It is for these reasons that it is likely that the "intensive-margin" effects of more generous prescription drug coverage on Medicare are likely to be small for insulin usage over the period studied. Further evidence that another intensive-margin effect - lowering the price of generic medications paid for by a Medicare+Choice plan - did not induce significant substitution towards using generic oral medications to manage diabetes is presented below. There appears to be no corresponding upward spike in self-reported usage of oral diabetic medications at age 65, which one would expect if the subsidies provided by Medicare+Choice plans were an important countervailing factor.

Before 2006, the previously uninsured who gained coverage via the Medicare program could also purchase supplemental insurance only offered to Medicare beneficiaries, commonly known as Medigap (since it fills the "gaps" in Medicare coverage). The number and type of Medigap plans varied (and varies) from state to state, but prior to 2006 only 3 out of 10 Medigap plans included prescription drug coverage, and even those that did were likely subject to similar restrictions on insulin coverage as above. In sum, only low-income, low-asset diabetics who qualified for Medicaid coverage in the period before 2006 were guaranteed full coverage for the costs of using insulin.

In the next section, I describe the strategy I employ to estimate the effect of coverage on insulin usage.

1.3 Empirical Strategy

The first central goal of this chapter is to provide evidence for a substantial negative effect of insurance for treatment on prevention. In this section, I will describe the regression-discontinuity design that I use to identify this effect, the difference-in-discontinuities design that allows this effect to change when the policy regime changes, and the quasi-difference-in-differences method I use for estimating the effects of this policy change on aggregate health outcomes.

1.3.1 The (Fuzzy) Regression-Discontinuity Estimator

In this subsection, I will outline the usual conditions for consistency of the regression discontinuity (RD) estimator.

Suppose we have a panel of observations with individuals indexed $i \in \{1, ..., N\}$ in periods indexed $t \in \{1, ..., T\}$ for some outcome Y_{it}, a vector of covariates X_{it}, a running variable R_{it} with some discontinuous change in program assignment at $R_{it} = \bar{R}$, time-invariant unobserved heterogeneity η_i, idiosyncratic unobserved shocks v_{it}, some (usually polynomial) functions $f(.)$ and $g(.)$, which are continuous in R_{it} at $\bar{R}$ and have parameter vectors γ_0 and γ_1 respectively, and a dummy indicator variable for assignment to "treatment" (in this setting, coverage for medical treatment) D_{it}, and denote by h the bandwidth - the absolute distance from the cutoff that determines whether an observation is included in the sample or not, and by $K(.)$ some kernel function, both of which are chosen at the discretion of the econometrician,

$$Y_{it} = \beta_0 + \beta_1 1[R_{it} \geq \bar{R}] + f(R_{it}, \gamma_0) + g(R_{it}, \gamma_1) \times 1[R_{it} \geq \bar{R}] + \delta X_{it} + \zeta t + \eta_i + v_{it} \text{ for } K\left(\left|\frac{R_{it} - \bar{R}}{h}\right| < 1\right); \tag{1.1}$$

I use local linear regression (so that $f(.)$ and $g(.)$ are linear) and the Uniform kernel throughout, so that $K(.)$ is just an identity function, and so the regressions are restricted to observations for

which $\left|\frac{R_{it}-\bar{R}}{h}\right| < 1$ (see the empirical specification, Equation 1.4 in Section 1.5). Local linear regression has the advantage of putting the least weight, of all local polynomial regressions, on observations far from the cutoff (Gelman and Imbens, 2018). The Uniform kernel does not have this same advantage - the Edge (Triangular) kernel places more weight than it does on observations near the cutoff, for example - but does have the advantage of transparency, since the weights that are placed on different observations by other kernels are often difficult to interpret, which leads to difficulty interpreting differences across results that use different kernel weighting functions. It is for this reason that Lee and Lemieux (2010) recommend using the Uniform kernel and presenting a variety of results using different bandwidths, h, to make the empirical analysis easier to assess. I present evidence in Section 1.7 on the effects of varying the bandwidth on the results. I do not, however, use this analysis for bandwidth selection, as this introduces pre-test bias.[3] Instead, I use the mean-squared error (MSE)-optimal bandwidth derived by Calonico et al. (2014a).

The (sharp) regression discontinuity (RD) estimand is a comparison of outcomes just above the cutoff, for some $R_i^+ > \bar{R}$, and just below the cutoff for some $R_i^- < \bar{R}$ (omitting the covariates for simplicity),

$$\lim_{R_{it}^+ \to R_{it}^-} [E[Y_{it}|R_{it} = R_{it}^+] - E[Y_{it}|R_{it} = R_{it}^-]]$$

$$= \beta_1 + \lim_{R_{it}^+ \to R_{it}^-} [f(R_{it}^+) - f(R_{it}^-) + g(R_{it}^+) - g(R_{it}^-) + [E[\eta_i + v_{it}|R_{it} = R_{it}^+] - E[\eta_i + v_{it}|R_{it} = R_{it}^-]]$$

which recovers β_1, the difference in average outcomes between those who are exposed to the treatment just above the cutoff and those who are just below the cutoff, if and only if (since $f(.)$ and $g(.)$ are

$$\lim_{R_{it}^+ \to R_{it}^-} [E[\eta_i + v_{it}|R_{it} = R_{it}^+] - E[\eta_i + v_{it}|R_i = R_i^-]] = 0$$

- equivalently, $E[\eta_i|R_i]$ is continuous in R_i at the cutoff $\bar{R}$. Intuitively, if no other unobserved characteristics change discontinuously at the cutoff, then the observed change in outcomes can be

[3] Armstrong and Kolesár (2017) derive critical values that are robust to this bias.

attributed to the observable change in policy at the cutoff. For example, if some other behavior changes discontinuously at the cutoff, then the observed difference in outcomes could be due to that behavior rather than the observed difference in treatment status. For example, if retired individuals are more likely to use insulin (due to the greater time available to them for the management of their disease), then in a cross section a spurious discontinuous increase in the proportion of those in work at age 65 in a cross-section would exaggerate the effect of gaining insurance.

If some individuals have access to the treatment of interest (gaining coverage, say) without being eligible for the program of interest, then the regression discontinuity design is said to be "fuzzy" instead of "sharp", and since it compares outcomes across individuals it is conventional to scale the difference in mean outcomes between the two groups for the proportion of individuals who change status at the cutoff. In this case the regression discontinuity estimand results from two-stage least squares (2SLS) estimation with D_{it} replacing the assignment indicator $1[R_{it} \geq \bar{R}]$ in Equation 1.1 and $1[R_{it} \geq \bar{R}]$ used to instrument for D_{it}. Define $Z^+ = \lim_{R_{it} \to \bar{R}} E[Z_{it}|R_{it} \geq \bar{R}]$, and $Z^- = \lim_{R_{it} \to \bar{R}} E[Z_{it}|R_{it} < \bar{R}]$ for any variable Z_{it}. Then the "fuzzy" treatment effect recovered by the 2SLS estimator of β_1, scaled for the difference in the proportion of individuals treated above and below the cutoff, is

$$\frac{Y^+ - Y^-}{D^+ - D^-},$$

if we define p to be the sample fraction of individuals who have access to "treatment" below the cutoff, so that $p = \hat{D}^-$, and all individuals above the cutoff are treated, we obtain that this denominator is $1 - p$, which in this context is the fraction of uninsured individuals before age 65. When comparing the fuzzy regression discontinuity estimates with the first-differences estimates, I will need to rescale them so as to obtain the implicit "treatment-on-the-treated" difference in means, which will be $1 - p$ multiplied by the estimate.

In practice, for a given bandwidth h, the estimators of Y^+ and Y^- for some kernel function $K(.)$ are (Hahn et al., 2001)

$$\hat{Y}^{+}(\bar{R}) = \frac{1}{n}\sum_{i=1}^{n} Y_{it} 1[0 < K\left(\frac{R_{it} - \bar{R}}{h}\right) < 1],\ \hat{Y}^{-}(\bar{R}) = \frac{1}{n}\sum_{i=1}^{n} Y_{it} 1[0 < K\left(\frac{\bar{R} - R_{it}}{h}\right) < 1];$$

which are consistent estimators of Y^{+} and Y^{-}.

1.3.2 Difference-in-Discontinuities

I also estimate the effect of Part D, the 2006 expansion of the Medicare program that provided prescription drug coverage at age 65, on insulin usage among diabetic Medicare beneficiaries. More generous coverage options following the implementation of Part D in 2006 are likely to have decreased the price that Medicare-eligible diabetics faced for insulin. This will require a version of Equation 1.1 augmented to allow for the change at the cutoff to differ by regime period, viz.:

$$Y_{it} = \beta_0 + \beta_1 D_{it} + f(R_{it}, \gamma_0) + g(R_{it}, \gamma_1) \times 1[R_{it} \geq \bar{R}] + \beta_2 1[t \geq 2006] + \beta_3 D_{it} \times 1[t \geq 2006] + \delta X_{it} + \zeta t + \eta_i + v_{it}; \tag{1.2}$$

with, as before, the included observations satisfying $K(|\frac{R_{it} - \bar{R}}{h}| < 1)$. Interest will then focus on β_3, the marginal effect of the regime change in 2006, and $\beta_1 + \beta_3$, the total effect in 2006. The difference in discontinuities at age 65 between the pre- and post-Part D era will require the following identifying assumption:

$$\lim_{R_{it}^{+} \to R_{it}^{-}} E[\eta_i + v_{it} | R_{it}^{+}, t \geq 2006] - E[\eta_i + v_{it} | R_{it}^{-}, t \geq 2006]$$

$$= \lim_{R_{it}^{+} \to R_{it}^{-}} E[\eta_i + v_{it} | R_{it}^{+}, t < 2006] - E[\eta_i + v_{it} | R_{it}^{-}, t < 2006]$$

This is a weaker identifying condition than the original RDD assumption or traditional difference-in-differences (DID) assumptions. We only require that the discontinuity in the unobservables in the period before the policy change is the same as the discontinuity in the unobservables after the policy change. This allows for discontinuities in the unobservables at $\bar{R}$ in each period (and hence is weaker than the standard RD assumptions) and does not require that the unobservables are conditionally independent of the interaction between the period indicator and the treatment indicator (and

hence is weaker than the standard DID assumptions). This identifying assumption would fail if, for example, another policy change coincided with the observable change in discontinuities so that the effect of crossing the threshold differs between the periods for multiple reasons, or the composition of individuals who cross the threshold changed between the two periods (i.e. a cohort effect coincidentally lined up with the implementation of the new policy regime). It is clear that checking for violations of this condition is more intricate precisely because it is only violated in more elaborate scenarios. I include robustness checks for whether the change in behavior can be attributed to one of (i) changes in the location of the retirement spike between the two eras or (ii) changes in patterns of selection into Medicare Advantage plans (which also changed as a result of the Medicare Modernization Act of 2003 that created Part D (McGuire et al., 2011)).

Despite these weaker identifying assumptions, the difference-in-discontinuities estimator may suffer from finite-sample problems that are common to the dynamic RDDs in the preceding subsections. These problems are inherent to two-stage-least-squares (2SLS) estimators of heterogeneous effects. In the difference-in-discontinuties design, the effects differ between the pre-2006 and post-2006 periods; in the dynamic RDD (see Appendix), the effects differ according to how long ago individuals became eligible for treatment. Since in the heterogeneous effects case there is more than one endogenous variable, we need more instruments so that there are at least as many instruments in the first stage as endogenous variables in the second stage. Since $1\{R_i \geq \bar{R}]$ will be used to instrument for insurance status pre-2006 (say) and $1[R_i \geq \bar{R}] \times 1[t = 2006]$ to instrument for $D_{it} \times 1[t = 2006]$, this leads to the problem of insufficient *independent* variation across instruments in the first stage, which leads to weak identification (Shea, 1997, Feir et al., 2016). In practice, this turns out to be a more serious problem for the dynamic RDDs than for the regressions investigating the effect of Part D. I address this and other problems of both finite-sample bias and bias in standard errors in the next subsection.

1.3.3 The Quasi-Difference-in-Differences Design

To examine the effect of Part D on the aggregate health of diabetics, I rely on a quasi-"difference-in-differences" design. This is a modification of the standard difference-in-differences design where

all units are exposed to treatment, but the extent of treatment varies exogenously across groups. In this setting, this corresponds to the larger effect of prescription drug coverage post-2006 on diabetic women's insulin usage. This will be seen to result from the larger proportion of diabetic women relative to men pre-65 who are uninsured - likely as a result of their lower participation in the labor market. The estimating equation is

$$Y_{it} = \beta_0 + \beta_1 1[Female = 1]_{it} + \beta_2 1[t \geq 2006]_{it} + \beta_3 1[Female = 1]_{it} \times 1[t \geq 2006]_{it} + \psi_i + \xi_{it}, \quad (1.3)$$

with β_3 as the parameter of interest. A key challenge for identification - having established that women respond differently to men to qualifying for Medicare coverage at age 65, and hence to changes in the composition of Medicare benefits - is to explain why differences in levels in say, heart disease between the genders can coexist with parallel trends in that same outcome (Kahn-Lang and Lang, 2019). I address these issues in more detail in Section 1.5.

1.4 Data

I use data from two principal sources: the Health and Retirement Study, a nationally representative longitudinal survey administered by the Institute for Social Research at the University of Michigan, as well as a cleaned version of a subset of the data called the "RAND HRS" dataset (Chien et al., 2013). Attention is restricted to 3 043 individuals diagnosed with diabetes from the 1998 wave of the HRS. In 1998, the individuals are drawn from four birth cohorts: the Oldest Old (born pre-1924), the Children of the Depression (born 1924-31), the original cohort from 1992 (born 1931-41) and the War Babies (born 1942-47), plus their co-habitants in the households in which they resided at the time of the survey. The HRS followed up respondents every two years, and so I also have data on surviving individuals from the 1998 wave in 2000, 2002, 2004, and 2006. The demographic characteristics of the sample are summarised in Table 1.2 (below).

Table 1.2 shows that the sample is primarily composed of either high school graduates or high

school dropouts. The sample has roughly two-thirds as many college graduates as the full HRS sample in 1998 (11.53% among diabetics versus 16.97% for the full sample). Strikingly, Table 1.3 reports that a majority of diabetics under 65 in the sample are not employed; this is in contrast to the nearly two-thirds of the full sample who report working for pay in 1998. This has adverse consequences for these diabetics' access to health insurance. Less than 1% of the sample report having some form of private insurance that isn't provided by an employer plan, which likely reflects the reluctance of insurers to accept enrollees with pre-existing conditions in the pre-ACA era. Baseline usage of insulin is in line with previous estimates for the total diabetic population in the United States, at 29.13% of under-65s and 25.99% of over-65s, a relative difference of 10.77%. (Compare Saaddine et al. (2002), who find that 30.9% of diabetics in the Third National Health and Nutrition Examination Survey (NHANES III) report using insulin to manage their condition). Those who receive Medicare before age 65 are equally split between individuals who report receiving Social Security Disability Insurance (SSDI) and those who do not. The latter are likely receiving Medicare for the treatment of End-Stage Renal Disease (ESRD), as this is the main alternative for accessing Medicare before age 65 for diabetics, but this is not asked in the HRS survey.

Table 1.2: Characteristics of the Sample at Baseline - Demographics, Health and Health Behaviors

Characteristic	Age < 65 ($N = 1308$)	Age ≥ 65 ($N = 1735$)
Age - Mean	57.99	74.40
BMI - Mean	30.73	27.86
Male (%)	45.34	45.82
High school graduate (%)	46.71	41.11
College graduate (%)	13.91	9.74
White (%)	65.62	76.89
Black (%)	28.10	18.96
Married (%)	68.46	54.56
Medical History (%)		
Current Smoker	19.88	8.18
Cancer	7.19	14.50
Heart Disease	25.61	39.32
Stroke	8.49	16.26
Using (%)		
Insulin	29.13	25.99
Medication	59.22	62.32
Diet	62.84	59.18
Vigorous Exercise	36.73	27.95

Notes: Drawn from the 3 043 self-reported diabetics in the 1998 Health and Retirement Study. All health conditions except for the "Current Smoker" indicator, which only applies to those who report smoking in 1998, are coded as "1" if a respondent reports that a doctor has ever diagnosed them with that condition, and "0" otherwise. The last four rows correspond to questions only asked of diabetics regarding their methods for managing their diabetes. Respondents who neither report their race as "White" or "Black" in the HRS are coded as "Other", and comprise the balance of the sample.

Table 1.3: Characteristics of the Sample at Baseline - Insurance and Employment Status

Coverage via:	Age < 65 ($N = 1308$)	Age ≥ 65 ($N = 1735$)
Medicaid (%):	13.0	16.5
Employer (%):	40.3	19.8
Spouse's Employer (%):	19.7	10.2
Union (%):	21.7	9.2
Medicare (SSDI) (%):	9.1	-
Medicare (not on SSDI) (%):	9.1	-
Private Insurance (%):	0.9	-
Not Covered (%):	21.5	2.1
Retiree Benefits? (%):	67.4	-
Working (%):	46.6	12.1
Considers Self Retired (%):	41.2	91.3

Notes: Drawn from the 3 043 self-reported diabetics in the 1998 Health and Retirement Study. I distinguish between individuals who report having access to Medicare and are on Social Security Disability Benefits (SSDI), and those who are not. The latter group almost certainly has access to Medicare via being in End Stage Renal Disease (ESRD) or kidney failure, which is not specifically recorded in the 1998 HRS but is one of the few routes to accessing Medicare before age 65 and a long-term consequence of diabetes.

Table 1.4: Incidents (%) of First Insulin Use vs. Insulin Cessation, Ages 65-66, 2000-2008

	2000	2002	2004	2006	2008
Ceased	12.50	17.07	11.90	1.89	3.45
Began	4.32	6.25	9.00	13.58	13.64

Notes: Each cell in the top row displays the percentage of respondents who reported using insulin at ages 63-64 two years prior who no longer report using insulin in the survey year in the header at ages 65-66. Each cell in the bottom row displays the percentage of respondents who reported not using insulin at ages 63-64 two years prior who report using insulin in the survey year in the header at ages 65-66. Author's own calculations from the Health and Retirement Study data, waves 4-9.

Table 1.4 shows that the percentage of diabetics who report not using insulin prior to receiving Medicare who begin reporting insulin usage once they qualify for Medicare coverage increases year-on-year between 2000 and 2006. In 2006, the percentage of newly qualified Medicare beneficiaries

who report using insulin at ages 63-64 and no longer report using insulin once they are 65 or older falls precipitiously, from an average of 13.82% across 2000-2004 to 1.89%. This descriptive evidence appears to make a *prima facie* case that Part D's introduction of more generous coverage for insulin significantly offset the crowding out of insulin usage by Medicare's coverage for treatment.

Moreover, when examining the percentages using insulin across the genders, a pattern emerges for the pooled sample in 1998-2004 that suggests that the crowding out effect of insurance is of first-order importance. Among non-Medicaid-dependent non-smokers aged 60-64 before 2006, female diabetics are between one and a half and twice as likely to be uninsured as male diabetics (see Section 1.2). At the same time, 25% of female diabetics with the same qualifiers report using insulin before age 65, compared with only 21% of men, a relative difference of nearly 25%. After age 65, these figures are 23.2% for women and 22.6% for men, a relative difference of approximately 2.6%. The relative difference in insulin usage between men and women is therefore smaller by nearly a factor of ten after age 65, when nearly 100% of individuals have access to insurance for treatment, relative to pre-age 65, when there are more uninsured female diabetics than male diabetics.

The following section gives the results of applying the empirical strategy explained in Section 1.3 to this data.

1.5 Results

In this section, I first present the results from the panel RDDs on the crowding out of prevention by insurance for treatment. I then extend the analysis to examine the offsetting effect of Medicare Part D, which made prescription drug coverage available at age 65, including more generous coverage for insulin. Lastly, I examine whether Medicare Part D affected aggregate health outcomes. The conclusions in this section are not altered in the dynamic regression-discontinuity designs, for reasons discussed alongside their presentation in the Appendix.

Throughout my analysis I exclude diabetic Medicaid recipients as they are eligible for full subsidies for their insulin by 1997 (with some mild restrictions in some states on the purchase

of auxiliary medical equipment such as blood glucose strips). I also exclude smokers, who have a weaker response to treatment with insulin and altered metabolism compared to the majority of diabetics (Eliasson, 2003). The regression-discontinuity and difference-in-discontinuity results are compared with their first-differenced counterparts to obtain lower and upper bounds for the effect of Medicare eligibility on health behaviors.

In the RDD results, I exclude Medicare Advantage and Medigap recipients who are over age 65 to deal with the "multiple treatments" problem (Caetano et al., 2017, Card et al., 2008). Not only do the previously uninsured gain insurance for the first time at age 65, the continuously insured also gain access to more generous coverage than before. The effect at the cutoff will therefore be a combination of these separate effects. Male diabetics do not exhibit the large changes in insulin usage that diabetic women do, and are also significantly less likely to be uninsured prior to age 65 when compared with female diabetics. This provides some evidence that the main results are due to changes at the extensive margin from being uninsured to being insured rather than at the intensive margin from less generous to more generous insurance. I also estimate regressions restricted to individuals who receive health insurance via their own or their spouse's employer, reported in the Appendix - the null results recorded there provide further evidence that changes in the composition and/or generosity of employer-provided coverage at age 65 is not the main mechanism behind the results. In addition, few of those in the sample who are uninsured before age 65 purchase supplemental insurance after age 65. Of those Medicaid-ineligible diabetics in the period 1998-2004 in the HRS who report buying supplemental insurance in the first two years of their Medicare eligibility, only 10.5% report having no source of health insurance two years prior. (By contrast, the first-differences regressions in Section 1.7 exploit variation within individuals rather than across individuals, and so can track what happens to individuals before and after qualifying for Medicare, or enrolling in supplemental Medicare insurance (Medigap) or a Medicare Advantage plan. Instead of dropping these observations for the first-differences regressions, I therefore control for the first difference of enrolment in Medigap or Medicare Advantage).

In the Part D results, matters are further complicated by the fact that aftter 2006 Medicare Advantage plans were required by the Medicare Modernization Act (MMA) to offer prescription

drug coverage that was at least equivalent to what could be obtained in a private Part D plan (McGuire et al., 2011). As a result, diabetics already enrolled in Medicare Advantage plans before age 65 may lead to underestimates of the extent to which the effect of crossing the age 65 threshold changes post-2006. In consequence, I exclude Medicare Advantage enrollees at all ages for the regressions that use the 1998-2008 sample. The resulting loss of observations is compensated by the greater sample size due to the addition of two waves of the HRS data.

To summarize: Medicaid recipients and smokers are present in none of the samples used for estimation. Supplemental insurance and Medicare Advantage enrollees are excluded if over 65 for the regression discontinuity design estimates in 1998-2004, and at all ages for the regression discontinuity design estimates in for 1998-2008 that determine the effect of Medicare Part D. In the first-differences regressions in Section 1.7, the first difference of indicators for enrolment in Medicare supplemental insurance (Medigap) or Medicare Advantage are included in one of the specifications as part of a set of control variables.

1.5.1 Crowding Out of Prevention by Insurance for Treatment: 1998-2004

In this subsection I document that the strongest evidence for *ex ante* moral hazard in insulin usage comes from female diabetics. The likely source of this difference is the much larger proportion of diabetic women who report having no source of health insurance relative to men prior to age 65, a difference of ten percentage points.

I now turn to evidence from panel RDDs, pooling together the years 1998-2004 (avoiding the policy regime change of 2006). This involves estimating the empirical counterpart of Equation 1.1 by two-stage least squares, which, with a Uniform kernel and local linear regression (used throughout this chapter), has the second-stage equation

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \gamma_0 (R_{it} - \bar{R}) + \gamma_1 (R_{it} - \bar{R}) \times 1[R_{it} \geq \bar{R}] + \delta X_{it} + \zeta t + \eta_i + v_{it} \text{ for } \left| \frac{R_{it} - \bar{R}}{h} \right| < 1, \quad (1.4)$$

where the subscript indicates an observation is for individual i in period t. I distinguish between

the time-invariant unobserved "fixed effect" η_i and the time-varying idiosyncratic error v_{it}. X_{it} is a vector of covariates and h denotes the bandwidth, chosen to minimize the mean-squared-error (MSE) criterion of Calonico et al. (2014a). I cluster standard errors at the individual level to account for the joint presence of persistence in treatment status and the error term (Bertrand et al., 2004).

Table 1.5: Unrestricted Panel RDDs, 1998-2004, Diabetic Women: Labor Market Outcomes

(1) Employed	(2) Retired	(3) Partly Retired	(4) Hours	(5) Earnings	(6) Social Security
0.03	-0.03	0.09	-0.31	731.71	-0.00
(0.76)	(-0.72)	(1.01)	(-0.13)	(0.87)	(-0.11)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications use local linear regression with age in months as the running variable and a Uniform kernel. Bandwidth used is 83.33, selected by the MSE criterion of Calonico et al. (2014a). Individuals enrolled in Medicaid at any age, or enrolled in supplemental insurance (Medigap) or a Medicare HMO (Medicare Advantage) after age 65, are excluded.

I now turn to testing for other changes at age 65 that might explain any differences in behavior other than Medicare eligibility. One potential threat to internal validity would be the coincidence of retirement at 65 with Medicare eligibility. There is some controversy over whether the spike in retirement status at 65 has disappeared in the United States (Card et al., 2008, Von Wachter, 2002, Johnson et al., 2013). This is the position of recent papers that use the Medicare eligibility age in a regression-discontinuity design such as Card et al. (2008, 2009). Whether or not this prevails for the U.S. population in general, a striking number of the diabetics in the sample under 65 are not employed (see Table 1.2). This likely contributes to my findings that there is little evidence of a spike in retirement at age 65 for diabetics (see Table 1.5). Another reason for this absence may be that the employed among this group are likely to retire later than at age 65. This is due to their condition giving them especially strong incentives to retain any benefits their employer-provided coverage may offer that are not provided on traditional Medicare. Several studies have found that

Table 1.6: Panel RDDs, 1998-2004, Diabetic Women: Other Health Behaviors and Outcomes

(1)	(2)	(3)	(4)
Any Hospital Stay	Nights in Hospital	Any Doctor Visit	No. Doctor Visits
-0.04	3.09	0.14*	2.34
(-0.21)	(0.66)	(2.43)	(0.29)
Kidney Problems	Poor Health	Diabetes Diagnosis	BMI
0.25	-0.09	-0.02	3.80
(1.89)	(-0.46)	(-0.37)	(1.67)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications use local linear regression with age in months as the running variable and a Uniform kernel. Regression discontinuity design is sharp. Bandwidth used is 83.33.

retaining health insurance benefits provided by employer-based plans are a significant influence on the timing of retirement in the United States, such as French and Jones (2011), Blau and Gilleskie (2008), Blau and Gilleskie (2006) and Rust and Phelan (1997). In 1997, around 32 percent of private-sector employers offered their employees retiree coverage (Buchmueller et al., 2006).[4] In the sample, a higher fraction of employed diabetics have retiree benefits than the general working population, but that nonetheless leaves a significant fraction do not, at around 32.6% (see Table 1.2). This suggests that the 1998 sample is relatively polarized between those not in work prior to age 65 and those who both work and have retiree coverage. These two forces will lead to a distribution that is polarized between individuals who retire at the Social Security claiming age of 62 on the one hand, and those who delay retirement as much as possible on the other, which will lead to a much less pronounced spike in retirement status at 65.

I also test for discontinuities in eight other behaviors and outcomes at age 65 (Table 1.6). The only statistically significant change is a discontinuous increase in the probability of reporting having had a doctor's appointment in the past two years, in line with the results found in Card et al. (2008)

[4]For a more recent treatment of the effects of retiree coverage on precautionary behaviour before the Affordable Care Act, see Clark and Mitchell (2014).

and Dave and Kaestner (2009).[5] Most significantly, there is no discontinuous change in the diagnosis of diabetes at age 65, which would otherwise potentially explain any negative effect as an increase in newly diagnosed diabetics who were not in need of insulin to manage their condition.

Table 1.7: Unrestricted Panel RDDs, 1998-2004, Diabetic Women: Insulin and Oral Medication Usage

	(1)	(2)	(3)	(4)
	Mean 60-64			
Insulin				
	0.26	-0.33*	-0.34*	-0.35*
		(-2.28)	(-2.42)	(-2.46)
Oral Medication				
	0.66	0.15	0.18	0.19
		(0.87)	(1.13)	(1.15)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Column (2) reports estimated discontinuities from specifications without covariates apart from the age in months and age in months interacted with the treatment indicator; Column (3) reports results including time dummies; Column (4) reports results including both time dummies and health status, marital status, work status and education fixed effects. All specifications use local linear regression with age in months as the running variable and a Uniform kernel. Bandwidth used is 83.33, selected by the MSE criterion of Calonico et al. (2014a). Individuals enrolled in Medicaid at any age, or enrolled in supplemental insurance (Medigap) or a Medicare HMO (Medicare Advantage) after age 65, are excluded.

The results for the unrestricted panel are summarised in Table 1.7.[6] The diet and exercise variables are either not available for the entire period 1998-2004 or, in the case of the exercise

[5] It is unlikely that these doctors' appointments can explain the reductions in insulin usage in this chapter. Once a patient is already using insulin, the "therapy of last resort", it would be contrary to the official guidelines for physicians (Nathan et al., 2009) to recommend that they discontinue using insulin to manage their condition. This would be more consistent with both the argument in Card et al. (2008) that a small reduction in smoking at age 65 may be attributed to this greater frequency of doctor's appointments and the medical literature on adherence to insulin, where the goal of most health providers is to encourage adherence to insulin once prescribed (cf. Weinger and Beverly (2010)).

[6] Separate results restricted to the original 1998 cohort can be found in the Appendix.

variable, are changed so as to make comparisons across time difficult. I nonetheless find little evidence of substitution towards these alternative investments in health as a result of crossing the age 65 threshold (see Appendix). The null hypothesis that there is no substitution towards oral medication to manage diabetes at age 65 is not rejected. By contrast, there is consistently strong evidence in favor of a decrease in insulin usage at age 65 across the years 1998-2004. Given the first stage estimate of 0.24 for insurance status, the smallest coefficient of -0.33 implies a $-0.33 \times 0.24 = -7.92$ percentage point decrease in insulin usage among female diabetics at age 65 from a baseline of 26% at ages 60-64, a relative reduction of 30.5%.

1.5.2 Mechanisms: Crowding Out via *Ex Ante* Moral Hazard

There are two mechanisms which could produce a discontinuous change in behavior in the month of Medicare eligibility. One of these, which I do not model explicitly in this chapter, is a precautionary motive. Hospitalizations do not increase discontinuously, in line with the findings of Card et al. (2009) (see Table 1.6), so the incentive to use insulin is unlikely to come from avoiding emergency treatment. It is more likely that the change in incentives for insulin usage comes from reduced uncertainty regarding the threat of a costly medical incident that, while difficult to defer, need not require immediate medical attention. This is due to the unpredictability of the timing of complications due to poorly managed diabetes. One of the consequences of episodes of abnormally high blood sugar levels (hyperglycemia) is direct damage to the cardiovascular system (Barrett-Connor et al., 2004), which raises the likelihood of an adverse cardiovascular event such as a heart attack. This is supported by the results in Section 1.5.5, which finds a forgone increase of 4.6 percentage points in the rate of heart disease among diabetic women due to the provision of prescription drug coverage under Medicare Part D in 2006.

Insulin usage may play the same role as precautionary behavior in this context. Studies of precautionary saving find similar results for exogenous variation across individuals in the replacement rate provided by unemployment insurance (Engen and Gruber, 2001) and access to a consumption floor via social insurance (Hubbard et al., 1995). A discontinuous change in the month of Medicare eligibility, similarly, is consistent with an exogenous reduction in uncertainty regarding the risk of

being liable for medical expenses for complications of diabetes at that time.

As individuals approach the threshold of Medicare eligibility, the risk of liability for large medical expenses decreases. As a result, if this is the mechanism behind the discontinuous decrease at the threshold, there could in principle also be a continuous downward trend before age 65 (cf. De Preux (2011) on "anticipatory moral hazard"). However, a consistent estimator of this age effect is difficult to obtain in practice; in the regression-discontinuity design, the running variable is not exogenous and we cannot obtain consistent estimators of the age profile of the outcome of interest (otherwise there would be no need for the discontinuity in the first place). In addition, using panel data presents the problem of distinguishing age effects from period-specific and cohort-specific effects. I do not attempt to obtain a consistent estimator of the extent of anticipatory moral hazard in this chapter, since the existence of per-period reductions leading up to age 65 is not mutually exclusive with the existence of a discontinuous change in individuals' incentives at age 65.

One test for whether there is a discontinuous change in the risk of medical expenses is to test for discontinuous reductions in the dispersion of medical expenses or the mean medical expenditures at age 65. This would then track the precautionary motive that diabetics have to use insulin, and would shed further light on the mechanism at work. It may be that the sequential reduction of uncertainty month-by-month approaching the date of Medicare eligibility is too small to be distinguished from noise, but that the discontinuous reduction in uncertainty in the month of Medicare eligibility is large enough to produce statistically significant changes in behavior. Barcellos and Jacobson (2015) find both a discontinuous 53 percent reduction in the 95th percentile of medical expenses at age 65 in the Medical Expenditure Panel Survey and similar declines in medical expenditures risk in the HRS, the same data set used in this study. In the Appendix, I examine similar evidence for changes in two measures of financial and medical expenditure risk at age 65, but obtain less conclusive results, likely because of the smaller sample size in this study.

It is also possible to explain the change in behavior as intertemporal substitution in a model without uncertainty. In the model in Section 1.6, I show that if prevention lowers the marginal utility of medical expenses in the future (i.e. it decreases future demand for healthcare, and is a substitute for treatment *ex post*), then an anticipated reduction in the price of treatment in the

next period will lower the optimal amount of prevention in that period. The responses found in this chapter can therefore arise in an environment of pure certainty as well (as there is no uncertainty in the aforementioned model). It is sufficient that self-insurance *ex ante* and insurance against financial losses *ex post* are substitutes. This mechanism also has the advantage of allowing me to reconcile the results in this chapter with those in other studies. I leave further discussion of the model to Section 1.6.

In Chapter 3 of this dissertation, I find that individuals fail to revise their reported future medical spending in the months leading up to their Medicare eligibility. The only effect on reported future medical spending found in Chapter 3 is a modest decrease (10 percentage points) in the reported probability of out-of-pocket medical expenses exceeding $1500. This may be consistent with individuals maintaining investments in health only up to the point where there is a large drop in the uncertainty of their medical expenses (cf. Barcellos and Jacobson (2015)). Consumption smoothing may then imply no mean change in total medical spending before and after age 65, while the drop in the probability of spending more than $1500 out of pocket corresponds entirely to the lower variance of medical expenditures when covered by Medicare. The results in Chapter 3 are also more consistent with large changes in insulin usage at the threshold than with a gradual decrease in insulin usage leading up to age 65 as uncertainty is resolved month by month. By contrast, if individuals reacted smoothly to the sequential resolution of uncertainty, there would be a lower probability of detecting a discontinuous change in insulin usage at the cutoff as much of the adjustment would have taken place already by the time they qualify for Medicare.

1.5.3 The Effect of Prescription Drug Coverage on Ex Ante Moral Hazard and Aggregate Outcomes

I now turn to analysing the effect of introducing generous subsidies for the purchase of insulin under Medicare Part D in 2006. The key result in this subsection is that Medicare Part D appears to have increased the demand for insulin among diabetic women by enough to more than offset the *ex ante* moral hazard effect of traditional fee-for-service Medicare. These results perform four separate functions in this chapter. First, they buttress the initial results on the negative impact

of coverage without insulin subsidies on usage by showing that this effect is reversed just when subsidies are introduced. Second, they suggest a method for combating *ex ante* moral hazard - lower the expected price of health-preserving behaviors in tandem with lowering the expected price of health care. Third, strong changes in oral medication usage are not observed with the onset of Part D, likely because of their significantly lower cost and better coverage options before 2006 (see Section 1.2) compared to those available for insulin. This allows me to attribute the prevention of an increase in heart disease among diabetic women found in Section 1.5.5 (below) to the insulin subsidies available on Medicare Part D specifically, rather than its broader coverage for other medications.

In the "difference-in-discontinuities" regressions, I estimate the empirical counterpart of Equation 1.2,

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \gamma_0(R_{it} - \bar{R}) + \gamma_1(R_{it} - \bar{R}) \times 1[R_{it} \geq \bar{R}] + \beta_2 1[t \geq 2006] + \beta_3 D_{it} \times 1[t \geq 2006] + \delta X_{it} + \zeta t + \eta_i + v_{it} \tag{1.5}$$

In Section 1.3, I discuss the weaker identifying assumptions necessary to identify the effects of interest than in the preceding subsection. Although two-stage least squares regressions that use highly correlated instruments are more susceptible to weak identification (Shea, 1997), in practice I find strong evidence against the null hypothesis that the set of instruments is weak.

There are three additional empirical challenges in this subsection. The first results from the fact that the passage of the Medicare Modernization Act (MMA) was in 2003, so there is at least one survey year (2004) in which individuals' behavior may have already been affected due to their anticipation of the availability of prescription drug coverage two years thereafter (Alpert, 2016). I discuss how much of the results can be accounted for by this mechanism in Section 1.5.4. Second, the MMA also changed the regulations governing private plans on Medicare Advantage (Part C), and there was a corresponding rapid increase in the take-up of these plans relative to their decline in the period 1997-2003 (McGuire et al., 2011). As per the discussion in Section 3 (above), I include robustness checks for changes in enrolment in Medicare HMOs (Medicare Advantage), retirement

behavior, and frequency of diagnosis, as well as time dummies to absorb anticipatory behavior by those not yet eligible for Medicare (Table 1.8). Third, as in the case of the dynamic equations (see Appendix), heterogeneous effects of two-stage least squares require at least as many sources of exogenous variation as endogenous variables. Although the identification conditions are weaker than for two-stage least squares (see above), the finite-sample issues are the same. Since the variation in $1[R_i \geq \bar{R}]$ is similar to that in $1[R_i \geq \bar{R}] \times 1[t = 2006]$, Shea's R^2 may be low, reflecting little independent variation in the first-stage (Shea, 1997). Since this variation is monotonically increasing in the sample size, the optimal bandwidth for the purposes of maximizing Shea's R^2 is $h = \infty$. It turns out that in this case there is little evidence of weak identification even at the MSE-optimal bandwidth of $h = 66.23$, as the Cragg-Donald statistics used to test for the presence of weak instrument sets all exceed conventional critical thresholds used to reject the null hypothesis that the set of instruments is weak (Stock and Yogo, 2005).[7]

It appears that there are mild differences in the effect of reaching age 65 on employment outcomes and diagnosis of diabetes between the period 2006-08 and 1998-2004 (Table 1.8). The latter is a reasonable response to gaining health insurance at age 65. As pointed out by Kenkel (2000), not all forms of prevention are the same: investments in health, as in the bulk of this chapter, are substitutes for having coverage *ex post*. By contrast, screenings are complementary to having coverage, since knowledge of one's condition is more useful if one can pay to treat it once it is discovered. Providing more generous coverage for prescription drugs will therefore provide an extra incentive to screen for conditions such as diabetes, due to individuals' increased ability to pay for the maintenance of one's health on discovering latent diabetes.

Of the changes in retirement and Social Security claiming behavior, none are particular to women and not to men (the increased rate of retirement post-2006 is among men only, and the rise in Social Security claims occurs across both sexes). The change in diabetes diagnosis is close to identical in both sexes. It seems unlikely - with the exception of increased retirement among men post-2006 - that these changes can explain disparities between the genders in their responses to qualifying for Medicare and the change in the Medicare program after 2006. There are two further

[7] Further discussion of weak identification in regression-discontinuity designs can be found in Feir et al. (2016).

Table 1.8: Differences in Discontinuities in Medicare Advantage Enrolment and Employment Measures, Pre- and Post-2006

	(1)	(2)	(3)	(4)
	Med. Adv.	Employed	Retired	Partly Ret.
Women				
	0.08	-0.03	-0.00	-0.04
	(1.37)	(-1.27)	(-0.00)	(-1.16)
Men				
	0.10	-0.03	0.06*	-0.06
	(1.56)	(-1.12)	(2.24)	(-1.79)
	Hours	Earnings	Soc. Sec.	Diagnosis
Women				
	-0.11	-574.21	0.05***	0.05**
	(-0.11)	(-0.63)	(4.54)	(2.69)
Men				
	-0.31	-2252.86	0.05***	0.05**
	(-0.27)	(-1.24)	(4.60)	(2.69)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Estimates are from the interaction term between the indicator for the year 2006 and the treatment indicator for a sharp regression discontinuity design. Standard errors are clustered at the individual level. "Medicare Advantage" results are from regressions where the dependent variable is equal to 1 if survey respondents answer "Yes" to the question "Do you receive your Medicare through an HMO?" and 0 otherwise. Earnings are measured in constant 1998 dollars. All results are from local linear regressions using the Uniform kernel. Bandwidth used is 66.23, selected by the MSE criterion of Calonico et al. (2014a).

reasons to suspect that these differences in discontinuities do not explain changes in insulin usage instead of differences in insurance status and the composition of available insurance packages. First, as in Card et al. (2008), the change in retirement status amongst men is too small to explain a dramatic change in the behavior of women either relative to men or in absolute terms. Second, to the extent that increased frequency of diagnosis of diabetes introduces bias into the estimator of the coefficient on $\widehat{D}_{it} \times 1[t \geq 2006]$, $\widehat{\beta}_3$, this bias is likely to be towards zero, since newly diagnosed

diabetics are not typically prescribed insulin as it is a therapy of last resort (see Section 1.2). As a result, the effect of Part D will be *under*stated if the difference-in-discontinuities in diagnosis of diabetes plays a large role in producing the results to follow.

Table 1.9: Difference-in-Discontinuities, Diabetic Women: Effect of Part D on Insulin and Oral Medication Usage

	(1)	(2)	(3)
Insulin			
$\widehat{\beta}_1$	-0.34*	-0.35**	-0.35**
	(-2.50)	(-2.62)	(-2.62)
$\widehat{\beta}_3$	0.68*	0.68*	0.60*
	(2.41)	(2.42)	(2.28)
Cragg-Donald Stat.	33.70	35.22	40.77
Oral Medication			
$\widehat{\beta}_1$	0.08	0.10	0.11
	(0.48)	(0.65)	(0.72)
$\widehat{\beta}_3$	0.03	0.03	0.05
	(0.11)	(0.12)	(0.17)
Cragg-Donald Stat.	33.66	35.27	40.63

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Column (1) reports estimated discontinuities from specifications without covariates apart from the age in months and age in months interacted with the treatment indicator; Column (2) reports results including time dummies; Column (3) reports results including both time dummies and work status, marital status, education, and health status indicators. All results are from local linear regressions using the Uniform kernel. Bandwidth used is 66.23, selected by the MSE criterion of Calonico et al. (2014a). Individuals enrolled in Medicaid, supplemental insurance (Medigap), or a Medicare HMO (Medicare Advantage) at any age are excluded.

Table 1.9 shows the effect of Part D on the change in the effect of qualifying for Medicare. I use a bandwidth of 66.23, selected by the MSE criterion derived in Calonico et al. (2014a) as before. The effect of qualifying for Medicare in 2006-08 is found to have a significantly more positive net impact on insulin usage than in the pre-2006 part of the sample, as examination of Table 1.3 in Section 1.5 would suggest. A Wald test cannot reject either the hypothesis that $\beta_1 + \beta_3 = 0$ or that $\beta_3 = -2\beta_1$. In sum, it appears that Part D increased the demand for insulin to an extent that completely offset the *ex ante* moral hazard effect of coverage for treatment on Medicare Parts A and B. The precise size of this effect is difficult to ascertain, but the results leave room for the possibility that it not only completely offset the negative "crowding out" of insulin usage at age 65 but also led to an equally large increase in uptake at 65 post-2006.

1.5.4 Mechanisms: Prescription Drug Coverage Under Part D Post-2006

In this subsection I discuss potential alternative mechanisms that can account for the net positive effect of qualifying for Medicare coverage on insulin usage post-2006.

There are two potential challenges to interpreting these results against which I can find no direct evidence in the data. The first is that a new long-acting (requiring only once-daily usage) insulin, insulin detemir (Levemir), was approved by the United States Food and Drug Administration in 2005, the year before Part D was implemented. At least one other long-acting insulin compound, insulin glargine (Lantus), had been available since April 2000. Since these types of insulin were not differentially available to over-65s, accounted for a small share of the market for insulin over the sample period, and require large implicit non-monetary costs of insulin usage to explain the results here, it seems unlikely that their introduction is responsible for the results in this subsection. Another reason to be sceptical that this can explain a large share of the difference in the treatment effect in 2006 is that the more significant innovation of long-acting insulin had taken place three years before the passage of the Medicare Modernization Act and six years before the rollout of Medicare Part D. The approval of insulin detemir was a significantly smaller contribution to the therapeutic options available to diabetics relative to the invention of long-acting insulin, which was already available at the time.

The second is that take-up of Part D was generally slow and made difficult by a chaotic sign-up process. There are three potential explanations for the results in this section even with this observation. First, the subset of individuals who did sign up for Part D were precisely those with a high enough demand for coverage for insulin to outweigh the moral hazard effect. Second, there is evidence that the effect of Part D went beyond the direct effect on prices faced by enrollees and had spillover effects that lowered the prices faced by other Medicare beneficiaries (Duggan and Scott Morton, 2010). Third, the 2006 wave of the Health and Retirement Study was collected between March 2006 and February 2007; while the beginning of the survey period had some overlap with the period during which enrolment in Part D had been lower than anticipated, by July 2006 nearly 22.5 million senior citizens had enrolled in Part D (Cubanski and Neuman, 2007). Hence both the 2006 and 2008 waves of the Health and Retirement Study are likely to have been gathered after significant problems with the rollout of Part D had been resolved.

One explanation for the post-2006 results is intertemporal substitution, similar to the explanation in this chapter for the large effect sizes pre-2006 (cf. Section 1.6). Alpert (2016) studies the effect of Part D's prescription drug coverage on the demand for non-essential medications and finds that those close to the Medicare eligibility age and those who qualify for Medicare coverage before 2006 strategically delay purchases of prescription drugs until they become cheaper once subsidies under Part D are rolled out. This accords with statistically significant negative coefficients on the Post-2006 indicator variable in the RDD results (though its counterpart is not significant in the first-differences regressions). In principle, this could mean that there is no net effect adherence to insulin therapy; instead, there could be merely a redistribution of the timing of initiating insulin therapy so that the negative effect found in 1998-2004 is not offset at all by the subsidies for insulin available after 2006 on Medicare Part D. There are three reasons to believe that the lifetime increase in insulin usage exceeds the measured intertemporal substitution effect. First, individuals' lifetime income was increased by the passage of the Medicare Modernization Act which created Part D, which would increase the lifetime demand for insulin even absent any price effects. Not only were there no tax increases implemented to pay for Medicare Part D (that could have in principle neutralized this effect), the Tax Increase Prevention and Reconciliation Act of 2005 extended the

horizon to which the tax cuts of 2001 and 2003 applied, effectively until the end of the working lives of those near Medicare eligibility. Second, as mentioned when discussing the problems with the rollout of Part D (above), Part D lowered the prices of covered prescription drugs after its implementation more generally (Duggan and Scott Morton, 2010). Third, the net effect is likely to be understated due to the increased frequency of diagnoses of diabetes upon qualifying for Medicare post-2006 (see discussion in previous subsection).

In sum, it seems extremely unlikely that Part D did not increase lifetime insulin usage among diabetics. In the next section, I examine the implications of Part D encouraging insulin usage for aggregate health outcomes and health care costs.

1.5.5 Aggregate Effects of Prescription Drug Coverage: 1998-2014

Despite the large changes in behavior documented above that ought to affect blood sugar levels and fluctuations, which are known to damage both the large and small blood vessels, I can only discern evidence in the HRS data for a reduction among diabetic women in the most common complication of diabetes, which is heart disease. In this section I document a relative decrease of 4.6% in the trend in heart disease rates among female diabetics over 65 relative to their male counterparts in the post-Part D era. This also accords with the results of the previous section, where the largest behavioral effects are observed for women. I then calculate a conservative estimate of forgone health costs based on this decrease in heart disease, as well as a larger estimate based on a previous study of cost containment attributable to reductions in blood sugar levels among diabetics.

Accordingly, the main difference that can be attributed to higher take-up of insulin following Part D is an improvement in the rate of heart disease: though diabetic men saw their rate of heart disease rise by 4.6% in 2006-2014, the proportion of diabetic women who had contracted heart conditions remained constant. In this subsection I graph the year-on-year deviations from the 1996 average of heart disease for diabetics over and under 65 separately for each gender, and report a difference-in-difference specification to quantify the extent of the differences after 2006. Throughout this subsection, I exclude Medicaid recipients for the same reasons as in the preceding subsections.[8]

[8] In the Appendix, I present evidence that there are no significant differences in the trends of diagnosis of diabetes

Figure 1.2 shows the differences in rates of cardiovascular disease between men and women over and under the age of 65 relative to the year 1996. We need evidence for two assumptions to attribute a given change in disease trends to Part D. The first is that there is no corresponding change in trends among under-65s who are not affected by changes to Medicare. The second is that men and women's trends before 2006 moved in parallel, so that their trends would have been parallel in the counterfactual where Medicare Part D was not introduced. The bottom panels of Figure 1.2 show flat trends for diabetic under-65s, supporting the first assumption. The second assumption is supported by the flat trends in cardiovascular disease for both genders among over-65s prior to 2006, when the rates for men start increasing relative to 1996 at an increasing rate. Table 1.10 estimates the difference-in-differences between men and women before and after 2006, and finds a statistically significant forgone increase in heart disease of 4.6% among diabetic women. The estimating equation is Equation 1.3, viz.:

$$Y_{it} = \beta_0 + \beta_1 1[Female = 1]_{it} + \beta_2 1[t \geq 2006]_{it} + \beta_3 1[Female = 1]_{it} \times 1[t \geq 2006]_{it} + \psi_i + \xi_{it}, \quad (1.6)$$

The main competing explanation for these changes is the passage of the Affordable Care Act (ACA) in 2010, which mandated changes in the United States' public provision of health insurance over the period 2010-2014. There are three reasons for scepticism that this can explain the patterns observed in Figure 1.2 and Table 1.10. The first is that we should observe similar differences among under-65s, to whom the Affordable Care Act - unlike Medicare reform - applied to the same extent. The second is that the main expansions of insurance under the Affordable Care Act were expansions of the Medicaid program, whose recipients are excluded from the analyses in this and the preceding sections. The third is that the timeline of the changes implemented by the Affordable Care Act cannot cannot explain either the modest divergence in trends between men and women in 2006-2010, before the passage of the ACA, or the larger divergence by 2012. In 2012, the United States Supreme Court ruled that the ACA was constitutional in July and President Obama was re-elected

or take-up of Medicaid among diabetics that can be attributed to Part D. In the first case, the trend is positive and not significantly different post-2006; in the second case, all of the increase in Medicaid take-up among diabetic men can be attributed to the Affordable Care Act (ACA). I discuss elsewhere in this section why the ACA cannot explain the main difference in trends pre- and post-2006.

that November. Prior to those events, the implementation of the main portion of the ACA - the creation of health insurance exchanges backed by an individual mandate to purchase insurance - was in significant doubt due to the scale of political opposition to the Act. These would go on to be implemented in 2014, by which time the divergence in trends documented in this section had already arisen. The part of the ACA most relevant to diabetics pre-2012 is the provision of pre-existing condition plans (PCIPs), which had low overall enrolment. Frean et al. (2017) only find modest effects of the ACA on access to health insurance in 2012-3, with larger effects in 2014-5. Even their largest estimate of increase in enrolment in 2014-5 - 10.8 percentage points for single adults - is smaller than the percentage changes at age 65 estimated in this section in the percentage of diabetic women who are insured. These changes would also apply with equal force to under-65s, and so still cannot explain the differential trends between under- and over-65s found in this subsection.

Table 1.10: Heart Disease Rates Among Diabetics and Non-Diabetics by Gender, Pre- and Post-2006

	Diabetics		Non-Diabetics	
	(1)	(2)	(3)	(4)
	Over 65	Under 65	Over 65	Under 65
Female	-0.0539**	-0.0569**	-0.0909***	-0.0565***
	(-2.89)	(-3.03)	(-11.31)	(-11.25)
Post-2006	0.0474**	-0.0227	0.0479***	-0.00502
	(3.25)	(-1.32)	(6.41)	(-0.84)
Female × Post-2006	-0.0457*	0.00959	-0.0142	0.0229**
	(-2.26)	(0.43)	(-1.50)	(3.18)
Constant	0.416***	0.260***	0.305***	0.127***
	(31.19)	(18.23)	(48.08)	(30.62)

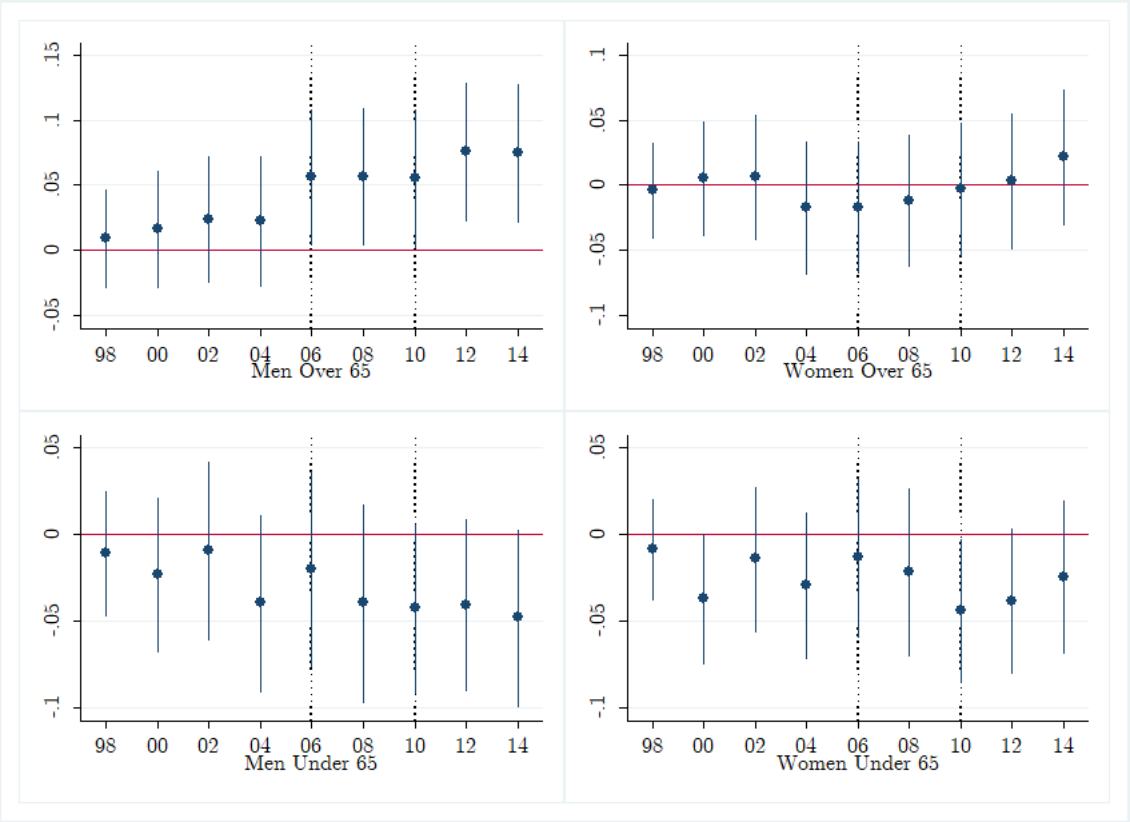
t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Dependent variable is an indicator for whether an individual responds "Yes" to the question "Has a doctor ever told you that you had a heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems?". Medicaid recipients are excluded.

The difference in mean rates of heart disease between the genders may still be cause for concern.

Figure 1.2: Trends in Heart Disease Relative to 1996 Among Diabetics in the HRS (Excl. Medicaid Recipients), 1998-2014



Notes: Plots are of coefficients from pooled OLS regressions of the outcome (proportion responding "Yes" to the question "Has a doctor ever told you that you had a heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems?") on time dummies for the years 1998-2014, with 1996 as the reference category. Since there is overlap among individuals across the different waves of the survey, standard errors for these regressions are clustered at the individual level. Dotted lines correspond to major health care reforms - Medicare Part D being implemented in 2006, and the Affordable Care Act being passed in 2010. Medicaid recipients are excluded.

Kahn-Lang and Lang (2019) argue that this calls counterfactuals in difference-in-differences and shift-share analyses into question, since one has to explain why the mechanism that produces (or allows for) the difference in levels does not also produce a difference in trends prior to the policy change. The counterfactual in this context is a similar upward trend for male and female diabetics' rates of heart disease over time. Some evidence that this would have occurred for diabetic women is given by columns 3 and 4 of Table 1.10, which provide evidence on trends in heart disease among non-diabetics. Over-65s' rates of heart disease are higher post-2006 for both genders without a differential trend, with a larger difference in levels between the sexes. This latter fact reflects previous findings that diabetes decreases women's advantage in vulnerability to heart disease (Juutilainen et al., 2004). Among non-diabetics under 65 (column 4 of Table 1.10) we can observe convergence in rates of heart disease post-2006 (which are unlikely to be attributable to Medicare, since they are typically ineligible). These observations together with the flat trends for both genders before 2006 (Figure 1.2) should strengthen our confidence that diabetic women would have shared in their male counterparts' increased propensity for heart disease after 2006 were it not for Part D.

A reduction in heart disease rates of this size is likely to have large cost savings. Using relatively conservative modelling assumptions, Barton et al. (2011) calculate that a 1% reduction in cardiovascular disease in the United Kingdom would result in cost savings to that country's National Health Service (NHS) of at least $48 million (£30 million at 2011 rates) per year in 2011 dollars. It is necessary to inflate the figures found in that study by a factor of 2 to obtain comparable numbers for the United States, since it has been found that the U. S. health care system pays roughly twice as much on average for comparable procedures to those in the rest of the OECD (Papanicolas et al., 2018). The population of American diabetics over age 65 is somewhat smaller, at around 20% of 46 million over-65s with diagnosed diabetes Centers for Disease Control and Prevention (2017), while the overall population of over-65s in the United States is similar in size to the overall population of the United Kingdom. Since Medicare tends to reimburse cardiovascular procedures at relatively high rates , preventing an increase in cardiovascular disease among diabetic women of 4.6% over 8 years is likely to have saved roughly $4.6 \times 0.2 \times 2 \times \$48$ million per year = $88.32 million per year in 2011 dollars over the same period.

This may yet underestimate of the net effect of lower blood sugar levels on health care costs. The HRS data may be underpowered to capture significant effects on other health outcomes, which have been found in other settings (i.e. Gilmer et al. (1997)). It is relatively easy to find larger estimates of forgone health care costs for even mild improvements in control of blood sugar levels, a broader criterion than examining changes in a specific adverse outcome such as heart disease. Using the lowest of Gilmer et al. (1997)'s estimates of $670 per person per annum in forgone health care costs of a 1 percentage point decrease in fasting blood sugar levels (at their data's average blood sugar levels) and the numbers of diabetics changing their behavior as a result of Part D, one can obtain forgone health care costs due to better control of blood sugar levels of up to $487 million per annum. Suppose we take the largest estimate of the effect of Part D of a net change of 15.8% (since I cannot reject the null hypothesis that the positive effect of Part D was twice as large as the negative "crowding out" effect). This gives 15.8 percent of the female diabetic population (23 million $\times 0.2 \times 15.8\%$ = approx. 726 800 people) each forgoing $670 per annum in health care costs from better glycaemic control, yielding approximately $487 million per annum in forgone health care expenditures.

Given the number of people involved, providing better coverage for insulin under Part D appears to reduce health care costs by between one-tenth and one-third as much as discouraging a similar number of people from smoking cigarettes. Choi et al. (2016) calculate that reducing the number of cigarette smokers by 2.5 million translates into a $4.6 billion reduction in health care costs. This is equivalent to a reduction of $1.3 billion in health care costs for a population of 726 800. Since tobacco control is considered one of the most cost-effective methods of improving population health, this suggests that encouraging insulin usage to the extent that prescription drug coverage under Part D did is among the more effective methods of holding down health care costs. A broad estimate of the savings indicates that the forgone costs may be up to 36% as large as smoking cessation in an equivalently large population. Given that smoking cessation is considered among the most cost-saving measures in public health, this places subsidising insulin in the first rank of policies aimed at containing health care costs.

The return on investment in insulin in the United States has likely decreased year-on-year as the

price of insulin in that country has increased dramatically between 2006 and the time of writing. The Medicare Modernization Act of 2003 prohibited the U. S. federal government from using its size as a purchaser of pharmaceuticals to bargain the prices of prescription medications downwards, as is done in most countries that provide prescription drug benefits. This component of the legislation may explain why average real annual expenditures per insulin user on insulin nearly tripled over the period 2006-2013, while average annual quantities of insulin demanded only increased by one-seventh (Hua et al., 2016). Given the cost savings calculated above, the returns to subsidizing insulin are likely to be higher in countries such as the Netherlands, where increases in the price of insulin have been less dramatic.

1.5.6 Summary of Findings

The empirical exercises above have three aims. First, they provide evidence that the negative crowding-out effect of insurance for treatment on prevention is strictly negative. Second, they estimate the extent to which this effect is counteracted if prevention is itself subsidized. Third, they provide evidence on the extent to which counteracting the crowding-out effect matters for health outcomes and spending on health care. The answers provided in this section were, first, that 30.5% of female insulin users, who undergo much larger changes in the proportion of uninsured individuals than males, stop using insulin when they qualify for Traditional Medicare coverage at age 65 in the pre-2006 sample. Second, this effect is either offset exactly or to the extent that there are equally large *positive* responses to qualifying for Medicare coverage post-2006, when that coverage included subsidies for insulin. Third, we can attribute a 4.6 percentage point forgone increase in heart disease among female diabetics and up to $487 million per annum in forgone health care costs to the change in behavior induced by the change in the Medicare program in 2006.

In the next section, I provide a simple theoretical model that allows me to interpret the results in this chapter and reconcile them with the rest of the literature on *ex ante* moral hazard in health behaviors.

1.6 Theoretical Framework: A Model of the Intertemporal Allocation of Prevention

In this section, I introduce a model where an agent chooses how to allocate her lifetime expenditures among consumption, prevention *ex ante*, and spending on non-preventive medical services. It turns out that all that is needed for the model to be able to rationalize the large effect sizes in this chapter is a sufficiently strong degree of substitutability between prevention and other medical spending. I conclude the section by drawing out the model's implications for which individuals are at the margin in this setting (and hence the distributional impact of the crowding-out effect) as well as its implications for the magnitudes of quantities that are not investigated in this chapter, particularly the income elasticity of demand for prevention.

Consider a two-period model where an agent decides between the allocation of her expenditures between consumption C_1 in period 1, medical services M_1 in period 1, and their period 2 counterparts, as well as continuous amounts of prevention ϕ_1, ϕ_2. Her lifetime utility that results from her choices is

$$U(C_1) + V(M_1, \phi_1) + \beta\{U(C_2) + V(M_2, \phi_2)\} \tag{1.7}$$

where the sub-utility functions satisfy the usual conditions. Denote, by (for example) V_ϕ the derivative of $V(.)$ with respect to ϕ in a given period (so that the sub-utility functions are the same for both periods). I will assume the derivatives of $V(M_t, \phi_t)$ with respect to ϕ_t have the following signs for $t = 1, 2$:

$$V_\phi > 0; \tag{1.8}$$

$$V_{\phi\phi} < 0; \tag{1.9}$$

$$V_{\phi M} < 0; \tag{1.10}$$

ϕ, prevention, has two roles: first, it is intrinsically valuable (i.e. for its role in producing health), and exhibits diminishing marginal utility as do the other arguments of the objective function. Second, the third derivative shows that the marginal utility of medical services is lower when ϕ is

higher. This captures the fact that the demand for medical services, and the share of the budget spent on medical services relative to consumption, are lower when the agent is in better health.[9] Setting the interest rate equal to zero for simplicity's sake, the lifetime budget constraint with initial assets A_1 is

$$C_2 + P^M M_2 + P^\phi \phi_2 = A_1 - C_1 - P^M M_1 - P^\phi \phi_1; \tag{1.11}$$

To obtain the elasticity of intertemporal substitution of prevention with respect to the price of treatment P^M, which will give agents' willingness to substitute prevention away from periods when the price of treatment is low and toward periods when the price of treatment is high, we have to hold the marginal utility of wealth (here denoted by μ) constant. This is because agents will move along their lifetime profile of treatment prices which, having been anticipated in advance, involves executing planned changes in prevention efforts conditional on the agent's lifetime resources. This means taking $\frac{\partial \mu}{\partial P^M} = 0$ when implicitly differentiating the first-order condition with respect to ϕ_2. Doing this yields the simple expression for the (Frisch) elasticity of intertemporal substitution

$$\epsilon^F_{\phi_2, P^M} \equiv \left(\frac{P^M}{\phi_2}\right)\left(\frac{\partial \phi_2}{\partial P^M}\right)\Bigg|_{\frac{\partial \mu}{\partial P^M}=0} = -\left(\frac{P^M}{\phi_2}\right)\left(\frac{V_{\phi M}\left(\frac{\partial M_2}{\partial P^M}\right)}{V_{\phi\phi}}\right), \tag{1.12}$$

which is positive, since:[10]

[9] Note that this is because I use "prevention" interchangeably with "health investments" in this chapter. Screenings for conditions can also be referred to as "prevention", but would have the opposite sign for the third derivative in this setting since screenings are complementary to health expenditures: it makes more sense to purchase medical services when one is aware of a condition than unaware, and being better able to purchase treatment makes the return to *information* regarding one's eventual health status higher.

[10] In contrast with the elasticity of intertemporal substitution of labor supply, which is larger than the Hicks elasticity, in this case the intertemporal substitution response is *smaller* than the Hicks elasticity, which is $\left(\frac{P^M}{\phi_2}\right)\left(\frac{P^\phi U_{CC}\left(\frac{\partial C_2}{\partial P^M}\right) - V_{\phi M}\left(\frac{\partial M_2}{\partial P^M}\right)}{V_{\phi\phi}}\right) > 0$ since $U_{CC}, V_{\phi\phi} < 0$ and $\left(\frac{\partial C_2}{\partial P^M}\right) > 0$ since consumption and medical spending are substitutes. For more details on these elasticities' relative magnitudes in the labor supply context, see Keane (2011). The reason for this difference is that in the labor supply model, "hours worked" are a "bad" rather than a "good", and so the utility function has to be convex with respect to hours of work, whereas here investment in health is a "good", and so $V_{\phi\phi} < 0$. If we flip the sign of $V_{\phi\phi}$ in the elasticities in this chapter, we obtain that the elasticity of intertemporal substitution is larger than the Hicks just as in the labor supply case. The intuition is that for labor supply, if lifetime consumption possibilities can be altered by working more hours, diminishing marginal utility of consumption will dampen the response to an increase in the wage since higher consumption is traded for the "bad" of less leisure time. This dampening effect is not present if the marginal utility of wealth is constant. By constrast, in this case, prevention expands lifetime consumption possibilities (by depressing the marginal utility of treatment) and is also valued in itself, and so there is even more reason to use it if the price of treatment rises and the marginal utility of wealth is not held constant.

$\left(\frac{\partial M_2}{\partial P^M}\right) < 0$ due to the Law of Demand;

$V_{\phi\phi} < 0$ due to diminishing marginal returns to prevention;

$V_{\phi M} < 0$ since prevention *ex ante* and treatment *ex post* are substitutes (more prevention lowers the marginal utility of treatment).

It follows that $\epsilon^F_{\phi_2, P^M} > 0$: an anticipated fall in the price of medical services unambiguously decreases the incentives to use prevention, and an anticipated rise in the price of medical services unambiguously increases the incentives to use prevention.

Since almost all Americans expect to be entitled to Medicare upon turning 65, the marginal utility of wealth does not change when they qualify for Medicare. Medicare RDD papers therefore estimate this elasticity, which is unambiguously positive (all else equal), unlike the experimental studies which estimate the Marshallian elasticity. Similarly, the main empirical exercises in this chapter estimate an extensive-margin elasticity of intertemporal substitution. This illustrates an advantage of using a Medicare RDD rather than an experiment to study this behavior. Note that, as per footnote 10, this change in the timing of preventive efforts is a *lower* bound for the "long-run" or "lifetime" change in the use of prevention, which is measured by the Hicks elasticity.

The usual interpretation of the elasticity of intertemporal substitution provides some intuition for its role in this context. Individuals move along their life-cycle consumption profiles each period according to their planned allocation of their lifetime resources across different periods. In the labor supply context, the elasticity of intertemporal substitution measures the proportional planned increase in hours worked for a proportionally higher wage relative to other periods with proportionally lower wages (since wages vary over the life cycle). In this context, the elasticity measures the response of the planned division between spending on prevention in the current period versus spending on treatment in a subsequent period to anticipated variation in the price of treatment over the life cycle. Over the life cycle, individuals will want to allocate more prevention to periods in which the price of treatment is high and less to those in which the price of treatment is low. Given some estimate of how the price of treatment differs in the Medicare-eligible portion of the life cycle

versus the uninsured portion of the life cycle, we can estimate how much variation over the life cycle in prevention spending individuals are willing to undertake to track the price of treatment. Since moving across the Medicare eligibility age is anticipated by almost all United States residents, it is this elasticity for which Medicare RDD results are relevant.[11]

The simple expression for the Frisch elasticity reconciles previous results and commentary on the literature on prevention with the results found in this chapter. The size of the effect depends on the relative sizes of the second derivative of utility with respect to ϕ itself - or the rate at which marginal utility from health diminishes - and the strength of the link between health investments and medical expenses, $V_{\phi M}$. The usual explanation for why $\epsilon^{F}_{\phi_2,P^M}$ may be small is that $V_{\phi\phi}$ is large - individuals are highly risk averse with respect to their health (Cutler and Zeckhauser, 2000, Kenkel, 2000). In this chapter, even if this holds, this can be outweighed by a sufficiently tight link between medical expenses and health investments, hence greater substitutability between prevention *ex ante* and

[11]One reason this link might fail is if the population in question cannot borrow to smooth expenditures across periods. Then consumption will track changes in the budget constraint each period, instead of changes in the lifetime budget constraint (Deaton, 1992). In that case, Medicare RDD papers would also estimate a Marshallian elasticity, income effects and all, rather than an intertemporal elasticity of substitution. Whether the estimates in this chapter underestimate or overestimate the true elasticity of intertemporal substitution depends on two opposing mechanisms. (In a similar vein, Keane and Wolpin (2001) argue that the presence of liquidity constraints has led to underestimates of the elasticity of intertemporal substitution in consumption). First, the income effect increases prevention at the age of Medicare eligibility, offsetting the negative cross-price effect. Second, liquidity constraints strengthen the precautionary motive Deaton (1992), since agents need to have a larger buffer against consumption fluctuations if they cannot borrow in a crisis. Therefore agents who are unable to borrow against future income will have a stronger motivation to use prevention, and exhibit larger decreases in prevention in the face of an exogenous decrease in risk. Hence it is ambiguous whether the true effect of Medicare eligibility on prevention is over- or under-estimated due to the presence of credit constraints. The argument that the responses in this chapter are Marshallian elasticities rather than elasticities of intertemporal substitution creates more of a puzzle, since it implies both that the estimators in this chapter recover the same elasticity as in the RAND or Oregon health insurance experiments, and yet that the extent of *ex ante* moral hazard in those cases is significantly smaller.

Smoothing consumption across covered and uncovered periods requires access to a store of liquid wealth that can be used to finance more expensive medical care in the uncovered period. To examine the evidence for this, I calculate the amount of liquid wealth held by households in which uninsured diabetic women reside at ages 60-64. The Health and Retirement Study surveys contain a large number of questions regarding pension wealth, assets, and debts. To proxy for a lack of access to liquidity, I calculate the number of uninsured diabetic women aged 60-64 in each wave of the HRS between 1998 and 2006 who live in households with neither housing wealth that could serve as collateral for a loan nor a positive amount of non-housing financial wealth. I use the definition of non-housing financial wealth in the RAND HRS data, which comprises stocks, bonds, checking accounts and certificates of deposit minus debts, and excludes the value of IRAs, Keogh plans, real estate, business wealth, and vehicles (Chien et al., 2013). 18% of this group are liquidity constrained according to this definition. If this forces consumption to track income instead of being smoothed over the age of 65, then we should see evidence of deviations from the permanent income hypothesis in other categories of expenditure as well. In the Appendix, I report regression discontinuity results for the non-medical expenditures on durable goods of households inhabited by female diabetics who are not insured by Medicaid. I am unable to reject the null hypothesis that non-medical durable consumption does not change discontinuously at age 65.

treatment *ex post* (i.e. $V_{\phi M}$ is large in magnitude), as exists among diabetics.[12]

This model is useful for its ability to shed light on heterogeneous effects of changing the price of medical services in a subsequent period on current-period incentives to use prevention (here, as in the rest of this chapter, used as a synonym for investment in health capital *ex ante*). Agents whose intrinsic utility from preventive care is more concave (larger $V_{\phi\phi}$) have weaker preventive care responses to the price of medical services. That is, agents that are more risk-averse with respect to their future health have smaller responses in their preventive behavior to a reduction in the price of medical services. This effect is counteracted by the relative effectiveness of prevention in reducing demand for future medical services in the second period, $V_{\phi M}$. The larger this term is - and so the tighter is the link between prevention and future demand for medical services - the larger is the reduction in prevention for a given fall in the price of future medical services.

This framework allows us to reconcile previous findings of small changes in health investment behavior upon qualifying for Medicare coverage (Dave and Kaestner, 2009, Card et al., 2008) and the large effects found in this chapter. Though smokers and binge drinkers may be less risk-averse with respect to their future health status than others (and so have smaller $V_{\phi\phi}$), the link between their behavior and their future medical expenses at the margin at age 65 is likely to be relatively weak (and so they also have smaller $V_{\phi M}$). The link between investment in health and future medical expenses is much stronger for diabetics than most other subsets of the population, and so given that $V_{\phi M}$ is relatively large for this group, we also see larger responses to a reduction in the price of medical services in preventive behavior in this group.

We can also use this model to address a previous explanation for the weak *ex ante* moral hazard effects found in the literature. Cutler and Zeckhauser (2000) and Kenkel (2000) pointed out that even if agents are insured against the financial losses of illness, they are in general not insured against the expected utility losses of ill health. This is captured in the model by a higher risk

[12]One alternative to the approach in this section to reconciling the twin facts that individuals both seek medical care when ill and choose to increase their probability of falling ill in the future is to use hyperbolic discounting or, in the limit, the model proposed in Banerjee and Mullainathan (2010). In a version of that model derived by the author and considered for this chapter, individuals care about their current health and so seek medical care, but not their future health, and so spend their income on consumption rather than prevention. The model presented in its stead has the advantages of (i) more parsimonious assumptions, (ii) a straightforward ability to link the empirical results in this chapter with those in the rest of the literature, and (iii) quantitative predictions for behavioral responses that have yet to be studied.

aversion over health (larger $V_{\phi\phi}$) leading to a smaller response of prevention to the price of medical services. The previous theoretical explanation in this model corresponds to high risk aversion with respect to health across individuals. This model shows that if the connection between investment in health and eventual medical expenses is particularly strong (high $V_{\phi M}$), as it is for diabetics to a far greater extent than for the general population, this previous explanation can still be valid for the broader near-elderly population without ruling out large responses of the kind found in this chapter.

In sum, a simple two-period model that introduces prevention as a choice variable that affects the marginal utility from medical services can reconcile the following observations: (1) Agents receive utility from medical services, and demand them when sick; (2) Even given the value that they place on medical services, if prevention reduces their future demand for medical services, lowering the price of medical services will reduce the incentive to use prevention; (3) The stronger the link between prevention and future demand for medical services, the stronger is the crowding-out effect.

We can therefore explain the weak effects found in previous studies, as well as why diabetics are a subset of the population among whom we would expect to find strong effects, with a simple two-period model of investment in health.[13]

I close this section with two further remarks that may be of use for future work. First, if $V_{\phi M}$ varies across individuals, this framework shows that the responsiveness of prevention to the price of medical services is strongest among those individuals with the strongest link between prevention and their medical expenses. From a policy perspective, this means that the crowding out of prevention by coverage for treatment is greatest among those individuals whose medical expenses are likeliest to increase to a large extent as a result. The adverse effects of coverage are concentrated among precisely the individuals that a policymaker would least want to discourage from using prevention. The model in this section therefore allows me to make qualitative statements regarding the marginal individuals for whom prevention is crowded out by insurance for treatment.

Second, this model can also show how Marshallian elasticities of prevention with respect to

[13]If we had data that allowed us to calculate the magnitude of the average fall in the price of insulin for the marginal diabetic upon qualifying for Medicare post-2006, we could also use the model to rationalize the relative sizes of the cross-price effect due to the price of treatment falling at age 65 and the own-price effect due to the price of insulin falling at age 65 in the post-Part D era.

insurance that are exactly zero can coexist with substantial magnitudes for the Hicks and Frisch elasticities. Note that by the Slutsky equation, the Marshallian derivative $\left(\frac{\partial \phi_2}{\partial P^M}\right)^M$ can be written

$$\begin{aligned}
\left(\frac{\partial \phi_2}{\partial P^M}\right)^M &= \left(\frac{\partial \phi_2}{\partial P^M}\right)^H - \frac{\partial \phi_2}{\partial A_1} M_2, \\
\implies \frac{P^M}{\phi_2}\left(\frac{\partial \phi_2}{\partial P^M}\right)^M &= \frac{P^M}{\phi_2}\left(\frac{\partial \phi_2}{\partial P^M}\right)^H - \frac{P^M}{\phi_2}\left(\frac{A_1}{A_1}\right)\frac{\partial \phi_2}{\partial A_1} M_2, \\
\implies \epsilon^M_{\phi_2, P^M} &= \epsilon^H_{\phi_2, P^M} - \frac{P^M M_2}{A_1}\epsilon_{\phi_2, A_1},
\end{aligned}$$

so that the Marshallian elasticity is equal to the Hicks elasticity less the product of the income elasticity of prevention ϵ_{ϕ_2, A_1} and future medical expenses' share of lifetime wealth $\frac{P^M M_2}{A_1}$. This is analogous to the case of the static labor supply model, except instead of a term depending on the ratio of labor income to lifetime wealth, we have a term depending on the ratio of future medical expenses to lifetime wealth. The main difficulty in recovering the Hicks elasticity from a given Marshall elasticity then comes from estimating $\frac{P^M M_2}{A_1}$ and ϵ_{ϕ_2, A_1}. For the sake of argument, suppose the Marshallian elasticity is 0, so that we can recover $\epsilon^H_{\phi_2, P^M} = \frac{P^M M_2}{A_1}\epsilon_{\phi_2, A_1}$. Banks et al. (2016) estimate that by age 70, medical expenses are on average 20% of household spending in the United States, which gives $\frac{P^M M_2}{A_1} = 0.2$.[14] There is as yet no consensus on the income elasticity of prevention, but if we take the upper end of estimates of the income elasticity for dental care - assuming that most dental care is preventive rather than palliative - from a survey of the literature by Getzen (2000), we obtain $\epsilon_{\phi_2, A_1} = 3.2$ from Silver (1972). This gives a Hicks elasticity of $\epsilon^H_{\phi_2, P^M} = 0.2 \times 3.2 = 0.64$. This is a much larger Hicks elasticity of health investments with respect to the price of health care than the majority of studies of labor supply find for the

[14] Since the 20% budget share of health care pertains to over-70s, one objection to its use is that it is an overestimate of the ratio of medical expenses to *lifetime* income. As a robustness check, I examine the distribution of the ratio of out-of-pocket medical expenses to household income for female diabetics under 65 not enrolled in Medicaid. The mean share is 23%, with an upper quartile of 16%. If I use the smaller figure of 16% for the calculations above, I obtain a Hicks elasticity of 0.512, only slightly above the upper bound for the Frisch elasticity (which is nonetheless consistent with the argument that the larger estimates are from an estimator biased away from zero). The minimal income elasticity consistent with my empirical results increases to 1.26, which is still within the range of elasticities considered by Finkelstein et al. (2012). This is subject to the caveat that current income may systematically differ from permanent income to different extents over the life cycle (Haider and Solon, 2006).

response of hours worked to wages (Keane, 2011). Failure to reject the null hypothesis of no *ex ante* moral hazard effect of health insurance on health investments in an experimental setting is consistent with large *ex ante* moral hazard effects after taking the income effect of providing coverage into account.

To relate the estimates in this chapter to the Frisch elasticity of prevention with respect to the price of treatment, first take the estimates which find between a 12.9% and 30.5% relative decrease in the proportion of insulin users.[15] We need to find the expected relative percentage decrease for a 1% difference in the price of health care when 65 relative to pre-65. Medicare Parts A and B typically cover between 60% and 80% of beneficiaries' health care costs. So the results in this chapter imply that $\epsilon^F_{\phi_2,P^M}$ is in the interval $[\frac{0.129}{0.8}, \frac{0.305}{0.6}] = [0.161, 0.508]$. In line with the theoretical model presented in this section, the elasticity of intertemporal substitution is smaller than the Hicks elasticity (see footnote 10, above). This comes with the caveats that the two estimates do not come from the same data, and I do not present a consistent estimator of the income elasticity of insulin in this chapter. The lower bound for the Frisch elasticity of 0.161 also allows for the possibility that the Hicks elasticity is smaller than 0.64 while maintaining the result that the Marshallian elasticity is zero. It therefore also tells us that, holding the budget share of future medical expenses constant, the smallest income elasticity of demand for insulin consistent with the results in this chapter, assuming a Marshallian elasticity of zero, is $\frac{0.161}{0.8\times0.2} = 1.01$. Hence we need insulin to be a luxury good on average to be consistent with the results found in this chapter. This is greater than the midpoint of the preferred range for the income elasticity of healthcare between 0 and 1.5 used by Finkelstein et al. (2012). There are nonetheless good reasons to believe this number is larger for the subcategory of preventive medicine than for health care overall; individuals' ability to delay usage is much greater for prevention than for treatment, for example. In addition, if we could observe which diabetics were Type I diabetics - who need insulin to survive - as opposed to Type IIs, the vast majority, we could separately estimate income elasticities of demand for the two subgroups. The income elasticity for Type I diabetics with respect to insulin is likely to be significantly lower

[15] The 12.9% figure comes from dividing the smaller first-differences result of a 4.4 percentage point decrease by the original cohort 60-64 average insulin usage of 34 percentage points, yielding a conservative estimate of the effect.

than for Type IIs for the reasons cited above.

The model used in this section cannot capture all elements of the health investment process. This may change some of the conclusions regarding the elasticities that are measured in the empirical section of this chapter. Imai and Keane (2004) and Keane (2016) show that the relative magnitudes of the compensated and intertemporal substitution elasticities changes when human capital accumulation is endogenous in labour supply models. The reason is that current actions affect future prices, and hence the timing of actions across periods as well. Therefore although in a labor supply model without learning-by-doing, the compensated labour supply elasticity is smaller than the elasticity of intertemporal substitution, when learning-by-doing is included this inequality can reverse. This is because deferring hours worked to later periods reduces human capital accumulated for all future periods, making substituting hours worked across time more costly even if the wage next period is higher than the current wage. By contrast, if wages in all future periods are higher, then there is a compounded increase in the return to hours worked today via their effect on accumulating human capital. This leads to larger responses to permanent wage changes than temporary wage changes. This line of argument may have an analogous effect in the setting considered in this paper. Though not derived here, we may expect that the extent to which the compensated elasticity of investment in health capital is larger than the intertemporal substitution elasticity will depend on the extent to which the return to investing in health capital cumulates over time. These considerations become more complicated by the fact that the regression-discontinuity estimates recover a Local Average Treatment Effect (Hahn et al., 2001). If there is heterogeneity in individuals' returns to investing in health capital before age 65, then the interpretation of the elasticities is conditional on the subset of individuals who are induced to change their insurance status by qualifying for Medicare coverage at age 65. Suppose these individuals have negligible returns from investing in health capital. Then they may be highly willing to substitute health capital investment across periods with different prices for insurance. At the same time, this subset of individuals may react to a *smaller* extent to unanticipated permanent changes in the price of insurance. This would reverse the conclusion from this section that the policy elasticity of interest (the compensated response) is larger than the measured response (the intertemporal substitution

elasticity). There are some reasons to think that the individuals with the highest rates of return on investment in health capital are those who become insured at age 65 after previously being uninsured. The previously uninsured may be expected to have relatively smaller endowments of health and hence have the most to gain from any investments in it, assuming diminishing marginal returns to investment. This is particularly true in the United States, where access to health insurance is closely tied to employment and pre-existing wealth. An adverse selection argument would cut against this reasoning and suggest that those who opted not to have insurance were those who were healthy enough to do without it. Which of these forces is more important is an empirical question that I leave to future research.

Another caveat is that investment in health enters directly into the utility function, and so the model cannot distinguish separate components of the costs and benefits of investing in health. All of the marginal costs of investment in health are absorbed into the price of prevention P^{ϕ}, and all of the marginal benefits are captured by the marginal utility of investment V_{ϕ}. Grossman (2000) cites Ried (1998) as a full treatment of comparative dynamics in a full Grossman model which includes the investment and consumption components of health capital. Ried (1998) derives Frisch decision functions for inputs into health, consumption and labour supply, and discusses the Frisch elasticity of the demand for treatment in period t for change in the price of medical treatment in period s, obtaining (retaining the notation from this section):

$$\left.\frac{dM_t}{dP_s^M}\right|_{\lambda=\bar{\lambda}} = \left.\frac{\partial M_t}{\partial P_t^M}\right|_{\lambda=\bar{\lambda}} + \left.\frac{\partial M_t}{\partial P_s^{\phi}}\right|_{\lambda=\bar{\lambda}} \times \frac{M_t}{\phi_t} \tag{1.13}$$

which, by symmetry, implies

$$\left.\frac{d\phi_t}{dP_s^M}\right|_{\lambda=\bar{\lambda}} = \left.\frac{\partial \phi_t}{\partial P_t^M}\right|_{\lambda=\bar{\lambda}} + \left.\frac{\partial \phi_t}{\partial P_s^{\phi}}\right|_{\lambda=\bar{\lambda}} \times \frac{\phi_t}{M_t} \tag{1.14}$$

the reason for this given in Ried (1998) is that a change in the price of medical care in s, P_s^M, implies a change in the marginal cost of gross investment P_s^{ϕ}. The reason for this is that in Ried's model,

the time path of the marginal cost of gross investment in health depends on the evolution of the health stock over time, which is itself affected by the change in the price of treatment P_t^M. Thus an elasticity that is unambiguously positive in the model used in this chapter may be either positive or negative in Ried's model. By abstracting away from the dynamics of investment in health capital, the model in this chapter implicitly assumes that this change in the time path of the marginal cost of investment is negligible. This may be the case if the individuals modelled do not expect to live long after the period when the anticipated price change occurs. Even so, the relationship between the compensated and intertemporal substitution (Frisch) elasticities is no longer unambiguous in theory as it is in this chapter's model. Khwaja (2010) estimates a similar (but not identical) model but does not discuss its comparative dynamics or cite Ried (1998).

1.7 Conclusion

The effect of insurance on health outcomes and behaviors depends on what is covered and to what extent. This chapter has provided evidence that insuring individuals against health risks can worsen those risks to a larger extent than previously thought. Before 2006, Medicare Parts A and B insured 60-80% of previously uninsured female diabetics' health care costs, but did not subsidize insulin. As a result, between 12.9% and 30.5% of the insulin users in this group would forego using insulin in the month of Medicare eligibility, when their risk of incurring large medical expenses would drop discontinuously. This chapter also provides an organizing theoretical framework to explain why previous studies have encountered more difficulty with recovering *ex ante* moral hazard in health behaviors. When the connection between a health behavior and health care costs is weaker, the underlying *ex ante* moral hazard effect is smaller. Individuals with the strongest responses will "hide in the herd" in data which contain a large number of individuals with weak responses. In addition, if the insurance is provided via randomized assignment, income effects are likely to be even larger for prevention than for other forms of health care, which further masks the negative effect of insurance for treatment on prevention.

The main policy implication of this chapter's findings is that policymakers have underestimated

the need for stronger incentives to use preventive care. In a universal health care system, it is nearly impossible - in the absence of invoking strong assumptions regarding counterfactual behaviour - to see the effect of coverage on prevention since there is no control group. In the United States, universal coverage for those above 65 and non-universal coverage for those below 65 provides some evidence on how coverage can crowd out prevention at the margin. Though many universal health care systems spend more on preventive services on average than public health programs in the United States, the evidence provided in this chapter suggests that the level of this spending may be an underestimate of what is necessary to minimise overall health care costs. This is due to the unseen crowding out effect which cannot be estimated in a universal health care system since there is no "control" group of uninsured individuals.

The results in this chapter suggest estimation of the Marshall, Hicks and Frisch elasticities of prevention with respect to the price of treatment as a profitable avenue for future research. With rich enough data, one could estimate all of the Marshall, Hicks and Frisch elasticities using a single data set. This would be an advance over comparisons of these elasticities across papers since differences across papers may also be due to unknown differences in data or estimation choices.

References

Aiyagari, S. R. (1994). Uninsured Idiosyncratic Risk and Aggregate Saving. *Quarterly Journal of Economics*, 109(3):659–684.

Alpert, A. (2016). The Anticipatory Effects of Medicare Part D on Drug Utilization. *Journal of Health Economics*, 49:28–45.

Armstrong, T. B. and Kolesár, M. (2017). A Simple Adjustment for Bandwidth Snooping. *Review of Economic Studies*, 85(2):732–765.

Armstrong, T. B. and Kolesár, M. (2018b). Optimal Inference in a Class of Regression Models. *Econometrica*, 86(2):655–683.

Attanasio, O. P. and Hoynes, H. W. (2000). Differential Mortality and Wealth Accumulation. *Journal of Human Resources*, pages 1–29.

Banerjee, A. and Mullainathan, S. (2010). The Shape of Temptation: Implications for the Economic Lives of the Poor. Unpublished manuscript.

Banks, J., Blundell, R., Levell, P., and Smith, J. P. (2016). Life-Cycle Consumption Patterns at Older Ages in the US and the UK: Can Medical Expenditures Explain the Difference? Technical report, National Bureau of Economic Research.

Barcellos, S. H. and Jacobson, M. (2015). The Effects of Medicare on Medical Expenditure Risk and Financial Strain. *American Economic Journal: Economic Policy*, 7(4):41–70.

Barrett-Connor, E., Giardina, E.-G. V., Gitt, A. K., Gudat, U., Steinberg, H. O., and Tschoepe, D. (2004). Women and Heart Disease: the Role of Diabetes and Hyperglycemia. *Archives of Internal Medicine*, 164(9):934–942.

Barton, P., Andronis, L., Briggs, A., McPherson, K., and Capewell, S. (2011). Effectiveness and Cost Effectiveness of Cardiovascular Disease Prevention in Whole Populations: Modelling Study. *British Medical Journal*, 343:d4044.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 119(1):249–275.

Blau, D. M. and Gilleskie, D. B. (2006). Health Insurance and Retirement of Married Couples. *Journal of Applied Econometrics*, 21(7):935–953.

Blau, D. M. and Gilleskie, D. B. (2008). The Role of Retiree Health Insurance in the Employment Behavior of Older Men. *International Economic Review*, 49(2):475–514.

Boland, B. (1998). The Evolution of Best-in-Class Pharmacy Management Techniques. *Journal of Managed Care Pharmacy*, 4(4):366–373.

Bommer, C., Heesemann, E., Sagalova, V., Manne-Goehler, J., Atun, R., Bärnighausen, T., and Vollmer, S. (2017). The Global Economic Burden of Diabetes in Adults Aged 20–79 Years: a Cost-of-Illness Study. *The Lancet Diabetes & Endocrinology*, 5(6):423–430.

Brown, J. B., Nichols, G. A., Glauber, H. S., and Bakst, A. (1999). Ten-Year Follow-Up of Antidiabetic Drug Use, Nonadherence, and Mortality in a Defined Population with Type 2 Diabetes Mellitus. *Clinical Therapeutics*, 21(6):1045–1057.

Buchmueller, T., Johnson, R. W., and Lo Sasso, A. T. (2006). Trends in Retiree Health Insurance, 1997–2003. *Health Affairs*, 25(6):1507–1516.

Caetano, C., Caetano, G., and Escanciano, J. C. (2017). Over-Identified Regression Discontinuity Design. Unpublished manuscript.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014a). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6):2295–2326.

Card, D., Dobkin, C., and Maestas, N. (2008). The Impact of Nearly Universal Insurance Coverage on Health Care Utilization and Health: Evidence from Medicare. *American Economic Review*, 98(5):2242–2258.

Card, D., Dobkin, C., and Maestas, N. (2009). Does Medicare Save Lives? *Quarterly Journal of Economics*, 124(2):597–636.

Carroll, C. D., Dynan, K. E., and Krane, S. D. (2003). Unemployment Risk and Precautionary Wealth: Evidence from Households' Balance Sheets. *Review of Economics and Statistics*, 85(3):586–604.

Centers for Disease Control and Prevention (2017). National Diabetes Statistics Report, 2017. *Atlanta, GA: Centers for Disease Control and Prevention, US Dept of Health and Human Services.*

Chetty, R. (2008). Moral Hazard Versus Liquidity and Optimal Unemployment Insurance. *Journal of Political Economy*, 116(2):173–234.

Chetty, R. and Szeidl, A. (2007). Consumption Commitments and Risk Preferences. *Quarterly Journal of Economics*, 122(2):831–877.

Chien, S., Campbell, N., Hayden, O., et al. (2013). *RAND HRS Data Documentation, Version M*. Santa Monica, CA: RAND Center for the Study of Aging.

Choi, A., Dave, D., and Sabia, J. J. (2016). Smoke Gets in Your Eyes: Medical Marijuana Laws and Tobacco Use. Unpublished manuscript.

Christian-Herman, J., Emons, M., and George, D. (2004). Effects of Generic-Only Drug Coverage in a Medicare HMO. *Health Affairs*, 23:455–468.

Clark, R. L. and Mitchell, O. S. (2014). How does Retiree Health Insurance Influence Public Sector Employee Saving? *Journal of Health Economics*, 38:109–118.

Cubanski, J. and Neuman, P. (2007). Status Report on Medicare Part D Enrollment in 2006: Analysis of Plan-Specific Market Share and Coverage. *Health Affairs*, 26(1):w1–w12.

Curcuru, S., Heaton, J., Lucas, D., and Moore, D. (2010). Heterogeneity and Portfolio Choice: Theory and Evidence. *Handbook of Financial Econometrics*, pages 337–382.

Cutler, D. M. and Zeckhauser, R. J. (2000). The Anatomy of Health Insurance. In *Handbook of Health Economics*, volume 1, pages 563–643. Elsevier.

Dave, D. and Kaestner, R. (2009). Health Insurance and Ex Ante Moral Hazard: Evidence from Medicare. *International Journal of Health Care Finance and Economics*, 9(4):367.

De Nardi, M., French, E., and Jones, J. B. (2010). Why Do the Elderly Save? The Role of Medical Expenses. *Journal of Political Economy*, 118(1):39–75.

De Preux, L. (2011). Anticipatory Ex Ante Moral Hazard and the Effect of Medicare on Prevention. *Health Economics*, 20(9):1056–1072.

Deaton, A. (1992). *Understanding Consumption*. Oxford University Press.

Diabetes Control and Complications Trial Research Group (1993). The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus. *New England Journal of Medicine*, 329(14):977–986.

Duggan, M. and Scott Morton, F. (2010). The Effect of Medicare Part D on Pharmaceutical Prices and Utilization. *American Economic Review*, 100(1):590–607.

Eeckhoudt, L., Gollier, C., and Schlesinger, H. (1996). Changes in Background Risk and Risk Taking Behavior. *Econometrica*, pages 683–689.

Einav, L. and Finkelstein, A. (2018). Moral Hazard in Health Insurance: What We Know and How We Know It. *Journal of the European Economic Association*, 16(4):957–982.

Eliasson, B. (2003). Cigarette Smoking and Diabetes. *Progress in Cardiovascular Diseases*, 45(5):405–413.

Engen, E. M. and Gruber, J. (2001). Unemployment Insurance and Precautionary Saving. *Journal of Monetary Economics*, 47(3):545–579.

Fang, H., Keane, M., Khwaja, A., Salm, M., and Silverman, D. (2007). Testing the Mechanisms of Structural Models: The Case of the Mickey Mantle Effect. *American Economic Review*, 97(2):53–59.

Feir, D., Lemieux, T., and Marmer, V. (2016). Weak Identification in Fuzzy Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 34(2):185–196.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics*, 127(3):1057–1106.

Frean, M., Gruber, J., and Sommers, B. D. (2017). Premium Subsidies, the Mandate, and Medicaid Expansion: Coverage Effects of the Affordable Care Act. *Journal of Health Economics*, 53:72–86.

French, E. and Jones, J. B. (2011). The Effects of Health Insurance and Self-Insurance on Retirement Behavior. *Econometrica*, 79(3):693–732.

Gelman, A. and Imbens, G. (2018). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, pages 1–10.

Getzen, T. E. (2000). Health Care is an Individual Necessity and a National Luxury: Applying Multilevel Decision Models to the Analysis of Health Care Expenditures. *Journal of Health Economics*, 19(2):259–270.

Gilmer, T. P., O'connor, P. J., Manning, W. G., and Rush, W. A. (1997). The Cost to Health Plans of Poor Glycemic Control. *Diabetes Care*, 20(12):1847–1853.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80(2):223–255.

Grossman, M. (2000). The Human Capital Model. In *Handbook of Health Economics*, volume 1, pages 347–408. Elsevier.

Guerrieri, V. and Lorenzoni, G. (2017). Credit Crises, Precautionary Savings, and the Liquidity Trap. *Quarterly Journal of Economics*, 132(3):1427–1467.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209.

Haider, S. and Solon, G. (2006). Life-Cycle Variation in the Association Between Current and Lifetime Earnings. *American Economic Review*, 96(4):1308–1320.

Hua, X., Carvalho, N., Tew, M., Huang, E. S., Herman, W. H., and Clarke, P. (2016). Expenditures and Prices of Antihyperglycemic Medications in the United States: 2002-2013. *Journal of the American Medical Association*, 315(13):1400–1402.

Hubbard, R. G., Skinner, J., and Zeldes, S. P. (1995). Precautionary Saving and Social Insurance. *Journal of Political Economy*, 103(2):360–399.

Hurd, M., Rohwedder, S., Carroll, J., Mallett, J., and McCullough, C. (2015). Rand cams data documentation. *Version D, Release*, 2:11.

Imai, S. and Keane, M. P. (2004). Intertemporal Labor Supply and Human Capital Accumulation. *International Economic Review*, 45(2):601–641.

Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies*, 79(3):933–959.

Johnson, R. W., Smith, K. E., and Haaga, O. (2013). How Did the Great Recession Affect Social Security Claiming? *Program on Retirement Policy Brief*, 37.

Joyce, G. F., Goldman, D. P., Karaca-Mandic, P., and Zheng, Y. (2007). Pharmacy Benefit Caps and the Chronically Ill. *Health Affairs*, 26(5):1333–1344.

Juutilainen, A., Kortelainen, S., Lehto, S., Rönnemaa, T., Pyörälä, K., and Laakso, M. (2004). Gender Difference in the Impact of Type 2 Diabetes on Coronary Heart Disease Risk. *Diabetes Care*, 27(12):2898–2904.

Kahn-Lang, A. and Lang, K. (2019). The Promise and Pitfalls of Differences-in-Differences: Reflections on "Sixteen and Pregnant" and Other Applications. *Journal of Business & Economic Statistics*.

Kaplan, G. and Violante, G. L. (2014). A Model of the Consumption Response to Fiscal Stimulus Payments. *Econometrica*, 82(4):1199–1239.

Kautzky-Willer, A., Harreiter, J., and Pacini, G. (2016). Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. *Endocrine Reviews*, 37(3):278–316.

Keane, M. P. (2011). Labor Supply and Taxes: A Survey. *Journal of Economic Literature*, 49(4):961–1075.

Keane, M. P. (2016). Life-Cycle Labour Supply with Human Capital: Econometric and Behavioural Implications. *Economic Journal*, 126(592):546–577.

Keane, M. P. and Wolpin, K. I. (2001). The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment. *International Economic Review*, 42(4):1051–1103.

Kenkel, D. (2000). Prevention. In Culyer, A. J. and Newhouse, J. P., editors, *Handbook of Health Economics*, volume 1, chapter 31, pages 1675–1720. Elsevier, North Holland.

Khwaja, A. (2009). Investment in Human Capital, Longevity and Moral Hazard in a Stochastic Life-Cycle Model of Demand for Health. Unpublished manuscript.

Khwaja, A. (2010). Estimating Willingness to Pay for Medicare Using a Dynamic Life-Cycle Model of Demand for Health Insurance. *Journal of Econometrics*, 156(1):130–147.

Kolesár, M. and Rothe, C. (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, 108(8):2277–2304.

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

MaCurdy, T. E. (1981). An Empirical Model of Labor Supply in a Life-Cycle Setting. *Journal of Political Economy*, 89(6):1059–1085.

McGuire, T. G., Newhouse, J. P., and Sinaiko, A. D. (2011). An Economic History of Medicare Part C. *Milbank Quarterly*, 89(2):289–332.

Menke, A., Casagrande, S., Geiss, L., and Cowie, C. C. (2015). Prevalence of and Trends in Diabetes among Adults in the United States, 1988-2012. *Journal of the American Medical Association*, 314(10):1021–1029.

Nathan, D. M., Buse, J. B., Davidson, M. B., Ferrannini, E., Holman, R. R., Sherwin, R., and Zinman, B. (2009). Medical Management of Hyperglycemia in Type 2 Diabetes: A Consensus Algorithm for the Initiation and Adjustment of Therapy: a Consensus Statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care*, 32(1):193–203.

Papanicolas, I., Woskie, L. R., and Jha, A. K. (2018). Health Care Spending in the United States and Other High-Income Countries. *Journal of the American Medical Association*, 319(10):1024–1039.

Piette, J. D., Heisler, M., and Wagner, T. H. (2004). Problems Paying Out-of-Pocket Medication Costs Among Older Adults with Diabetes. *Diabetes Care*, 27(2):384–391.

Rajagopalan, R., Joyce, A., Ollendorf, D., and Murray, F. (2003). PDB5: Medication Compliance in Type 2 Diabetes Subjects: Retrospective Data Analysis. *Value in Health*, 6(3):328.

Ried, W. (1998). Comparative Dynamic Analysis of the Full Grossman Model. *Journal of Health Economics*, 17(4):383–425.

Rust, J. and Phelan, C. (1997). How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Markets. *Econometrica*, pages 781–831.

Saaddine, J. B., Engelgau, M. M., Beckles, G. L., Gregg, E. W., Thompson, T. J., and Narayan, K. V. (2002). A Diabetes Report Card for the United States: Quality of Care in the 1990s. *Annals of Internal Medicine*, 136(8):565–574.

Salas, M., Hughes, D., Zuluaga, A., Vardeva, K., and Lebmeier, M. (2009). Costs of Medication Nonadherence in Patients with Diabetes Mellitus: a Systematic Review and Critical Analysis of the Literature. *Value in Health*, 12(6):915–922.

Shea, J. (1997). Instrument Relevance in Multivariate Linear Models: A Simple Measure. *Review of Economics and Statistics*, 79(2):348–352.

Silver, M. (1972). An Economic Analysis of Variations in Medical Expenses and Work-Loss Rates. In Klarman, H., editor, *Essays in the Economics of Health and Medical Care*, pages 97–118. NBER.

Sommers, B. D., Gawande, A. A., and Baicker, K. (2017). Health Insurance Coverage and Health - What the Recent Evidence Tells Us. *New England Journal of Medicine*, 377(6):586–593.

Soumerai, S. and Ross-Degnan, D. (1999). Inadequate Prescription-Drug Coverage for Medicare Enrollees–a Call to Action. *New England Journal of Medicine*, 340(9):722–728.

Soumerai, S. B., Pierre-Jacques, M., Zhang, F., Ross-Degnan, D., Adams, A. S., Gurwitz, J., Adler, G., and Safran, D. G. (2006). Cost-Related Medication Nonadherence Among Elderly and

Disabled Medicare Beneficiaries: a National Survey 1 Year Before the Medicare Drug Benefit. *Archives of Internal Medicine*, pages 1829–1835.

Stock, J. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In Andrews, Donald & Stock, J., editor, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press.

Von Wachter, T. (2002). *The End of Mandatory Retirement in the US: Effects on Retirement and Implicit Contracts*. Center for Labor Economics, University of California, Berkeley.

Weinger, K. and Beverly, E. A. (2010). Barriers to Achieving Glycemic Targets: Who Omits Insulin and Why? *Diabetes Care*, 33:450–452.

Wild, S., Roglic, G., Green, A., Sicree, R., and King, H. (2004). Global Prevalence of Diabetes: Estimates for the Year 2000 and Projections for 2030. *Diabetes Care*, 27(5):1047–1053.

Zhuo, X., Zhang, P., Barker, L., Albright, A., Thompson, T. J., and Gregg, E. (2014). The Lifetime Cost of Diabetes and its Implications for Diabetes Prevention. *Diabetes Care*, 37(9):2557–2564.

Appendix

Robustness Checks for Section 1.5

Table A.1: Original Cohort RDDs, 1998-2004, Diabetic Women: Insulin and Oral Medication Usage

	(1)	(2)	(3)	(4)
	Mean 60-64			
Insulin				
	0.34	-0.51*	-0.44*	-0.47*
		(-2.55)	(-2.33)	(-2.49)
Oral Medication				
	0.64	0.13	0.15	0.16
		(0.62)	(0.78)	(0.82)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Column (2) reports estimated discontinuities from specifications without covariates apart from the age in months and age in months interacted with the treatment indicator; Column (3) reports results including time dummies; Column (4) reports results including both time dummies and health status, marital status, work status and education fixed effects. All specifications use local linear regression with age in months as the running variable and a Uniform kernel. Bandwidth used is 70.07, selected by the MSE criterion of Calonico et al. (2014a). Individuals enrolled in Medicaid at any age, or enrolled in supplemental insurance (Medigap) or a Medicare HMO (Medicare Advantage) after age 65, are excluded.

Table A.2: RDD Results with Age 62 as the Cutoff, 1998-2004, Diabetic Women: Insulin and Oral Medication Usage

	(1) Mean 60-64	(2)	(3)	(4)
Insulin				
	0.23	-0.53	-0.52	-0.38
		(-0.40)	(-0.39)	(-0.33)
Oral Medication				
	0.69	-0.44	-0.53	-0.56
		(-0.30)	(-0.36)	(-0.43)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Results treat 62 as the cutoff for the regression discontinuity design instead of 65 as in the rest of this chapter. Standard errors are clustered at the individual level. Column (2) reports estimated discontinuities from specifications without covariates apart from the age in months and age in months interacted with the treatment indicator; Column (3) reports results including time dummies; Column (4) reports results including both time dummies and health status, education fixed effects, and indicators for whether individuals are enrolled in a Medicare HMO (Medicare+Choice/Medicare Advantage, depending on whether pre- or post-2003) and whether they have purchased supplemental insurance (Medigap). All specifications use local linear regression with age in months as the running variable and a Uniform kernel.

Table A.3: RDD Results, 1998-2004, Continuously Insured Diabetic Women: Insulin and Oral Medication Usage

	(1)	(2)	(3)	(4)
	Mean 60-64			
Insulin				
	0.24	-0.02	-0.02	-0.02
		(-0.41)	(-0.44)	(-0.40)
Oral Medication				
	0.67	0.05	0.05	0.05
		(1.07)	(1.01)	(1.07)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Sample is restricted to individuals who are insured via either their or their spouse's employer. Standard errors are clustered at the individual level. Column (2) reports estimated discontinuities from specifications without covariates apart from the age in months and age in months interacted with the treatment indicator; Column (3) reports results including time dummies; Column (4) reports results including both time dummies and health status, education fixed effects, and indicators for whether individuals are enrolled in a Medicare HMO (Medicare+Choice/Medicare Advantage, depending on whether pre- or post-2003) and whether they have purchased supplemental insurance (Medigap). All specifications use local linear regression with age in months as the running variable and a Uniform kernel.

Table A.4: First-Differences Results, 1998-2004, Diabetic Women: Exercise and Diet

	(1)	(2)	(3)
△Exercise			
	0.06	0.06	0.07
	(1.17)	(1.18)	(0.89)
△Diet			
	-0.03	-0.03	-0.01
	(-0.61)	(-0.57)	(-0.08)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Column (1) reports estimated pre-post differences from specifications without covariates; Column (2) reports results including time dummies; Column (3) reports results including first-differences in earnings, work status, health status (an indicator equal to 1 if an individual reports being in "Fair" or "Poor" health, marital status, and enrolment in either Medicare Advantage/a Medicare HMO or supplementary coverage under Medicare (Medigap).

Table A.5: Unrestricted Panel RDD Results, 1998-2004 1998-2008, Diabetic Men, Insulin and Oral Medication Usage

	(1)	(2)	(3)	(4)	(5)	(6)
Insulin						
$\widehat{\beta}_1$	0.20	0.19	0.14	0.06	0.08	0.05
	(0.58)	(0.56)	(0.47)	(0.21)	(0.25)	(0.16)
$\widehat{\beta}_3$				0.35	0.38	0.21
				(0.74)	(0.79)	(0.48)
Cragg-Donald Stat.				10.98	10.79	13.07
Oral Medication						
$\widehat{\beta}_1$	-0.75	-0.73	-0.70	-0.55	-0.54	-0.56
	(-1.77)	(-1.75)	(-1.84)	(-1.49)	(-1.48)	(-1.62)
$\widehat{\beta}_3$				0.10	0.16	0.13
				(0.21)	(0.33)	(0.27)
Cragg-Donald Stat.				10.97	10.79	13.05

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications use local linear regression with age in months as the running variable and a Uniform kernel and exclude Medicaid recipients and smokers. Columns (1)-(3) are for the period 1998-2004; Columns (4)-(6) are for the period 1998-2008. Columns (1) and (4) report estimated pre-post differences from specifications without covariates; Columns (2) and (5) report results including time dummies; Columns (3)and (6) report results including first-differences in earnings, work status, health status (an indicator equal to 1 if an individual reports being in "Fair" or "Poor" health, marital status, and enrolment in either Medicare Advantage/a Medicare HMO or supplementary coverage under Medicare (Medigap).

Table A.6: First Differences Results, 1998-2004 1998-2008, Diabetic Men, Insulin and Oral Medication Usage

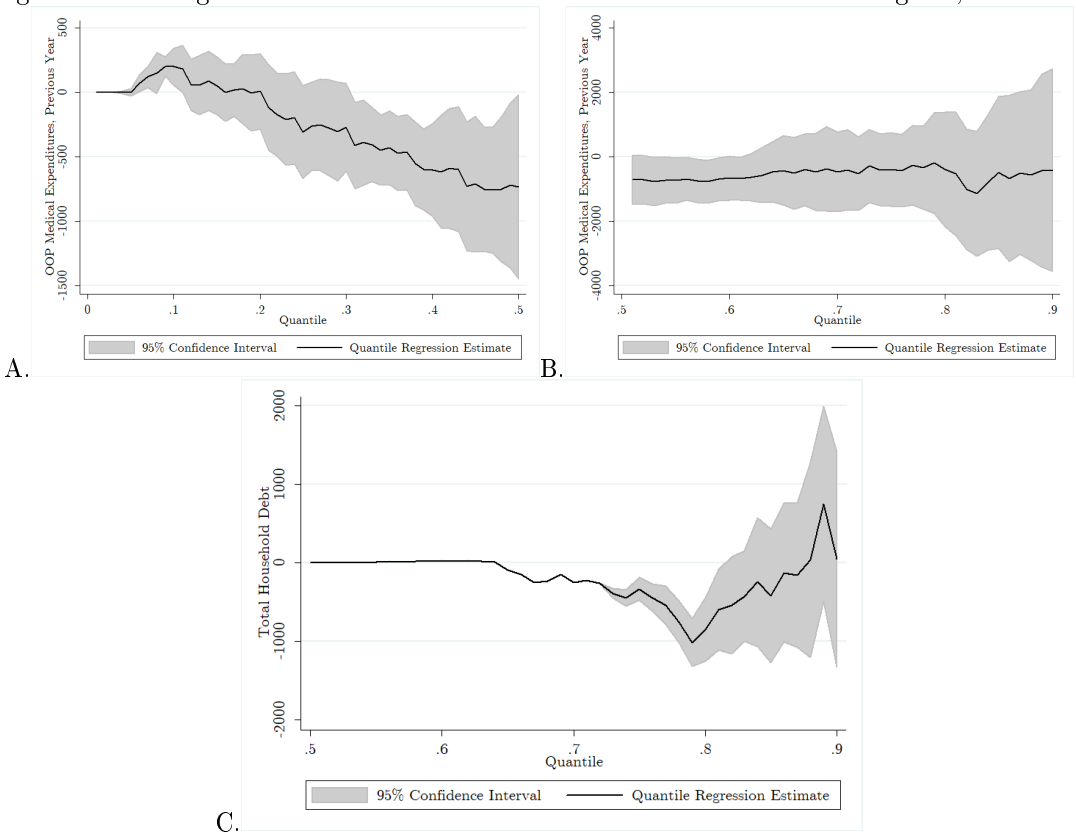
	(1)	(2)	(3)	(4)	(5)	(6)
△Insulin						
$\widehat{\beta}_1$	0.02	0.02	0.03	0.02	0.02	0.04
	(1.08)	(1.12)	(1.36)	(1.08)	(1.12)	(1.57)
$\widehat{\beta}_3$				-0.01	-0.01	-0.03
				(-0.25)	(-0.28)	(-0.92)
△Oral Medication						
$\widehat{\beta}_1$	0.00	0.00	0.01	0.00	0.00	0.01
	(0.15)	(0.16)	(0.18)	(0.15)	(0.16)	(0.48)
$\widehat{\beta}_3$				0.00	0.00	0.01
				(0.10)	(0.08)	(0.29)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

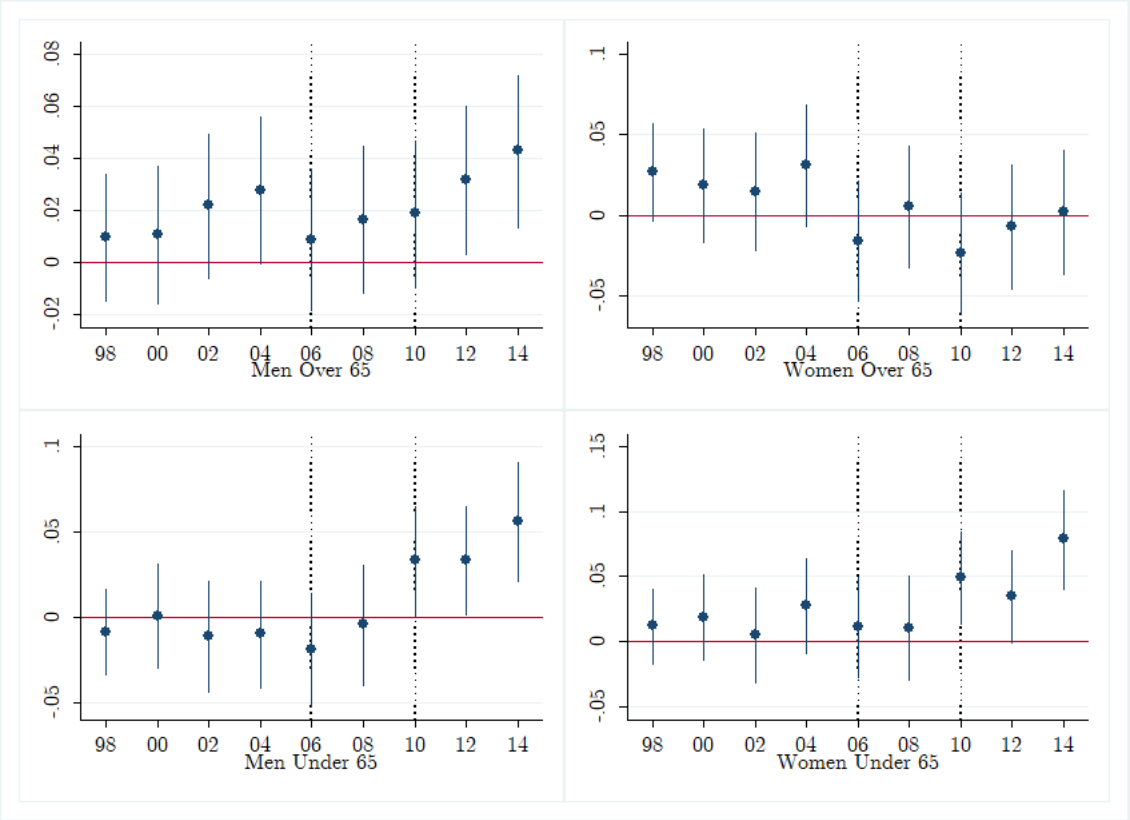
Notes: Standard errors are clustered at the individual level. All specifications use local linear regression with age in months as the running variable and a Uniform kernel and exclude Medicaid recipients and smokers. Columns (1)-(3) are for the period 1998-2004; Columns (4)-(6) are for the period 1998-2008. Columns (1)and (4) report estimated pre-post differences from specifications without covariates; Columns (2) and (5) report results including time dummies; Columns (3) and (6) report results including first-differences in earnings, work status, health status (an indicator equal to 1 if an individual reports being in "Fair" or "Poor" health, marital status, and enrolment in either Medicare Advantage/a Medicare HMO or supplementary coverage under Medicare (Medigap).

Figure A.1: Changes in Measures of Financial Risk for Female Diabetics at Age 65, 1998-2004



Notes: Figure A.1. displays point estimates and confidence intervals for quantile regression versions of the panel data RDD equation with the quantile τ varying over the set $\{0.01, 0.02, ..., 0.89, 0.90\}$. The dependent variable is out-of-pocket medical expenses from the preceding twelve months for Panels (A) and (B) and total household debt for Panel C.
All sharp regression-discontinuity results are obtained excluding those on Medicaid, Medicare Advantage or Medicare supplemental coverage (Medigap), use a Uniform kernel with local linear regression, and include time dummies, health status, and education fixed effects. The magnitude of the effects is generally smaller than those found in Barcellos and Jacobson (2015), likely because of the smaller sample size.

Figure A.2: Trends in Medicaid Enrolment Relative to 1996 Among Diabetics in the HRS, 1998-2014



Notes: Plots are of coefficients from pooled OLS regressions of the outcome (reporting receipt of Medicaid) on time dummies for the years 1998-2014, with 1996 as the reference category. Since there is overlap among individuals across the different waves of the survey, standard errors for these regressions are clustered at the individual level. Dotted lines correspond to major health care reforms - Medicare Part D being implemented in 2006, and the Affordable Care Act being passed in 2010. Medicaid recipients are excluded.

Table A.7: Medicaid Take-up Among Diabetics by Gender, Pre- and Post-2006

	Including 2010-14		Excluding 2010-14	
	(1)	(2)	(3)	(4)
	Over 65	Under 65	Over 65	Under 65
Female	0.133***	0.0901***	0.133***	0.0901***
	(11.38)	(7.34)	(11.38)	(7.34)
Post-2006	0.00765	0.0320**	-0.00349	-0.00606
	(1.04)	(3.20)	(-0.47)	(-0.52)
Female × Post-2006	-0.0358**	-0.00398	-0.0218	0.00373
	(-2.90)	(-0.26)	(-1.75)	(0.21)
Constant	0.0927***	0.0827***	0.0927***	0.0827***
	(13.87)	(11.47)	(13.87)	(11.47)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Dependent variable is an indicator for whether an individual reports receipt of Medicaid. First two columns include the years 2010-2014, which follow the passage of the Patient Protection and Affordable Care Act (ACA), while columns (3) and (4) restrict the Post-2006 observations to the 2006 and 2008 waves of the Health and Retirement Study.

Table A.8: Diagnosis of Diabetes by Gender, Pre- and Post-2006

	(1) Over 65	(2) Under 65	(3) Over 65	(4) Under 65
Female	-0.0316***	-0.0197***	-0.0461***	-0.0279***
	(-5.05)	(-3.44)	(-7.30)	(-4.95)
Post-2006	0.0821***	0.0596***	0.0804***	0.0563***
	(14.81)	(9.59)	(14.16)	(9.06)
Female × Post-2006	-0.00800	0.000928	-0.00803	-0.000878
	(-1.12)	(0.12)	(-1.10)	(-0.11)
Constant	0.195***	0.148***	0.191***	0.142***
	(40.14)	(33.32)	(38.46)	(32.02)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

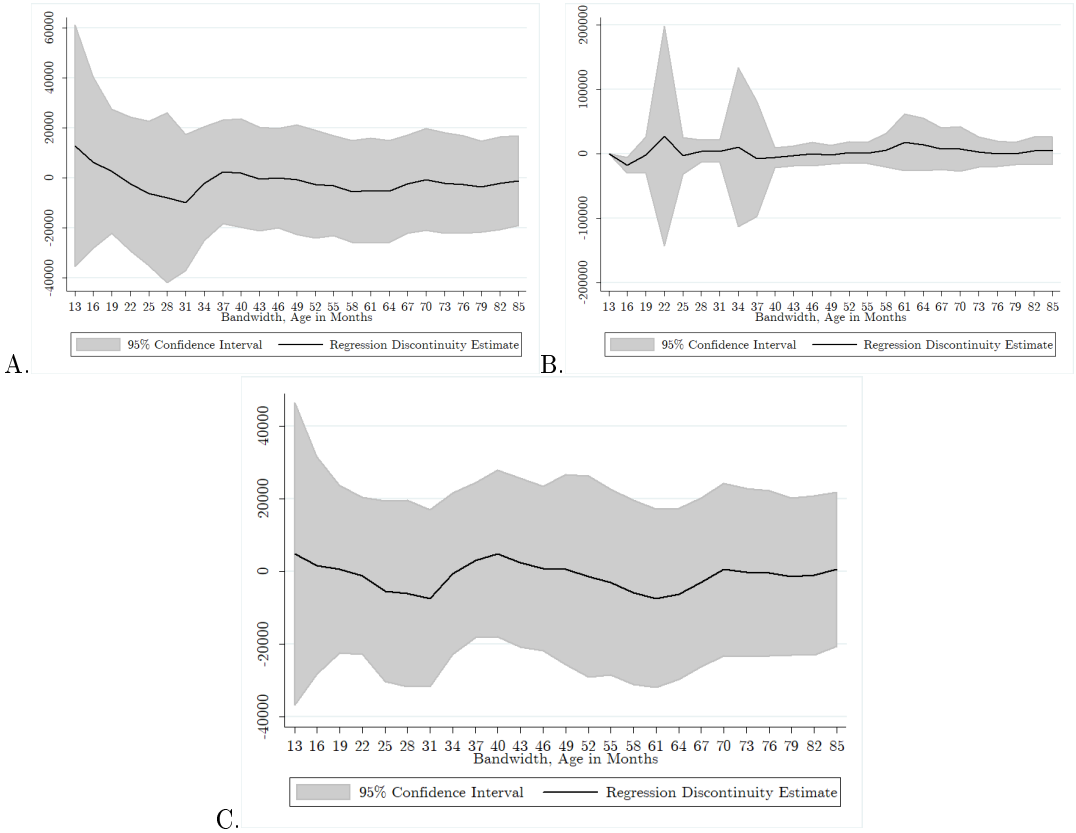
Notes: Standard errors are clustered at the individual level. Dependent variable is an indicator for whether an individual reports diagnosis of diabetes. First two columns pertain to the full sample; third and fourth columns restrict the analysis to the subsample that does not report being enrolled in Medicaid.

Evidence on (the Lack of) Credit Constraints

If individuals find it difficult to bring future income forward, they may struggle to exchange more prevention this year for less prevention next year. This would cast doubt on the power that intertemporal substitution has to explain the difference between this study and studies that do not use Medicare eligibility in an RDD. In those circumstances, the same agents would also find it difficult to bring income forward to purchase durable goods that can be financed on credit, and so there would be a spike in durable goods purchases as agents became eligible for Medicare.

I use the RAND CAMS dataset in this subsection. This is a cleaned version of the mail survey data run by the Health and Retirement Study (HRS) known as the Consumption and Activities Mail Survey (CAMS) (Hurd et al., 2015). I aggregate together total spending on durable goods, which comprises spending on refrigerators, washing machines, dishwashers, televisions, and computers, and add total spending on vehicles. I then replace insulin with total durable spending in a regression-discontinuity design using the years 1998-2004. I examine first the subsample of non-Medicaid-eligible diabetics who respond to the CAMS, then the subsample of that subsample that have nonpositive financial wealth and no housing wealth (hence no home equity with which to secure a loan), then the subsample that has positive housing wealth and so may exhibit "wealthy hand-to-mouth" behavior (as in Chetty (2008). There appears to be little evidence from these regressions that credit constraints are binding in the HRS data (see Figure A.3., below). This conclusion comes with the caveat that the overlap between the subset of HRS respondents who filled out the CAMS surveys and the subset who report being diabetic overlap very slightly, and so direct evidence on credit constraints among diabetics is difficult to obtain using these data.

Figure A.3: RD Estimates for Effect of Medicare Eligibility on Durable Goods Expenditures by Bandwidth, 1998-2004



Notes: All fuzzy regression-discontinuity results are obtained excluding those on Medicaid and use a uniform kernel with local linear regression. Standard errors are clustered within individuals. Sample consists of both male and female diabetics. Panel (A) corresponds to the broader subsample of diabetics who respond to the CAMS survey. Panel (B) corresponds to the subsample further restricted to individuals with nonpositive financial wealth and no housing wealth. Panel (C) follows Chetty (2008) and restricts the sample to individuals with positive housing wealth in case diabetics exhibit "wealthy hand-to-mouth" (Kaplan and Violante, 2014, Chetty and Szeidl, 2007) behavior. None of the specifications find significant increases in durable goods purchases at the age of Medicare eligibility.

Partial Identification Via Within-Individual Variation

In this section, I will exploit the fact that I have repeated observations on the same individuals in the data to compare the same individuals' insulin usage behavior before and after turning 65. I will use a combination of economic theory and institutional variation to show that a first-differences estimator that makes this within-individual comparison is biased upwards and towards zero by age effects, and that a corollary of this is that the regression-discontinuity estimator, which makes comparisons across individuals, is biased downward and away from zero (if at all).[16]

In this section, I first discuss the first-differences estimator and why it is likely to be biased upward (i.e. towards zero if the true effect is negative). Second, I discuss finite-sample biases in the standard regression-discontinuity design and why the arguments for the first-differences estimator being biased upward imply that the regression-discontinuity estimator is biased downward (i.e. away from zero if the true effect is negative). Third, I present simulations of the main theoretical mechanisms - differential mortality and shorter horizons towards the end of life - that provide further evidence for the arguments in the first two subsections. The fourth subsection presents the first-differences results and uses them to bound the effect of qualifying for Medicare coverage at age 65.

It will be useful to define the hypothetical outcomes Y_{it} for a given individual i in a period t when covered for treatment (denoted $D_{it} = 1$) and not covered (denoted $D_{it} = 0$), as

$$Y_{it} = \beta_0 + \delta X_{it} + \zeta t + \eta_i + v_{it} \text{ if } D_{it} = 0, \tag{1.15}$$

$$Y_{it} = (\beta_0 + \beta_1) + \delta X_{it} + \zeta t + \eta_i + v_{it} \text{ if } D_{it} = 1, \tag{1.16}$$

where X_{it} is a vector of covariates including age, R_{it}, t is a linear time trend, η_i is

[16] In a previous version of this chapter, I also estimated adjusted confidence intervals for the RDD results according to the method of Kolesár and Rothe (2018). The results (available on request) are robust to this modification for reasonable values of the tuning parameter required by their method. A key advantage of using the current method relative to both this method and "bias-corrected" estimators such as that of Calonico et al. (2014a) is that each of these either explicitly or implicitly makes some assumption regarding the curvature of the "true" conditional expectation function (CEF) (see Armstrong and Kolesár (2018b) for a proof). The method presented in this section eschews making any such assumption but forgoes point identification by doing so.

individual-specific time-invariant unobserved heterogeneity, v_{it} is an idiosyncratic shock in period t. From this we can derive the equation in levels

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \delta X_{it} + \zeta t + \eta_i + v_{it}, \tag{1.17}$$

and assuming individuals stay in the program once eligible, so that we can only have $\triangle D_{it} = 1$ or $\triangle D_{it} = 0$, the equations in first differences

$$\triangle Y_{it} = (\beta_1 + \zeta) + \delta \triangle X_{it} + \triangle v_{it} \text{ if } \triangle D_{it} = 1, \tag{1.18}$$

$$\triangle Y_{it} = \zeta + \delta \triangle X_{it} + \triangle v_{it} \text{ if } \triangle D_{it} = 0, \tag{1.19}$$

from which we can derive the equation

$$\triangle Y_{it} = \zeta + \beta_1 \triangle D_{it} + \delta \triangle X_{it} + \triangle v_{it}. \tag{1.20}$$

The assumption that allows both the equation in levels and the equation in differences to recover the same underlying effect is that the true difference in behavior between the covered and uncovered states is a fixed proportion. This assumption comes is supported by the fact that the change in agents' constraints is anticipated in advance, so that there is no shift of the lifetime participation profile either side of the age threshold of interest (cf. Section 1.6 and/or MaCurdy (1981)), and so the movement along that profile recovered by the first-differences estimator is the same as the movement along the participation profile measured using the cross-sectional differences between individuals either side of the age threshold.

Positive Bias in the First-Differences Estimator

In this subsection, I will show that the first-differences estimator in my setting is biased upward and towards zero for a negative effect of interest. This will be for two reasons. First, the "treatment" group of those who switch status to being covered by Medicare at age 65 is younger than the

"control" group of those who are continuously insured by Medicare in both adjacent periods, who are 67 or older since the Health and Retirement Study surveys are taken every two years. Second, given that the "control" group is older, they are less likely to begin using insulin, and more likely to stop using insulin if they are already using it. This will be because a greater proportion of the older group will be approaching the end of life, and so will have progressively smaller returns to investing in their health as they approach the end of their investment horizon (compare the model in Grossman (1972)). In the next subsection, I will show that both this argument and an argument from differential mortality will imply that the regression-discontinuity estimator is biased in the opposite direction, downward and away from zero.

Since I have repeated observations of the same individuals, I can estimate Equation 1.20, with an additional sample selection criterion that will buttress the arguments in this section,

$$\triangle Y_{it} = \zeta + \beta_1 \triangle D_{it} + \delta \triangle X_{it} + \triangle v_{it} \text{ if } \triangle D_{it} = 0 \implies R_{it} \geq \bar{R},$$

where $\triangle D_{it}$ is equal to 1 if individual i qualified for coverage under Medicare in period t (and only 0 otherwise, since no one who becomes eligible ages out of the program), $\triangle X_{it}$ is a vector of first-differenced time-varying individual characteristics, $\triangle v_{it}$ is an idiosyncratic error term, and the condition $\triangle D_{it} = 0 \implies R_{it} \geq \bar{R}$ implies that the only individuals with $\triangle D_{it} = 0$ are continuously insured by Medicare at ages 67 or older (where R_{it}, age, exceeds the age of eligibility, $\bar{R}$). The effect of qualifying for coverage on the change in the probability of using insulin between the uncovered and the covered period is β_1, which is identified if and only if

$$E[\triangle v_{it} | \triangle D_{it}] = 0 \text{ (Strict exogeneity)}, \tag{1.21}$$

We can obtain relatively tight bounds on the effect of qualifying for Medicare by restricting the sample so that violations of strict exogeneity are likely to create bias in the opposite direction of finite-sample bias for the regression-discontinuity estimator. If we restrict the sample to individuals who have just aged into the program (who have $\triangle D_{it} = 1$) and individuals who are older than them who are continuously enrolled in the program (so that $\triangle D_{it} = 0$), then the "treatment group" is younger than the "control group". On the other hand,

regression-discontinuity designs will compare individuals for whom $D_{it} = 1$ with individuals younger than them with $D_{it} = 0$ so that the "treatment group" is older than the "control group". It is therefore likely that if we can establish the sign of the bias for one of these estimators, the bias for the other estimator will be of the opposite sign.[17]

Suppose that the true conditional expectation functions (CEF) for the trends in insulin usage $\triangle Y_{it}$ among those who switch status ($\triangle D_{it} = 1$) and those who do not ($\triangle D_{it} = 0$) is given by (omitting covariates for simplicity):

$$E[\triangle Y_{it} | \triangle D_{it} = 0] = \zeta, \tag{1.22}$$

$$E[\triangle Y_{it} | \triangle D_{it} = 1] = \zeta + \beta_1, \tag{1.23}$$

then

$$\beta_1 = E[\triangle Y_{it} | \triangle D_{it} = 1] - E[\triangle Y_{it} | \triangle D_{it} = 0]; \tag{1.24}$$

Now consider that $\triangle Y_{it} \in \{-1, 0, 1\}$ for all i, t. Let n^T be the number of individuals for which $\triangle D_{it} = 1$, n^C the number of individuals for which $\triangle D_{it} = 0$, n_k^T, n_k^C the number of individuals with $\triangle Y_{it} = k$ for $\triangle D_{it} = 1$ and $\triangle D_{it} = 0$ respectively. Then it is easy to show that

$$\beta_1 = \left(\frac{n_1^T - n_{-1}^T}{n^T} \right) - \left(\frac{n_1^C - n_{-1}^C}{n^C} \right), \tag{1.25}$$

and the sample counterpart of β_1, $\widehat{\beta}_1$, is obtained by replacing n^T, n^C, n_k^T, n_k^C with their sample counterparts (for each value of k in the latter two cases). Hence any discrepancy between the sample counts of these groups and the counts that would be obtained from an experiment with true random sampling will lead to bias in the first-differences estimator. Intuitively, the first-differences estimator compares the numbers of those who initiate insulin therapy between

[17] Note that it is not logically impossible for the biases to be of the same sign. If decisions exhibit negative serial correlation, then higher probabilities of an outcome in a given period also produce higher probabilities of reversal in the subsequent period. In this case, equations in levels and equations in differences may still exhibit opposite biases if the treatment and control groups are the same for both. The evidence from both the medical literature (Brown et al., 1999, Rajagopalan et al., 2003) and the Appendix, where I estimate bounds for the effect of past decisions on present decisions, indicates that decisions are positively serially correlated in this context.

ages 65-67 after reporting not using it in the last wave with the number of those who cease using insulin at ages 65-67 after having used it in the last wave of the survey. This net initiations minus cessations figure is then compared with the same figure for the "control" group, the rest of the sample. If the control groups figure for initiations - cessations is not a consistent estimator for that of the treatment group's counterfactual, strict exogeneity is violated and the estimator is biased overall.

If the bias is unambiguously signed, this will be because of an unambiguously signed difference in the number of individuals at the margin of switching their insulin usage status for whom $\triangle D_{it} = 1$ and those for whom $\triangle D_{it} = 0$. If those with $\triangle D_{it} = 0$ do not have the same distribution of unobserved heterogeneity as those who have just qualified and have $\triangle D_{it} = 1$, then they are a control group that differs systematically from the treatment group for reasons that are not related to being treated (provided with coverage). Rewriting β_1 as the difference in means between groups, we obtain

$$\beta_1 = \left(\frac{n_1^T}{n^T} - \frac{n_1^C}{n^C}\right) - \left(\frac{n_{-1}^T}{n^T} - \frac{n_{-1}^C}{n^C}\right), \tag{1.26}$$

so that the bias is unambiguously positive if and only if we underestimate the proportion of individuals who would have begun insulin in the absence of qualifying for Medicare $\left(\frac{n_1^C}{n^C}\right)$ and overestimate the number of individuals who would have stopped insulin therapy in the same counterfactual $\left(\frac{n_{-1}^C}{n^C}\right)$. In this case $\widehat{\beta}_1$ will be an underestimate of β_1 if $\beta_1 < 0$ and an overestimate if $\beta_1 > 0$. This would imply that the control group for whom $\triangle D_{it} = 0$ has a weaker trend towards insulin usage than in the true counterfactual, which therefore leads us to underestimate the strength of the departure from trend upon qualifying for Medicare.

I argue that since those with $\triangle D_{it} = 0$ are older than 67 by construction, these conditions are likely to hold.

Individuals who have already begun insulin therapy will have weaker incentives to continue using insulin over time as they approach the end of life. In the limit, total certainty about not surviving to the next period reduces the return to investing in health to zero (Grossman, 1972).[18] The difference

[18] This effect of expected lifespan on investment in health is termed the "Mickey Mantle effect" in Fang et al. (2007) and Khwaja (2009).

between those with $\triangle D_{it} = 0$ and a randomly assigned control group is that the first-differences estimator's "control" group has a spuriously negative trend (or spuriously weak positive trend) and this will lead to an underestimate of the true counterfactual trend.[19] I will assume that the effects of insulin on lengthening life are second-order (as evidence in the literature seems to suggest (Weinger and Beverly, 2010)). Thus the first-order effect of aging within insulin adherents will be to reduce the return to using insulin. This is because as individuals approach the end of life, the expected return from using insulin in the future decreases (Grossman, 1972, 2000).[20]

It may be argued that insurance for treatment also reduces mortality, and so increases individuals' horizons and hence their propensity to invest in their health. There are two problems with this argument. First, since Medicare is a long-expected change rather than an unexpected change, it will already have been built into individuals' subjective mortality expectations, and hence will not increase life *expectancy per se*. Second, the effects of insurance on mortality in the literature tend to be small. The largest such effects are found within a small window of Medicare eligbility by Card et al. (2009), and that paper finds a reduction in one-week mortality of between 0.8 and 1 percentage points for a subset of 12% of their sample. It seems unlikely that this extends individuals' horizons more generally to the extent that it has a first-order effect on their decisions regarding investment in health.

I formalize these ideas in the following Proposition:

Proposition 1. Suppose that $\left(\frac{n_1^T - n_{-1}^T}{n^T}\right)$ is estimated consistently by its sample counterpart but that

$$\frac{\widehat{n}_1^C}{\widehat{n}^C} < \frac{n_1^C}{n^C} \text{ and } \frac{\widehat{n}_{-1}^C}{\widehat{n}^C} > \frac{n_{-1}^C}{n^C},$$

then $\widehat{\beta}_1 > \beta_1$, implying $E[\triangle v_{it} | \triangle D_{it} = 0] < E[\triangle v_{it} | \triangle D_{it} = 1]$, hence that the first-differences

[19] Results from first-differences regressions that include individuals younger than 65 with $\triangle D_{it} = 0$ are available on request from the author. In line with the arguments in this subsection, the results are more negative than in the regressions that restrict the sample to those who change status and older individuals who have $\triangle D_{it} = 1$.

[20] In fact, we require that the net present value of investing in health decreases at an increasing rate towards the end of a finite lifespan. This is true under relatively mild conditions in human capital models (Grossman, 2000), and follows from the fact that investment in a single period raises the stock of human capital in every subsequent period. If there is only one period in which to reap a return on investment in health capital, this is larger than the return of zero from investment in health in the last period of life, and smaller than the return from investment with two periods left, and so forth. A proof of this argument (Proposition 3) can be found at the end of the Appendix.

estimator $\widehat{\beta}_1$ is biased upward due to a failure of strict exogeneity.

Proof: Write

$$\begin{aligned}\widehat{\beta}_1 &= \left(\frac{\hat{n}_1^T - \hat{n}_{-1}^T}{\hat{n}^T}\right) - \left(\frac{\hat{n}_1^C - \hat{n}_{-1}^C}{\hat{n}^C}\right) \\ &= \left\{\left(\frac{n_1^T - n_{-1}^T}{n^T}\right) - \left(\frac{n_1^C - n_{-1}^C}{n^C}\right)\right\} + \left\{\left(\frac{n_1^C}{n^C} - \frac{\hat{n}_1^C}{\hat{n}^C}\right) - \left(\frac{n_{-1}^C}{n^C} - \frac{\hat{n}_{-1}^C}{\hat{n}^C}\right)\right\} \\ &> \left(\frac{n_1^T - n_{-1}^T}{n^T}\right) - \left(\frac{n_1^C - n_{-1}^C}{n^C}\right) = \beta_1,\end{aligned}$$

where the last line follows from the assumptions that $\frac{\widehat{n}_1^C}{\widehat{n}^C} < \frac{n_1^C}{n^C}$ and $\frac{\widehat{n}_{-1}^C}{\widehat{n}^C} > \frac{n_{-1}^C}{n^C}$,. Hence $E[\widehat{\beta}_1 - \beta_1] = E[\triangle v_{it} | \triangle D_{it} = 1] - E[\triangle v_{it} | \triangle D_{it} = 0] > 0$, so the bias in the estimators of $\frac{n_1^C}{n^C}$ and $\frac{n_{-1}^C}{n^C}$ can be explained as a violation of strict exogeneity■.

Hence selection on unobservables works to bias the first-differences estimator towards zero. Intuitively, the counterfactual trend towards insulin usage is underestimated for the treatment group that qualifies for coverage at age 65. I exploit this in the rest of the chapter in order to bound the effect of qualifying for Medicare coverage between the first-differences results and the cross-sectional RDD results. The reason for the difference in the sign of the bias for the two methods, given the same institutional variation, is that the "treatment" group in the RDD regression is older than the "control" group (right-hand side vs. left-hand side of the cutoff age of 65), whereas the "treatment" group in the first-differences regressions (those who switch status, aged 65-66) is younger than the "control" group (individuals continuously insured by Medicare, 67 or older).

Negative Finite-Sample Bias in the Regression Discontinuity Estimator

In practice regression-discontinuity designs rely on comparing average outcomes for individuals "close" to the cutoff, within some bandwidth h. A recent literature has focused on how best to choose this bandwidth so as to optimally trade off finite-sample bias and inefficiency in the resulting estimator (Calonico et al., 2014a, Imbens and Kalyanaraman, 2012, Armstrong and

Kolesár, 2017). If one focuses on individuals too close to the cutoff, one risks having too few individuals to have sufficient power to reject the null hypothesis, especially with a discrete running variable which has only so many individuals at each integer value (Kolesár and Rothe, 2018). Examine individuals too far away from the cutoff, and there is a greater risk that non-comparability due to unobserved traits drives differences in average outcomes rather than the discontinuity of interest.

Suppose we have finite-sample bias in the regression discontinuity design. This will result in estimators biased downward and away from zero in this context. Since the running variable is age, individuals far enough to the right of the cutoff in a given cross-section will have had to survive for longer to be observed than those far to the left of the cutoff. It has been known for some time that estimating life-cycle profiles with cross-sectional data will be subject to bias from "differential mortality" - those who survived had characteristics that produce spurious patterns in the data that are likely to differ from the actual life-cycle profiles we could construct if we could observe the same individuals' actions over time (Attanasio and Hoynes, 2000, De Nardi et al., 2010). In the current context, healthier diabetics are more likely to survive and less likely to use insulin (since it is a preventive measure "of last resort"). If we could see which diabetics ought to be using insulin as a subset of diabetics (which I cannot see in the data), we could distinguish non-users from non-adherents. Since we cannot do this, non-users are more likely to be healthier and to survive for longer than insulin users. (*Among the subset of diabetics who are recommended insulin*, by contrast, those who adhere to insulin will have fewer health complications than non-adherents). Hence as we move to the right of the cutoff age of 65, we encounter individuals who are less and less likely to benefit from using insulin due to the relative mildness of their disease, and so we will spuriously estimate a lower average for insulin usage among this group, biasing the estimator downward and away from zero.

Moreover, the argument from the previous subsection regarding individuals who are closer to the end of life being less likely to start or continue using insulin applies to the proportion using insulin as well as the trend. The latter was used to argue that the bias of the first-differences estimator was upward and toward zero, but here leads to bias in the opposite direction, which reinforces the effect

of differential mortality on the negative bias of the estimator. The reason for this is that in the case of the first-differences estimator, the "treatment" group are individuals aged 65-66 who have switched status, and the "control" group comprises older individuals who are continuously insured by Medicare. By contrast, the regression-discontinuity estimator, which examines outcomes in levels, compares a "treatment" group of individuals older than 65 with a "control" group of individuals younger than 65. As a result, the same argument for why the first-differences estimator is biased towards zero is also an argument for why the RDD estimator is biased away from zero, since the two estimators have the opposite ranking for whether the "treatment" or the "control" group is older.

With a discrete running variable, we are caught between an RDD estimator with large efficiency losses due to the small sample size when restricting attention to observations close to the cutoff, and an RDD estimator biased downward and away from zero when using a wider bandwidth. By contrast, if the estimator based on variation within individuals over time is biased upward and towards zero (as argued above), and still gives statistically significant results, then we can bound the size of the effect of gaining coverage at age 65.

In sum: older individuals need to survive to be observed; this will bias the levels of their insulin usage towards zero, hence the effect of Medicare away from zero, since being healthier they will have lower usage on average. Hence I will assume that the finite-sample bias of the RDD is such that

$$E[\eta_i | R_{it} \geq \bar{R}] < E[\eta_i | R_{it} < \bar{R}], \tag{1.27}$$

In the following subsection, I separately simulate the effects of the two mechanisms cited in this and the previous subsection (differential mortality and shrinking horizons towards the end of life). It turns out that the biases in the two estimators have the expected signs argued for in this and the previous subsection.

Monte Carlo Evidence: Biases in First-Differences and Regression Discontinuity Estimators

This subsection simulates the effects of differential mortality and approaching the end of life on the regression discontinuity and first-differences estimators. The effects on the biases of the estimators are as they should be according to the arguments in the preceding two subsections.

I simulate data from 1000 draws for each bandwidth in the set $\{1, 2, 3, ..., 10\}$ of a Standard Normal random variable labeled $\varepsilon_i \sim N(0, 1)$. This is used to draw the initial distribution of disutilities of usage for a cohort of 50 individuals, who start all at the same age. Higher values of ε_i will correspond to higher disutilities of usage and therefore lower likelihoods of usage. I then generate a panel of 21 time periods for these individuals with ε_i constant in each period, but with $R_{it} = t - 11$ - so that the "age" running variable is normalized to 0;

$$Y_{it} = 1[\varepsilon_i - \frac{R_{it}}{20} \leq 1.645] \text{ if } R_{it} < 0 \text{ (equiv. } t < 10); \tag{1.28}$$

$$Y_{it} = 1[\varepsilon_i - \frac{R_{it}}{20} \leq 0] \text{ if } R_{it} \geq 0 \text{ (equiv. } t \geq 10); \tag{1.29}$$

so we have an outcome that shifts in expectation by 0.45 at the threshold value of $R_{it} = 0$ (since 1.645 is the 95^{th} percentile of the standard Normal distribution). The first set of simulations use this set-up to estimate the sharp regression discontinuity design with a local linear regression

$$Y_{it} = \beta_0 + \beta_1 1[R_{it} \geq 0] + \beta_2 R_{it} + \beta_3 R_{it} \times 1[R_{it} \geq 0] + \eta_i + v_{it} \text{ if } |R_{it}| < h \tag{1.30}$$

where $h \in \{1, 2, 3, ..., 10\}$, η_i is time-invariant unobserved individual-level heterogeneity due to unobserved permanent differences in the disutilities of usage, and v_{it} is functional form misspecification error, which results from estimating a linear probability model when the true model is a Probit.

In the second specification, I introduce differential mortality by ε_i, so that individuals with lower disutilities of usage (lower ε_i) are selected out of the sample to the right of the cutoff of $R_{it} = 0$. To ensure that the selection process does not affect the true effect of crossing the threshold of

$R_{it} = 0$, I restrict the process to individuals who have $R_{it} \geq 1$. "Deaths" are drawn according to the Accelerated Failure Time model, where T_i is the last period of life for individual i,

$$lnT_i = \theta + \varepsilon_i; \tag{1.31}$$

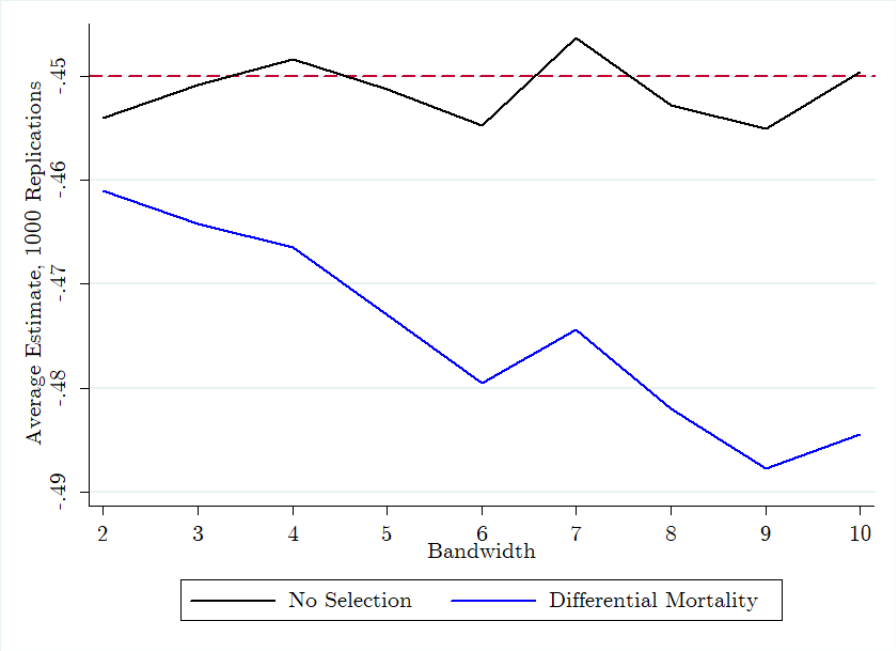
I set $\theta = 3.5$. The results are displayed in Figure A.4. I do not display confidence intervals since I have enough observations in the simulations for relatively small confidence intervals even at smaller bandwidths, thus abstracting away from the efficiency problems that arise in this chapter. The local linear estimates without selection are on average the value of the true change in probabilities at the cutoff. When individuals with lower disutilities of usage get selected out of the sample, the RDD estimator is progressively more negatively biased the wider the bandwidth used. Note that the magnitude of the biases depends not only on the rapidity of differential selection but also the distribution of reservation utilities. A distribution of reservation utilities that is more densely concentrated so that a minority of individuals accounts for the majority of cases where $Y_{it} = 1$ will result in larger biases if that minority attrites from the sample than in the case where reservation utilities are Uniformly distributed. Figure A.5 shows the empirical counterpart of Figure A.4. As expected, estimates produced with observations from within 24 months of the cutoff are closer to zero than those obtained with larger bandwidths. One difference between the figures is that the estimates from the HRS data are asymptotic to roughly -0.5, which is not predicted by the simpler framework used in the Monte Carlo experiments.

The second mechanism is that while diabetics will be more likely to require insulin over time, as they approach the end of life they will also have shorter horizons and hence lower returns from using insulin. In order to examine the effect of this mechanism on first-differences and regression-discontinuity estimates, I change the latent index model equations to

$$Y_{it} = 1[\varepsilon_i - \frac{R_{it}}{20} + \frac{R_{it}^2}{10} \leq 1.645] \text{ if } R_{it} < 0; \tag{1.32}$$

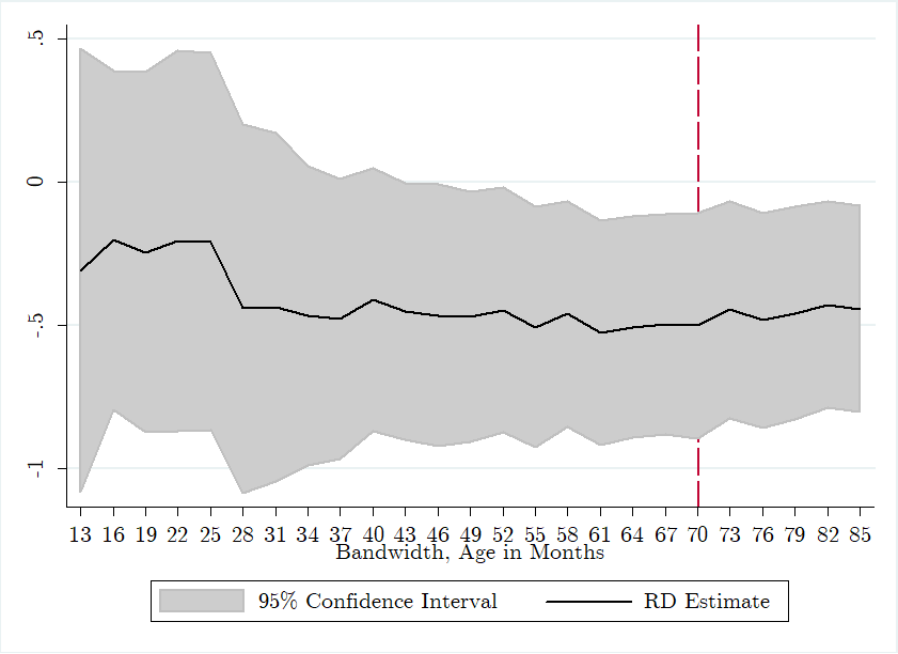
$$Y_{it} = 1[\varepsilon_i - \frac{R_{it}}{20} + \frac{R_{it}^2}{10} \leq 0] \text{ if } R_{it} \geq 0; \tag{1.33}$$

Figure A.4: Simulated Effects of Differential Mortality on Bias in RDD Estimates



Notes: Lines connect average estimates of $\widehat{\beta}_1$ for Equation 1.30 where $\beta_1 = -0.45$ for 1000 simulated processes of aging and extensive-margin usage under differential sample selection. Individuals with higher disutilities of usage are more likely to stay in the sample. Both lines correspond to a local linear regression with a Uniform kernel, using the bandwidth displayed on the x-axis, but the blue line corresponds to simulations where differential sample selection is present and the black line to simulations where it is absent.

Figure A.5: RD Estimates for Effect of Medicare Eligibility on Insulin Usage by Bandwidth, 1998-2004 - Original 1998 Cohort



Notes: All fuzzy regression-discontinuity results are obtained excluding those on Medicaid, Medicare Advantage or Medicare supplemental coverage (Medigap), use a Uniform kernel with local linear regression, and include time dummies, health status, and education fixed effects. Standard errors are clustered within individuals. Red dashed line corresponds to the optimal bandwidth of 70.07.

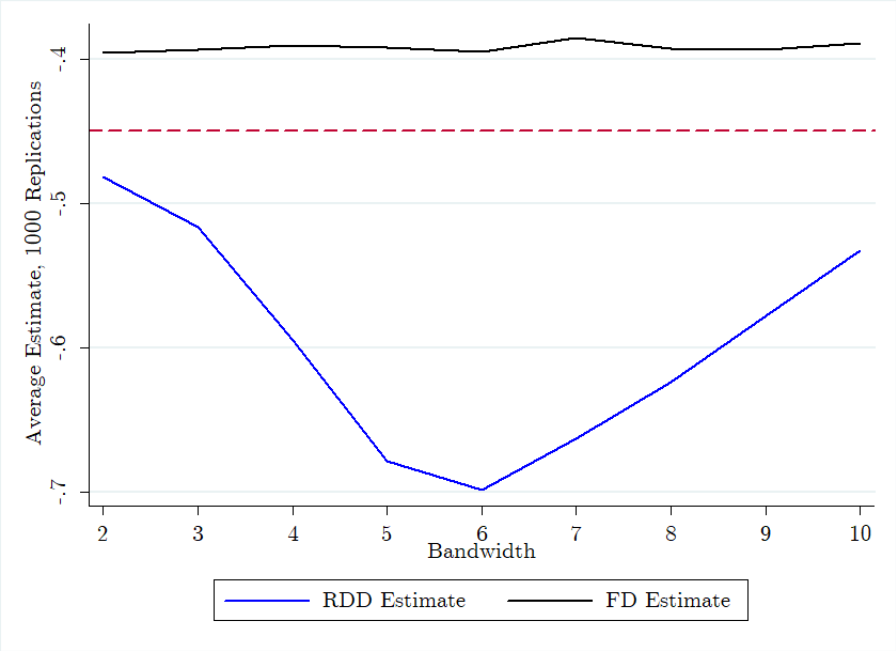
In order to examine the effect of this mechanism in isolation, I no longer make use of differential sample selection according to individual-specific mortality. I also introduce results from estimating the first-differences specification

$$\triangle Y_{it} = \beta_1 \triangle 1[R_{it} \geq 0] + \beta_2 \triangle R_{it} + \beta_3 \triangle R_{it} \times 1[R_{it} \geq 0] + \triangle v_{it}, \tag{1.34}$$

while restricting the sample to $R_{it} \geq -1$. Note that this condition guarantees that the idiosyncratic error $\triangle v_{it}$ is different for individuals who are in the "treatment" group versus the "control" group, since the first difference of some i.i.d. functional form misspecification error will differ for observations around the cutoff of $R_{it} = 0$ and the set of observations for which $0 < R_{it} \leq 10$ (compare Figure A.5). The results are displayed in Figure A.6. As argued in the preceding subsections, the first-differences estimator is biased upward and towards zero. The RDD estimator is biased downward and away from zero over the range of the running variable considered in the simulations, but the bias appears to decrease for bandwidths wider than $h = 6$. This shows that the relative strength of differential mortality and the within-person horizon effect matter for the extent of the bias in the RDD estimator (and potentially whether it is even biased downward at all). The reason the bias starts decreasing at higher bandwidths when there is no differential mortality is that if the "hump-shaped" life-cycle profile of usage is symmetric about the cutoff, then widening the bandwidth will result in the inclusion of a greater number of zeroes to the left as well as the right of the cutoff (see Figure A.7). For the purposes of this chapter it seems reasonable to assume that the pattern seen to the left of $h = 6$ in Figure A.6 holds over the range of bandwidths considered, since the combined effects of differential mortality and shorter horizons to the right of the cutoff are unlikely to be matched by similarly rapid changes at younger ages.

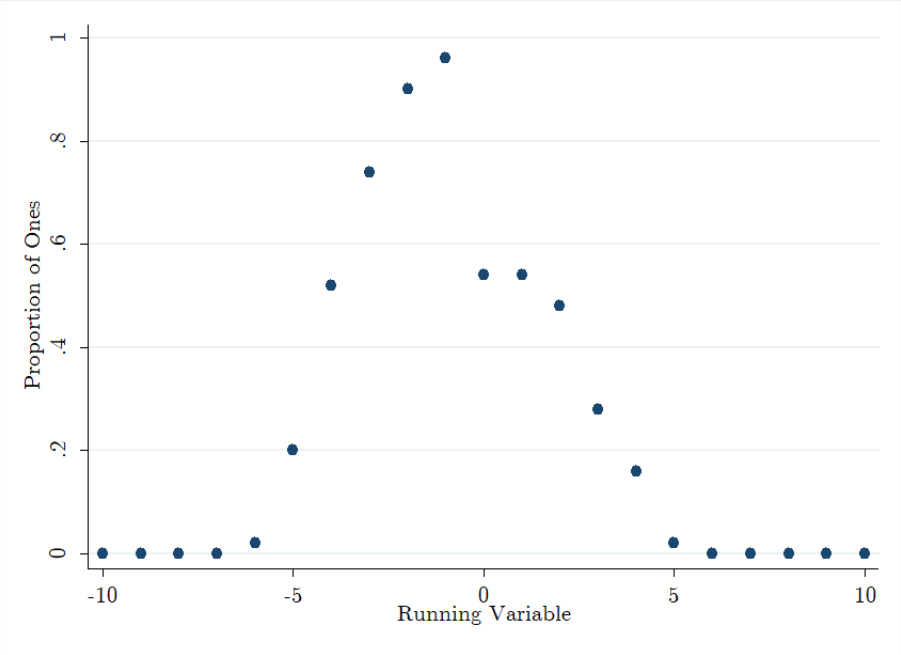
In the next subsection, I estimate first-differences equations to provide a lower bound for the effect of qualifying for Medicare coverage.

Figure A.6: Simulated Effects of Shrinking Horizons on Bias in RDD and FD Estimates



Notes: Lines connect average estimates of $\widehat{\beta}_1$ for Equations 1.30 (blue) and 1.34 (black) where $\beta_1 = -0.45$ for 1000 simulated processes of aging and extensive-margin usage. Both lines correspond to a local linear regression with a Uniform kernel, using the bandwidth displayed on the x-axis, but the true conditional expectation function (CEF) is quadratic since within a given individual, the disutility of usage first rises, and then falls with "age". The blue line corresponds to simulations of a local linear regression with a Uniform kernel and the bandwidth listed on the x-axis. The black line corresponds to simulations that use the same simulated data to estimate a model in first-differences which only includes data for $R_{it} \geq -1$ that would recover the same underlying population parameter as the regression-discontinuity estimates if both estimators were unbiased.

Figure A.7: RDD Graph, One Draw of Simulated Shrinking Horizons, Quadratic Linear Index Model



Notes: The figure represents binned data by simulated "age" from a simulated model where the disutility of "usage" first rises, then falls with "age". The discontinuity at the centered value of the "age" variable is -0.45 by construction. The true conditional expectation function (CEF), as can be seen in the figure, is quadratic.

First-Differences Results

In this subsection, I present the empirical results from first-differences regressions that allow me to obtain partial identification even if point identification fails. The estimating equation becomes the empirical counterpart of Equation 1.34, augmented to allow pre- and post-2006 effects to differ,

$$\triangle Y_{it} = \zeta + \beta_1 \triangle 1[R_{it} \geq \bar{R}] + \beta_2 1[t \geq 2006] + \beta_3 \triangle 1[R_{it} \geq \bar{R}] \times 1[t \geq 2006] + \delta \triangle X_{it} + \triangle v_{it}, \quad (1.35)$$

The sample is restricted to those 63 and over to ensure that the "control group" of those continuously insured by Medicare is older than the "treatment group" of those who change status at ages 65-

Table A.9: First-Differences Results, Diabetic Women: Insulin and Oral Medication Usage

	(1)	(2)	(3)
△Insulin			
$\widehat{\beta}_1$	-0.04*	-0.04*	-0.06*
	(-2.26)	(-2.27)	(-2.23)
$\widehat{\beta}_3$	0.08**	0.08**	0.10**
	(2.99)	(2.92)	(2.93)
△Oral Medication			
$\widehat{\beta}_1$	0.04	0.04	0.05
	(1.44)	(1.39)	(1.34)
$\widehat{\beta}_3$	0.01	0.01	-0.01
	(0.34)	(0.29)	(-0.27)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. Sample is restricted to individuals 63 and older. Column (1) reports estimated pre-post differences from specifications without covariates; Column (2) reports results including time dummies; Column (3) reports results including first-differences in work status, health status (an indicator equal to 1 if an individual reports being in "Fair" or "Poor" health, marital status, and enrolment in either Medicare Advantage/a Medicare HMO or supplementary coverage under Medicare (Medigap).

66 (since the HRS survey is taken every two years). $\triangle X_{it}$ includes changes in health status, earnings, work status, marital status, and choice of supplemental insurance (Medigap) or choice of Medicare+Choice/Medicare Advantage/Medicare HMO. The results are displayed in Table A.9. As expected, the estimated effects are smaller than those estimated using a regression-discontinuity design, but are nonetheless statistically significant. Common to both sets of results is the failure to find significant changes in usage of oral medication. The smallest among these results implies a 4.4 percentage point decline following qualifying for Medicare (note that the first-differences results are on a different scale compared to the fuzzy RD results, since they make direct comparisons within

individuals before and after qualifying for Medicare and so control flexibly for other changes in available insurance at 65 as well). If we are especially conservative and use the average usage of 34 percentage points between ages 60 and 64 from the original 1998 cohort (see Appendix), then this translates into a relative difference of 12.9% between those just under and just over 65 in insulin usage. This produces the lower bound quoted in the Introduction. Reassuringly, the effect of Part D is similar to the one reached from the RDD results, and inspection appears to indicate that the magnitude of the relative offset is also similar. Formally, a Wald test cannot reject the null hypotheses that $\beta_1 = -\beta_3$ or that $\beta_1 = -2\beta_3$ in this context either.

To summarize: this section has provided a method for obtaining bounds on the effect of qualifying for Medicare coverage on insulin usage. Implementing this method results in a lower bound of a 4.4 percentage point decrease pre-Part D, and an upper bound of a 7.9 percentage point decrease. The relative offset of this negative effect due to Part D subsidies for insulin post-2006 is similar in magnitude in both cases.

Discrete Investments in Health and the First-Differences Estimator

In this section, I will show that the probability of making a discrete investment in human capital falls at an increasing rate towards the end of a finite lifetime, and that this implies that a first-differences estimator that compares an older "control" group with a younger "treatment" group can be expected to have weaker positive trends (or stronger negative trends) in health investment in the older control group due to unobservable differences in the net returns to investment in health. Throughout this subsection I shall assume there are no differences in long-run endowments of health or lifespan (which are differenced out in a first-differences specification), no heterogeneity in the net return to investment in health apart from that induced by distance from the end of life, and nondecreasing costs to forgoing investment in health with age.

Proposition 3. Suppose we have, as in Grossman (2000), a law of motion for health capital, with H_t the stock of health capital in period t, $\delta_t \in (0,1)$ the rate of depreciation between t and $t+1$, and I_t gross investment in health capital in period t,

$H_{t+1} - H_t = I_t - \delta_t H_t$,

with investment occuring over a finite lifespan so that $t = 1, ..., T$. Then the net present value of the return on gross investment in health capital in period t, defined

$NPV_t = \sum_{s=t}^{T-1} \frac{\partial H_{s+1}}{\partial I_t}$,

falls at an increasing rate as $t \to T$.

Proof. It is easy to show that $NPV_t = 1 + \sum_{r=t}^{T} \prod_{s=t}^{r} (1 - \delta_s)$. Working back from period T, we can find that $NPV_{T-1} = 1$, and $NPV_{T-2} = 1 + (1 - \delta_{T-2})$, so that $NPV_{T-1} - NPV_{T-2} = -(1 - \delta_{T-2}) < 0$, and $NPV_{T-2} - NPV_{T-3} = -(1 - \delta_{T-2})(1 - \delta_{T-3}) < 0$, which is smaller in absolute value than the decrease in net present value from $T-2$ to $T-1$ since $\delta_t \in (0, 1)$ for all t. The rest follows by induction.■

In consequence, we know that if we denote a general function $g(t)$ to mean the net present discounted value of investing in health at age t, assuming that the costs of investment in health are nondecreasing in t, then $g'(t) < 0$ and $g''(t) < 0$. Now suppose we have some latent reservation utility for individual i denoted $\varepsilon_i \sim N(0, 1)$ for simplicity and the decision rule for whether we observe an individual making a particular investment in health at time t is $Y_{it} = 1[\varepsilon_i \leq -a_0 - a_1 D_{it} - g_i(t)]$, so that, denoting the cumulative distribution function of ε by F_ε,

$Pr(Y_{it} = 1) = F_\varepsilon(-a_0 - a_1 - g_i(t))$ if $D_{it} = 1$ and $F_\varepsilon(-a_0 - g_i(t))$ otherwise. Clearly $Pr(Y_{it} = 1)$ declines in t at an increasing rate regardless of the value of D_{it}. It follows that if, as in the set-up in the text, we have that $t_0 < t < t_k$ for those who have $D_{it} = 1$ and $D_{it-1} = 0$ and $t_k < t'_0 < t < t'_k$ for all those with $D_{it} = D_{it-1} = 1$ (the only case by construction of the sample where $\triangle D_{it} = 0$), then the unobserved net present discounted value of investing in health will lead to a smaller value of $E[\triangle Y_{it} | \triangle D_{it} = 1] - E[\triangle Y_{it} | \triangle D_{it} = 0]$ if this difference is negative.

Chapter 2

Dynamic Adjustment Following Discontinuous Regime Changes

Abstract

I study regression-discontinuity designs where the assignment variable evolves deterministically with time. I show that under mild assumptions, it is possible to identify whether the short-term effect isolated by studying the discontinuous change that occurs at the cutoff differs from the long-term effect of the regime change on the outcome of interest. The test for whether the two effects differ requires estimators of the derivatives either side of the cutoff, which are available under the usual regression-discontinuity assumptions. If the derivative of the conditional expectation changes discontinuously at the cutoff, then unless the first difference of the outcome of interest is negatively serially correlated this will change the trajectory as well as the level of the variable of interest at the cutoff, and so the long-run effect will differ from the short-run effect. I then illustrate the properties of the test with Monte Carlo evidence and use it in an empirical application: recovering the effect of Medicare coverage on insulin usage.

2.1 Introduction

A large literature uses discontinuous regime changes that are quasi-randomly assigned via an "assignment" / "forcing"/ "running" variable to identify the effect of the regime change, known as the "treatment", on some outcome around the discontinuity (Lee and Lemieux, 2010). A well-known limitation of these studies is that the effects they recover are only identified in a small neighbourhood of the discontinuous change, or "cutoff". This has prompted recent work that uses either matching on observables (Angrist and Rokkanen, 2015) to extrapolate effects away from the cutoff or that identifies how the effect would change if there were a small perturbation to the location of the cutoff (Dong and Lewbel, 2015). I show in this chapter that the latter technique has an additional use when the running variable evolves deterministically with time, as in the many regression-discontinuity designs where the assignment variable is age (e.g. Card et al. (2008, 2009), Battistin et al. (2009), Chen and Van der Klaauw (2008), Lalive (2007, 2008)). In this case, relatively mild assumptions allow the researcher to test whether the long-run effect of exposure to the discontinuous change at the cutoff is larger or smaller than the short-run effect estimated near the cutoff.

The intuition behind the result is relatively straightforward. If we only observe the level of the outcome of interest shifting discontinuously at the onset of some policy change, then unless we add further assumptions to the regression-discontinuity design about what happens away from the cutoff no more can be said about the relationship between the short-run and long-run effects of the discontinuity. By contrast, if the rate of change of the outcome also changes discontinuously at the threshold then more can be said regarding the long-run effect of the discontinuity - provided the researcher is prepared to assume that discontinuous changes in the trajectory of the outcome variable at the cutoff are persistent. In that case, it is justified to infer that the long-run effect of the policy differs from the short-run effect. In many economic applications, outcomes are positively serially correlated over time, and this by itself is sufficient (but not necessary) for the long-run effect to exceed the short-run effect.

This chapter lies at the intersection of two literatures: one on extrapolating effects away from the cutoff in a regression-discontinuity design, and the literature on estimating dynamic effects of

shocks. Many economic variables of interest are serially correlated, such as individual consumption (Hall, 1988, Flavin and Nakagawa, 2008) and labor supply (Hyslop, 1999, Keane and Sauer, 2009). This has led to a large literature on estimating the extent of serial correlation in these variables (Arellano and Bond, 1991, Blundell and Bond, 1998). In the Appendix, I provide a method for using a parametric AR(1) specification to identify serial correlation in an outcome of interest via the variation provided by a regression-discontinuity design. This is not the main focus of this chapter since, as discussed in the Appendix, the identification of the exact relationship between the short-run and long-run effect of crossing the threshold comes from a strong functional form assumption that the outcome of interest does not vary with the assignment variable in the absence of any intervention. I sacrifice the ability to make any precise statements about what the long-run effect of a discontinuous change is by avoiding this strong functional form assumption. I am nonetheless able to make statements about whether the long-run effect cumulates or dissipates over time under the weaker assumption that changes to the slope of the outcome variable at the cutoff persist over time and are not reversed at a later stage.

The rest of this chapter is organized as follows. Section 2.2 presents the econometric framework and the test for whether the short-run effect differs from the long-run effect. Section 2.3 provides Monte Carlo evidence to illustrate the conditions under which the test is appropriate. Section 2.4 applies the test developed in this chapter to data on insulin usage before and after diabetics qualify for health insurance via Medicare in the United States. Section 2.5 concludes.

2.2 Nonparametric Identification of Differences Between Short-Term and Long-Term Effects

First I define the notation I will use as well as the objects of interest. Let Y_{it} be the outcome of interest for individual i at time t. Let $D_{it} = 1$ if individual i is exposed to a treatment regime at time t and 0 otherwise. Defining Y_{it}^1 to be the potential outcome if individual i is exposed to treatment in period t, and Y_{it}^0 the potential outcome if they are not, in any given period t the observed outcome can be written $Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$. I assume that there is an assignment

rule depending on some assignment variable R_{it}; an individual i is eligible to have $D_{it} = 1$ in period t if and only if $R_{it} \geq \bar{R}$ (with $\bar{R}$ normalised to be 0 in most applications).

I shall assume that

Assumption A1. We observe Y_{it}, R_{it} and D_{it} for each individual.

Assumption A2. (Smoothness). The conditional expectation functions of the potential outcomes, $f_1(R_{it}) = E[Y_{it}^1|R_{it}]$ and $f_0(R_{it}) = E[Y_{it}^0|R_{it}]$, are twice continuously differentiable in R_{it}.

Assumption A3. R_{it} is a deterministic function of time, so that for any $s > t$ $R_{is} > R_{it}$.

Assumption A4. R_{it} is a continuous random variable, but is observed at discrete points of its support, so that we observe $R_{it} \in \{R_{min}, R_{min} + 1, ..., \bar{R} - 1, \bar{R}, \bar{R} + 1, ..., R_{max}$.

I next define the population objects of interest:

Definition 1. The short run effect of exposure to treatment τ^{SR} is defined

$$\tau^{SR} = \lim_{\epsilon \to 0} E[Y_{it}^1|R_{it} = \bar{R} + \epsilon] - E[Y_{it}^0|R_{it} = \bar{R} - \epsilon]$$

Denoting the usual regression-discontinuity estimand in the sharp design by τ^{RD}, by definition $\tau^{RD} = \tau^{SR}$. It will also be useful to define the Treatment Effect Derivative (TED) at R_{it} for a given placement of the cutoff value $\bar{R}$, $\frac{\partial \tau}{\partial R_{it}}(R_{it}, \bar{R})$. I will denote by $f_1^{(m)}(.)$ and $f_0^{(m)}(.)$ the m^{th} derivatives of $f_1(.)$ and $f_0(.)$ respectively.

Definition 2. The Treatment Effect Derivative (TED) $\frac{\partial \tau}{\partial R_{it}}(R_{it}, \bar{R})$ of $\tau = E[Y_{it}^1|R_{it}] - E[Y_{it}^0|R_{it}]$ with respect to R_{it} given the cutoff value is $R_{it} = \bar{R}$ is

$$\frac{\partial \tau}{\partial R_{it}}(R_{it}, \bar{R}) = \frac{\partial}{\partial R_{it}}\{E[Y_{it}^1|R_{it}] - E[Y_{it}^0|R_{it}]\} = f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it})$$

In Dong and Lewbel (2015), this parameter plays a crucial role, as under their assumption of local policy invariance (that this derivative does not vary with the placement of the cutoff value $\bar{R}$) this also recovers the effect of infinitesmally perturbing the value of the cutoff on the Local Average Treatment Effect (LATE) identified at the cutoff. Here I am concerned with a

different extrapolation: for a given cutoff, what can we infer from the Treatment Effect Derivative about the relationship between the effect at the cutoff and the long-run effect of the discontinuity? Additionally, which assumptions do we need to make to make that inference?

Definition 3. The long-run effect of treatment, τ_q^{LR}, is defined here as the cumulative effect of the discontinuous change q discrete points of support of the running variable to the right of the cutoff value $\bar{R}$,

$$\tau_q^{LR} = \tau^{SR} + \sum_{R_{it}=\bar{R}+1}^{\bar{R}+q} \{(E[Y_{it}^1|R_{it}] - E[Y_{it}^0|R_{it}]) - (E[Y_{it}^1|R_{it}-1] - E[Y_{it}^0|R_{it}-1])\},$$

Given that we only have information on the discontinuities in the level and the slope of $E[Y_{it}|R_{it}]$ at the cutoff (as in Dong and Lewbel (2015)), our ability to identify the long-run effect of the discontinuous change q points of support ("ages") from the cutoff, τ_q^{LR}, will depend on how persistent any change to the slope of the conditional expectation function (should one be detected) at the cutoff is. In the absence of any discontinuous change in the slope of $E[Y_{it}|R_{it}]$, we will be unable to rule out that there are undetected changes to the long-run trajectory of the average outcome, and so the null is that we cannot identify anything beyond the Local Average Treatment Effect (LATE) around the cutoff - the effect on individuals who are shifted into the treatment regime by the assigment rule within a small neighbourhood of the cutoff.

The following assumption regarding the persistence of discontinuous changes in the slope of the conditional expectation function is useful for making inferences regarding the relationship between the long-run and short-run effects of discontinuous changes. Recall that $E[Y_{it}^1|R_{it}] = f_1(R_{it})$, $E[Y_{it}^0|R_{it}] = f_0(R_{it})$, and $f_j^{(1)}(R_{it})$ be the first derivative of $f_j(.)$ with respect to R_{it} for $j = 0, 1$. Then if we assume that

Assumption A5. (Persistence of Trajectory Changes). $\lim_{\epsilon \to 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) > 0$ implies $f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it}) \geq 0$ whenever $R_{it} \geq \bar{R}$ and $\lim_{\epsilon \to 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) < 0$ implies $f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it}) \leq 0$ whenever $R_{it} \geq \bar{R}$.

(A5) also places restrictions on the higher-order derivatives of the treatment effects considered,

since the class of functions whose first derivatives have the same sign everywhere over some or all of their domain have one of the following properties:

Definition 4. A function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is *absolutely monotone* iff for all k, either $h^{(k)} > 0$ or $h^{(k)} < 0$, where $h^{(k)}$ denotes the k^{th} derivative of h.

Definition 5. A function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is *totally monotone* iff for all k, either$(-1)^k h^{(k)} > 0$ or $(-1)^k h^{(k)} < 0$, where $h^{(k)}$ denotes the k^{th} derivative of h.

Examples of totally monotone functions include the familiar classes used as non-satiated utility functions such as power functions $f(x) = \frac{x^{1-\rho}}{1-\rho}$ for $\rho \in \mathbb{R}$ and the exponential functions of the form $f(x) = K - e^{-x}$ for $K > 0$. If the conditional expectation functions $E[Y_{it}^0|R_{it}]$ and $E[Y_{it}^1|R_{it}]$ are linear, then the first derivative of $E[Y_{it}^1|R_{it}] - E[Y_{it}^0|R_{it}]$ is constant everywhere, and so (A5) is automatically satisfied.

Note that (A5) applies to all the points of support where $R_{it} \geq \bar{R}$, not just at the discrete points where R_{it} is observed. This will be important for the following result.

Proposition 1. If (A5) holds, then $\lim_{\epsilon \rightarrow 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) > 0$ implies $\tau^{SR} < \tau_q^{LR}$ for all $q > 0$, and similarly $\lim_{\epsilon \rightarrow 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) < 0$ implies $\tau^{SR} > \tau_q^{LR}$ for all $q > 0$.

Proof: Suppose (A5) holds. Take the case where $\lim_{\epsilon \rightarrow 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) > 0$. Then $f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it}) > 0$ everywhere that $R_{it} \geq \bar{R}$. By the Mean Value Theorem, for any interval between points in the support of R $[R_{it}, R_{it+1}]$ there exists some point in the support of R, R_{it}^*, not necessarily observed, such that

$$0 < [f_1^{(1)}(R_{it}^*) - f_0^{(1)}(R_{it}^*)][R_{it+1} - R_{it}] = [(f_1(R_{it+1}) - f_0(R_{it+1})) - (f_1(R_{it}) - f_0(R_{it}))],$$

and given that this holds for any arbitrarily chosen interval, it holds for each term in the sum over the intervals in$\{[\bar{R}, \bar{R}+1], [\bar{R}+1, \bar{R}+2], ..., [\bar{R}+(q-1), \bar{R}+q]\}$,

$$\sum_{R_{it}=\bar{R}+1}^{\bar{R}+q} \{(E[Y_{it}^1|R_{it}] - E[Y_{it}^0|R_{it}]) - (E[Y_{it}^1|R_{it}-1] - E[Y_{it}^0|R_{it}-1])\} > 0,$$

which implies that the second term in the expression for τ_q^{LR} is positive regardless of q and hence that $\tau^{SR} < \tau_q^{LR}$ for any q. The proof is identical for the case where $\lim_{\epsilon \to 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) < 0$. ■.

As will be borne out in the simulation results below, this holds when there is positive serial correlation in the dependent variable (almost trivially, in the case of an AR(1) process: given the formula for a geometric series, the change in the long-run mean will be $\frac{\tau^{SR}}{1-\rho}$ for an AR(1) process with persistence parameter $\rho \in (0,1)$, which increases the magnitude of the long-run effect relative to the short-run effect).

Lemma 1. If two random variables X and Y are positively correlated and have finite first and second moments, then there is an affine relationship between $E[X]$ and $E[Y]$ such that there is an affine function $g(.)$ for which

$$\begin{aligned} E[Y] &= g(E[X]) \\ &= a + bE[X], \end{aligned}$$

with $a \in \mathbb{R}$ and $b > 0$. Similarly,

$$E[Y|Z] = \tilde{a} + \tilde{b}E[X|Z]$$

with $\tilde{a} \in \mathbb{R}$ and $\tilde{b} > 0$ if Z is another random variable.

Proposition 2. If

(i) Y_{it} is positively serially correlated, i.e. we have $Corr(Y_{it}, Y_{is}) > 0$ for any $t \neq s$, and

(ii) the first derivative of the affine function $g(.)$ governing the positive serial correlation between Y_{it}, Y_{is} is common to both the treated and untreated states, so that

$$g^{(1)}(E[Y_{it}^1|R_{it}]) - g^{(1)}(E[Y_{it}^0|R_{it}]) = 0,$$

then Assumption (A5) holds.

Proof: Suppose Y_{it} exhibits positive serial correlation. Then if we take the case where $\lim_{\epsilon \to 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) > 0$, we want to show that at an arbitrary $R_{it} \geq \bar{R}$ $f_1^{(1)}(R_{it}) >$

$f_0^{(1)}(R_{it})$, i.e. the actual slope is greater than the counterfactual slope. By positive serial correlation, we know that for any interval $[t-1,t]$ there exists some affine function $g(.)$ such that $E[Y_{it}^j|R_{it}] = g_j(E[Y_{it-1}^j|R_{it-1}])$ and $g_j^{(1)}(E[Y_{it-1}^j|R_{it-1}]) > 0$ for $j = 0, 1$. Hence take

$$f_1(R_{it}) = E[Y_{it}^1|R_{it}] = g(E[Y_{it-1}^1|R_{it-1}]) = g(f_1(R_{it-1})),$$

and

$$f_0(R_{it}) = E[Y_{it}^0|R_{it}] = g(E[Y_{it-1}^0|R_{it-1}]) = g(f_0(R_{it-1})).$$

By the Chain Rule, we can write

$$f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it}) = g^{(1)}(f_1(R_{it-1}))f_1^{(1)}(R_{it-1})) - g^{(1)}(f_0(R_{it-1}))f_0^{(1)}(R_{it-1}) > 0,$$

and by Lemma 1 and part (ii) of the statement of the Proposition,

$$g^{(1)}(f_1(R_{it-1})) = g^{(1)}(f_0(R_{it-1})) = b,$$

for some $b > 0$. So

$$f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it}) = b \times [f_1^{(1)}(R_{it-1}) - f_0^{(1)}(R_{it-1})] > [f_1^{(1)}(R_{it-1}) - f_0^{(1)}(R_{it-1})]$$

by positive serial correlation. Since the interval $[t-1,t]$ was arbitrary, it follows that if $f_1^{(1)}(R_{it}) - f_0^{(1)}(R_{it}) > 0$ at any point R_{it} in the support of R then any subsequent point in the support of R, R_{is} with $s > t$ also has $f_1^{(1)}(R_{is}) - f_0^{(1)}(R_{is}) > 0$. Since we supposed that $\lim_{\epsilon \to 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) > 0$, this holds in particular for all $R_{it} \geq \bar{R}$. The proof is identical for the case where $\lim_{\epsilon \to 0}(f_1^{(1)}(\bar{R}+\epsilon) - f_0^{(1)}(\bar{R}-\epsilon)) < 0$. ■.

In a regression-discontinuity design, the treatment regime is applied if individual i's assignment variable R_{it} exceeds some threshold $\bar{R}$ at time t, so that $D_{it} = 1$ if (but not necessarily only if) $R_{it} \geq \bar{R}$. If this is the case only if $R_{it} \geq \bar{R}$, and for all individuals with $R_{it} \geq \bar{R}$, the regression-discontinuity design is said to be "sharp"; otherwise, it is said to be "fuzzy". The estimating equation that aims to recover the discontinuous effect of reaching the threshold $\bar{R}$ in the sharp design is

$$Y_{it} = \beta_0 + \beta_1 1[R_{it} \geq \bar{R}] + (1 - 1[R_{it} \geq \bar{R}])f_0(R_{it}) + 1[R_{it} \geq \bar{R}]f_1(R_{it}) + X_{it} + v_{it}, K\left(\left|\frac{R_{it} - \bar{R}}{h}\right| < 1\right), \tag{2.1}$$

where $f_0(.)$ and $f_1(.)$ are continuous (usually polynomial) functions, X_{it} is a vector of observable characteristics, and v_{it} is unobserved. $K(.)$ is a kernel function, chosen to be the Uniform kernel throughout this study, so that only observations within a neighbourhood of length equal to the bandwidth h are used in the estimation. To go from the empirical specification in this equation to identification of the long-run and short-run effects, we will need the standard regression-discontinuity identification assumption,

Assumption A6. (Continuity). The unobservable component of Y_{it}, labelled v_{it}, is continuous in R_{it} at $\bar{R}$, the eligibility cutoff.

Identification of the local average treatment effect comes from continuity of the levels of the functions $f_1(.)$ and $f_0(.)$ at the cutoff $R_{it} = \bar{R}$ together with Assumptions (A1) and (A6), since then, conditional on observables, $\widehat{\beta_1} = \lim_{\epsilon \to 0} E[Y_{it}^1 | R_{it} = \bar{R} + \epsilon] - E[Y_{it}^0 | R_{it} = \bar{R} - \epsilon] = \tau^{SR}$ (see Lee and Lemieux (2010) for a simple proof). Moreover, a discontinuous change in the first derivative of the conditional expectation function at the cutoff, $\lim_{\epsilon \to 0} f_1^{(1)}(\bar{R} + \epsilon) - f_0^{(1)}(\bar{R} - \epsilon)$, the Treatment Effect Derivative at $R_{it} = \bar{R}$, is consistently estimated by the interaction terms between the local polynomial terms and the indicator for whether $R_{it} \geq \bar{R}$ (see Dong and Lewbel (2015) for a proof). For example, in a local linear specification the relevant estimator is the coefficient on $R_{it} \times 1[R_{it} \geq \bar{R}]$. If these are all equal to zero, we cannot make any statements regarding the relationship between the RD estimate and the long-run effect of treatment on outcomes. If these indicate that the slope is more positive after treatment, then (A5) implies that the long-run effect is greater than the RD estimate, and if is more negative then the long-run effect is less than the RD estimate.

The bias of the estimator of the difference between slopes is larger than that for the difference between levels (the object of interest for the classical RD estimator) for a given bandwidth. This is easily shown by following the derivation of the nonparametric estimator's bias in terms of the bandwidth, h, found in Imbens and Kalyanaraman (2012), which shows that the term that is $o(h^2)$ in the expression for the bias of $\widehat{\tau^{SR}}$ in Imbens and Kalyanaraman (2012) is $o(h^3)$ in the expression for the bias of $\widehat{\frac{\partial \tau^{SR}}{\partial R_{it}}}|_{R_{it}=\bar{R}}$, the estimator of $\lim_{\epsilon \to 0} f_1^{(1)}(R_{it} + \epsilon) - f_0^{(1)}(R_{it} - \epsilon)$ It therefore seems likely

that the optimal bandwidth for inference on $\widehat{\frac{\partial \tau^{SR}}{\partial R_{it}}}|_{R_{it}=\bar{R}}$ is smaller than that for the traditional RD estimator. I leave it to future work to derive this optimal bandwidth.

The assumptions used here to identify whether the long-run and short-run treatment effects differ are stronger than those used in Dong and Lewbel (2015). They are relatively innocuous unless the underlying data-generating process is such that exposure to treatment affects the curvature of the conditional expectation function $E[Y_{it}|R_{it}, R_{it} \geq \bar{R}]$ at different points in time. This might occur if the effect of the treatment regime on outcomes is a "flattening" or "stretching" of the conditional expectation function. In these cases, Assumption (A5) would fail since then we could have $\frac{\partial \tau^{SR}}{\partial R_{it}}(R_{it})$ positive over some range of the R_{it} and negative over another range or vice versa. In most applications, there are not strong theoretical reasons to believe that this is due to the treatment regime. There are also common cases where the assumption does hold, such as in the case of positive serial correlation in the outcome variable. By contrast, negative serial correlation in the outcome variable renders conclusions from examining the first derivative at the cutoff invalid. Positive serial correlation in the outcome variable may not be a necessary condition for the extrapolation proposed here. I leave it to future work to derive a comprehensive set of conditions for the appropriateness of Assumption A5.

2.3 Monte Carlo Evidence on the Performance of the Difference-in-Slopes Method

In this section, I present simulation evidence for cases where the test proposed in the previous section for differences between the short-term and long-term effects of the discontinuous change works well, in that it detects persistent changes where they exist and doesn't detect them where they do not. I also discuss two cases where it is not expected to perform well: negative serial correlation in the outcome variable and changes in the curvature of the conditional expectation function.

I will simulate the following equations:

$$Y_{it} = \rho Y_{it-1} + \beta_0 + \beta_1 1[R_{it} \geq \bar{R}] + \beta_2 R_{it} + \beta_3 R_{it} \times 1[R_{it} \geq \bar{R}] + v_{it}, \tag{2.2}$$

$$Y_{it} = \rho Y_{it-1} + \beta_0 + \beta_1 1[R_{it} \geq \bar{R}] + \beta_2 R_{it} + \beta_3 R_{it} \times 1[R_{it} \geq \bar{R}] + \beta_4 R_{it}^2 + \beta_5 R_{it}^2 \times 1[R_{it} \geq \bar{R}] + v_{it}, \tag{2.3}$$

$$Y_{it} = \beta_0 + \beta_1 1[R_{it} \geq \bar{R}] + \beta_2 R_{it} + \beta_3 R_{it} \times 1[R_{it} \geq \bar{R}] + \beta_4 R_{it}^2 + \beta_5 R_{it}^2 \times 1[R_{it} \geq \bar{R}] + v_{it}, \tag{2.4}$$

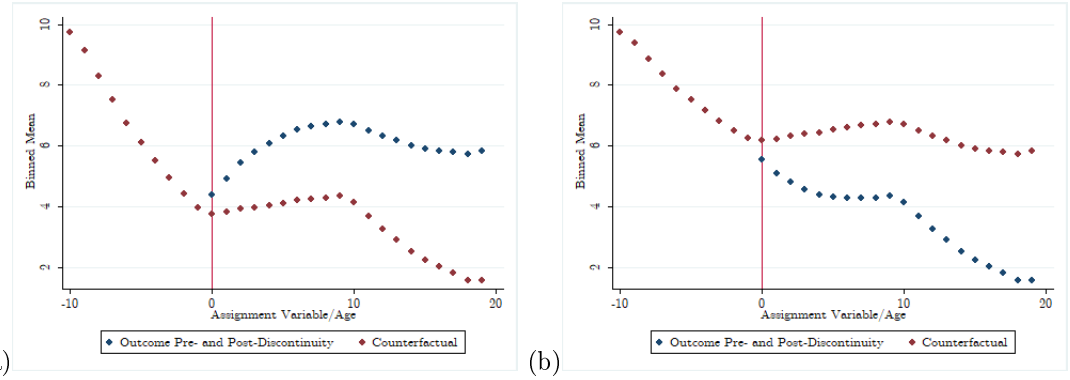
with Y_{it} being the outcome of interest, and R_{it} being the assignment variable, with $v_{it} \sim N(0,1)$ i.i.d. and exogenous throughout.

Equations 2.2 and 2.3 are included to examine Type I and Type II errors in cases where there is positive serial correlation, which produces a true divergence between the the short-run and long-run means, and negative serial correlation, which may lead to misleading inferences regarding the difference between short-run and long-run means. In addition, I separately simulate cases where the extent of positive serial correlation is either strong or weak; it may be expected that the power of the method to detect discontinuous changes in the slope of the conditional expectation function is different when there is only weak positive serial correlation.

Equation 2.4 is simulated to examine the performance of the method when faced with hump-shaped age profiles, which occur in several economic variables, such as wages (Keane, 2011) and consumption (Battistin et al., 2009). I separately simulate two separate cases: one where the entire age profile shifts by a constant after the threshold at which the discontinuity occurs is reached, and one where the age profile is distorted as a result of exposure to the treatment regime. For example, if the return to an activity is different at each age by age-specific increments due to the treatment regime, we may expect either flattening or bulging of the age profile relative to the counterfactual.

Even without confounding, there will still be spurious inferences from estimating discontinuous slope changes at the cutoff in the absence of any information regarding the effect of crossing the threshold on the age profile of the outcome of interest. In the simulations of Equation 2.2, I separately simulate cases where $\rho = 0.9$ (exemplified in Figure 2.1), $\rho = 0.1$, and $\rho = -0.9$ (exemplified in Figure 2.2), as well as cases where $\beta_1 = ln(2)$ and $\beta_1 = -ln(2)$, while keeping $\beta_0 = 0, \beta_2 = 0.02, \beta_3 = 0$ common across simulations of that equation. In the simulations of

Figure 2.1: Short- and Long-Run Effects of a Discontinuous Change with Positive Serial Correlation



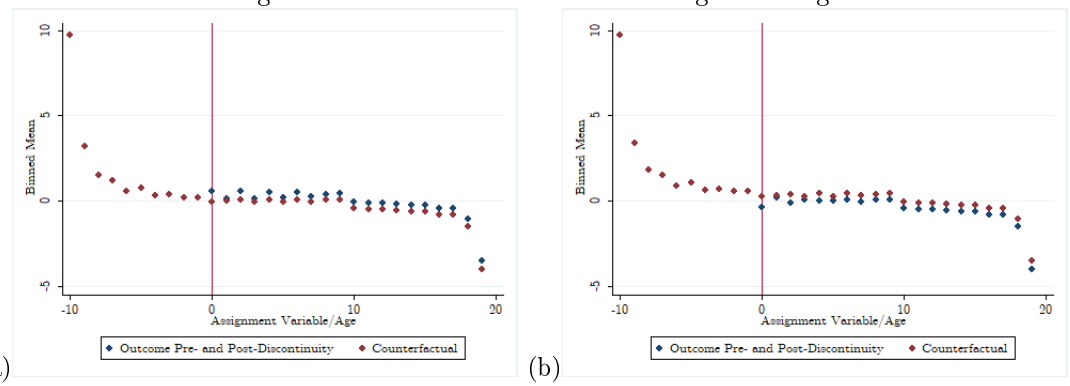
(a) (b)

Notes: Panel (a) shows simulated binned means from the equation $Y_{it} = 0.9Y_{it-1} + ln(2) \times 1[R_{it} \geq \bar{R}] + 0.02R_{it} + v_{it}$ and its counterfactual equation $Y_{it} = 0.9Y_{it-1} + 0.02R_{it} + v_{it}$; Panel (b) shows simulated binned means from the equations from the equation $Y_{it} = 0.9Y_{it-1} - ln(2) \times 1[R_{it} \geq \bar{R}] + 0.02R_{it} + v_{it}$ and the same counterfactual equation $Y_{it} = 0.9Y_{it-1} + 0.02R_{it} + v_{it}$.

Equation 2.3, I separately simulate two cases. The first case is one where there is only a level shift of the age profile at the cutoff, so that $\beta_0 = 10$, $\beta_1 = -ln(2)$, $\beta_2 = \beta_3 = \beta_5 = 0$, and $\beta_4 = -0.01$. The second case has, by contrast, $\beta_5 = 0.01$, so that the curvature after the cutoff is reduced to zero, flattening the age profile into a line parallel with the x-axis. This ensures that the counterfactual hump-shaped age profile initially lies below the post-treatment age profile before lying above it at points in the support of R further to the right of $\bar{R}$.

Figure 2.1 shows the results from one simulation where $\rho = 0.9$. As per the proof in the preceding section, the slope of the conditional expectation function is either more positive if the short-run effect is positive or more negative if it is negative at every point after the intercept shift than in the counterfactual. Figures 2.2 and 2.3 display cases of negative serial correlation and "flattening" of the conditional expectation function, respectively. These are cases where estimating the slope of the conditional expectation function around the cutoff can lead to misleading inferences regarding the difference between the short-term and long-term means. In the case of negative serial correlation, the estimated slope is positive for a positive short-run effect, and negative for a negative short-run effect, but the long-run difference between the realised and counterfactual means is smaller than is observed at the cutoff. In the latter case, where a hump-shaped CEF is flattened into a constant

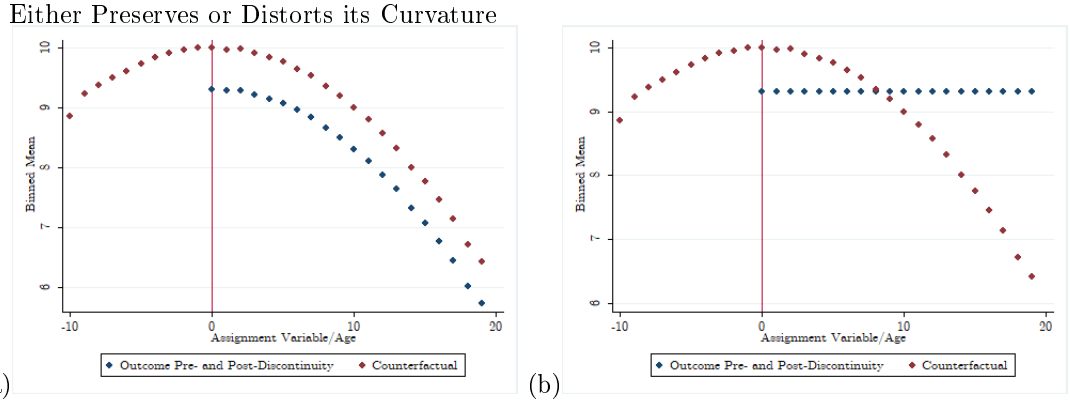
Figure 2.2: Short- and Long-Run Effects of a Discontinuous Change with Negative Serial Correlation



(a) (b)

Notes: Panel (a) shows simulated binned means from the equation $Y_{it} = -0.9Y_{it-1} + ln(2) \times 1[R_{it} \geq \bar{R}] + 0.02R_{it} + v_{it}$ and its counterfactual equation $Y_{it} = -0.9Y_{it-1} + 0.02R_{it} + v_{it}$; Panel (b) shows simulated binned means from the equations from the equation $Y_{it} = -0.9Y_{it-1} - ln(2) \times 1[R_{it} \geq \bar{R}] + 0.02R_{it} + v_{it}$ and the same counterfactual equation $Y_{it} = -0.9Y_{it-1} + 0.02R_{it} + v_{it}$.

Figure 2.3: Short- and Long-Run Effects of a Discontinuous Change to a Hump-Shaped Age Profile that Either Preserves or Distorts its Curvature



(a) (b)

Notes: Panel (a) shows simulated binned means from the equation $Y_{it} = 10 - ln(2) \times 1[R_{it} \geq \bar{R}] - 0.01R_{it}^2 + v_{it}$ and its counterfactual $Y_{it} = 10 - 0.01R_{it}^2 + v_{it}$; Panel (b) shows simulated binned means from the equation $Y_{it} = 10 - ln(2) \times 1[R_{it} \geq \bar{R}] - 0.01R_{it}^2 + 0.01 \times 1[R_{it} \geq \bar{R}] \times R_{it}^2 + v_{it}$ and its counterfactual $Y_{it} = 10 - 0.01R_{it}^2 + v_{it}$.

with age, the longer-term effect diminishes with age before becoming positive instead of negative, while at the cutoff the true difference in slopes is negligible since the peak of the hump coincides with the cutoff. Even more misleading inferences would occur in cases where the cutoff is at the left of the peak of the hump-shaped age profile, where the slope is increasing, so that if the curvature of the profile changes as a consequence of treatment the counterfactual slope could be more negative close to the cutoff but more negative far from the cutoff. In the case just described, Assumption (A5) fails - a discontinuous change in the slope of the conditional expectation function coincides with a difference of the opposite sign in the slope of the counterfactual CEF at another point where $R_{it} \geq \bar{R}$.

To examine the performance of the difference-in-slopes method, I simulate its use with five different types of data-generating processes: (1) positively serially correlated outcomes with high persistence, (2) positively serially correlated outcomes with low persistence, (3) negatively serially correlated outcomes, (4) hump-shaped age profiles with a level shift only, and (5) hump-shaped age profiles which have a different curvature for the treatment and control groups. (1) provides a case where the difference-in-slopes method should detect a true widening of the difference between the short-run and long-run effects. (2) provides a case where the short-run and long-run effects differ in magnitude, but where the weaker extent of serial correlation may result in a higher Type II error rate. (3) provides a case where differences in slopes are likely to provide misleading inferences regarding the difference between short-term and long-term effects. (4) provides another case where there is no difference between the short-term and the long-term effect, but where Type I errors may be mitigated by using a higher-order polynomial rather than a local linear specification. (5) provides a case where the difference between short-term and long-term effects changes over time, and so cannot in principle be detected by the difference-in-slopes method since Assumption (A5) fails.

Tables 2.1 and 2.2 shows estimates from 1000 replications of simulated local linear and local quadratic regressions with bandwidths close to the cutoff. For the local linear regressions, I set the bandwidth to $|R_{it}| \leq 2$, having normalised the running variable so that $\bar{R} = 0$, while for local quadratic regressions I use $|R_{it}| \leq 3$. The first set of results, in Table 2.1, are from 1000

estimations of Equations 2.2 and 2.3, all with $\beta_1 = ln(2)$, $\beta_0 = \beta_3 = \beta_4 = \beta_5 = 0$, $\beta_2 = 0.02$ but with different values of the persistence parameter ρ, set to either 0.9 (strong positive serial correlation), 0.1 (weak positive serial correlation) or -0.9 (negative serial correlation). The second set of results, in Table 2.2, are from 1000 estimations of Equation 2.4, with two variations: one with $\beta_0 = 10$, $\beta_1 = -ln(2)$, $\beta_3 = \beta_4 = \beta_5 = 0$ and $\beta_2 = -0.01$, so that the effect of exposure to treatment is just a level shift of a hump-shaped age profile, and one with $\beta_0 = 10$, $\beta_1 = -ln(2)$, $\beta_4 = -0.01$, $\beta_5 = 0.01$ and $\beta_2 = \beta_3 = 0$ so that untreated individuals have a hump-shaped profile which is "flattened" to a constant post-treatment. In each table I record the average t-statistic for the coefficients of interest (i.e. the linear and quadratic terms interacted with the eligibility indicator), as well as rates of either Type I or Type II errors depending on whether the underlying data-generating process exhibits more than a level shift. I record the rate of Type I errors for cases where Assumption (A5) does not hold and/or there is no difference between the short-run and long-run effects (Cases (3), (4), and (5)) and Type II errors for cases where Assumption (A5) holds and there is a difference between the short-run and long-run effects of exposure to the treatment regime (Cases (1) and (2)). In each simulation, I start with a single cross-section of 10000 individuals with initial condition $Y_{i1} \sim N(9.744, 0.261)$ to mimic the distribution of a variable such as the natural log of consumption, with the parameter values taken from the empirical distribution of log consumption in Battistin et al. (2009). I then simulate Y_{it} moving forward in time according to one of Equations 2.2, 2.3 or 2.4, with the pseudo-"age" variable normalised to be 0 at the cutoff as is standard practice in regression-discontinuity designs (Lee and Card, 2008) with a $Unif(-10, 10)$ distribution in the initial period $t = 1$. I simulate each process for ten periods, so that the age distribution in the final period is $Unif(0, 20)$. All the regressions on the simulated data pool the cross-sections together.

The results suggest that the difference-in-slopes method is relatively conservative. In cases where either Assumption (A5) fails or the short-run and long-run effects are equal, the number of false positives does not exceed the nominal rate. By contrast, the false negative rate is relatively high in cases where there is a difference between the short-run and long-run effect - particularly when a local quadratic specification is used instead of local linear regression. These simulation-

based results are consistent with the empirical application in the next section, where I do not find evidence of a difference between the short-run and long-run effects despite reasons from both theory and supporting evidence to believe that a difference exists.

Table 2.1: Average t-Statistics, Type I and Type II Error Rates of Tests for Differences in Slopes: AR(1) Processes

	(1)	(2)	(3)	(4)	(5)	(6)
	Local Linear Regression			Local Quadratic Regression		
	$\rho = 0.9$	$\rho = 0.1$	$\rho = -0.9$	$\rho = 0.9$	$\rho = 0.1$	$\rho = -0.9$
t-stat., R_{it}	23.12	2.27	-0.02	9.05	0.58	-0.32
t-stat., R_{it}^2	-	-	-	-2.51	-0.50	0.11
Type I	-	-	0.036	-	-	0.026
Type II	0	0.371	-	0	0.860	-

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors clustered at the unit level throughout. Each panel records average t-statistics for the difference in slopes either side of the cutoff, as well as the "Type I Error Rate" for the proportion of simulations in which the researcher would mistakenly reject the null hypothesis that the long-run and short-run means do not differ. Simulation results recorded are for the AR(1) process described by Equation 2.2.

In Table 2.1, I examine three cases, each with a local linear (a la Equation 2.2) or a local quadratic (a la Equation 2.3) specification: high positive serial correlation ($\rho = 0.9$), low positive serial correlation ($\rho = 0.1$) and negative serial correlation ($\rho = -0.9$). This table shows two main features that impact how conservative the difference-in-slopes test is: the strength of positive serial correlation when it is present and the order of the polynomial used to model the conditional expectation function either side of the cutoff. When a local linear specification is used for a process that exhibits low persistence (the $\rho = 0.1$ case), the Type II error rate is less than half the rate that results from using a local quadratic specification for the same process. For both local linear and local quadratic specifications, persistence as high as $\rho = 0.9$ is sufficient to drive the Type II error rate to 0. At the same time, the difference-in-slopes method's conservatism is displayed by its

lower-than-nominal Type I error rates for the case of negative serial correlation, which is a scenario in which Assumption (A5) is violated.

Table 2.2: Average t-Statistics and Type I Error Rates of Tests for Differences in Slopes: Hump-Shaped Age Profiles

	(1)	(2)	(3)	(4)
	Local Linear Regression		Local Quadratic Regression	
	$\beta_5 = 0$	$\beta_5 = 0.01$	$\beta_5 = 0$	$\beta_5 = 0.01$
t-stat., R_{it}	-2.10	-0.03	-1.43	0.02
t-stat., R_{it}^2	-	-0.03	-	0.55
Type I	0.554	0.026	0.295	0.032

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors clustered at the unit level throughout. Each panel records average t-statistics for the difference in slopes either side of the cutoff, as well as the "Type I Error Rate" for the proportion of simulations in which the researcher would mistakenly reject the null hypothesis that the long-run and short-run means do not differ. Cases where $beta_5 = 0$ are cases where there is only a level shift in the conditional expectation function after crossing the threshold for eligibility for treatment; cases where $beta_5 = 0.01$ are those where the hump-shaped CEF "flattens" and becomes a constant function after treatment while the counterfactual CEF retains its hump shape.

In Table 2.2, I simulate two instances of Equation 2.4. In the first instance, that of a level shift of a hump-shaped age profile, there is no difference between the short-run and long-run effects of the discontinuous change at the cutoff, and no corresponding discontinuous change in the slope of the conditional expectation function. In the second instance, the discontinuous change results in a "flattening" of the profile as well as a level shift. In this case, while there is an initial discontinuous decrease in the slope of the function just around the cutoff - from slightly positive to zero - there is no corresponding increase in the magnitude of the negative treatment effect with age. This is because the change in the curvature of the conditional expectation function violates Assumption (A5); if the way the slope changes itself is changed by exposure to treatment, we are no longer guaranteed that the observed change in the slope of the conditional expectation function at the

cutoff signals a similar change across the rest of the conditional expectation function for $R_{it} \geq \bar{R}$. In this particular case, the counterfactual slope of the conditional expectation function is actually more negative for $R_{it} \geq \bar{R}$, even though it is more positive for $R_{it} < \bar{R}$.

The results indicate that the difference-in-slopes method can have a high Type I error rate if only local linear regression is used (55.4% for the case with the level shift, and 29.5% for the case with the curvature change (Table 2.2)), even with a relatively small bandwidth and a large number of observations. This is consistent with the observation from the previous section that we should expect larger finite-sample biases for nonparametric estimators of derivatives than for estimators of levels of the same function. As in the results in Table 2.1, increasing the order of the local polynomial used to estimate the difference in slopes eliminates this problem.[1]

2.4 Application: The Insulin Usage of Diabetics Before and After Qualifying for Medicare Coverage at Age 65

In this section, I examine the consequences of applying the methods in this study to the substantive results of the earlier chapter in this thesis on the insulin usage of diabetics. This comes with the caveat that the outcome of interest here is a discrete choice (cf. Hyslop (1999), Keane and Sauer (2009). This means that the usual methods for direct estimation of the extent of serial correlation in the dependent variable such as the Arellano-Bond estimator (Arellano and Bond, 1991) are inappropriate. In the event, I do not recover evidence that the discontinuous decrease in insulin usage found in the earlier chapter has persistent effects beyond the level shift in the expected proportion of insulin users among self-reported diabetics. This may be surprising, since there are strong reasons to believe that human capital investment decisions are serially correlated (Grossman, 2000, Imai and Keane, 2004). Since failing to reject the null hypothesis is not the same as accepting the null, the results in this section leave open the possibility that insurance for treatment has

[1] The fact that the difference-in-slopes method performs better with conditional expectation functions that have lower curvature echoes a recent result in Kolesár and Rothe (2018) that inference in regression-discontinuity designs with discrete running variables requires some assumption, whether explicit or implicit, regarding the curvature of the conditional expectation function.

persistent effects on insulin usage decisions over time, but that the method used here is unable to detect them.

The requirement that the treatment effect is either absolutely or totally monotone places restrictions on the age profile of insulin in the treated and untreated states. The treatment effect must increase with age, even if the increase goes to 0 as age goes to infinity (as is the case for all totally monotone functions). Whether this is realistic will depend on if it is realistic for the difference between insured and uninsured persons' insulin usage to increase indefinitely. Consider the Appendix to Chapter 1, which uses the Grossman model to show that the returns to investment in health must be decreasing at an increasing rate with age. The return is zero regardless of insurance status in the period before the end of life, which implies that the effect of insurance on health investment behaviour is zero in that period. If the effect is negative at some point, and zero at another, then for a continuous age profile it must be that the first derivative is positive at some point. Hence the assumptions derived here to extrapolate from finding a negative first derivative of the conditional expectation function in this Chapter may fail in this setting. This may however not be empirically relevant if the range of ages considered in the empirical analysis is too low for the average difference between the insured and uninsured persons to begin to decline.

I use data from the Health and Retirement Study (HRS) and a cleaned version of the data called the RAND HRS (Chien et al., 2013). A detailed description of the data can be found in the preceding chapter. There are at least four reasons to believe that past decisions affect present decisions in health capital investment decisions. First, advice from medical professionals on whether to use a specific type of preventive medicine is likely to be correlated over time, so that individuals who follow a past recommendation are more likely to be recommended to continue using their regular method of prevention, absent some exogenous shock, and vice versa. Second, the difficulty of making adjustments to an individual's routine may generate habits in their demand for prevention, as in standard models of habits in consumption. Third, if individuals exhibit anticipation effects (as in Alpert (2016) or De Preux (2011)), then $Pr(Y_{it}|D_{it}, Y_{it-1}) \neq Pr(Y_{it}|D_{it})$ since the contemporaneous probability $Pr(Y_{it}|D_{it})$ will ignore the impact of past anticipatory actions (for example, delaying investment until a period when it becomes less expensive). Fourth, in general,

simple models of investment in human capital over the life cycle, investments in one period are complementary to investments in all other periods, and so the past history of decisions will affect each period's optimal choice for a given individual (as in Imai and Keane (2004) or Keane (2016)).

I first presage the main results with a test for positive serial correlation in the outcome variable, contrasting the OLS and Fixed-Effects estimators of the coefficient on the lagged dependent variable. I then present estimates of the coefficients of the age profile after the age of eligibility, which should be zero if there is no deviation in the trajectory of the dependent variable due to treatment.

The OLS and Fixed-Effects estimators are useful for bounding the value of the persistence parameter ρ since the OLS estimator $\widehat{\rho}_{OLS}$ is biased upward and the FE estimator $\widehat{\rho}_{FE}$ is biased downward (Nickell, 1981). A consequence of this is that the true value of ρ lies between the probability limits of these two estimators. They appear to produce somewhat different conclusions regarding the extent of persistence in the data depending on whether the 1998-2004 or 1998-2008 sample is used (Table 2.3, below). In both cases, the OLS estimator finds that the persistence parameter is less than one, and is around 0.8 (roughly at the upper end of the findings of prospective cohort studies of persistence in insulin usage (Brown et al., 1999)). For 1998-2004, the case where $\rho = 0$ cannot be ruled out by the FE estimator, whereas the estimate from 1998-2008 suggests that $\rho > 0$. The suggested interval for the values that ρ might take on is considerably wider than that suggested by studies in the medical literature of adherence to insulin therapy (cf. Brown et al. (1999), Rajagopalan et al. (2003)), which is an interval of $[0.64, 0.8]$.

Table 2.4 displays the results of applying the difference-in-slopes test to the empirical setting used in the preceding chapter. Despite the good theoretical reasons and supporting evidence for positive serial correlation in insulin usage decisions, I am unable to recover a significant difference in slopes (the only statistically significant result being a more precise estimate of a number close to 0), and hence the trajectory of insulin usage, whether using a local linear or local quadratic specification or the 1998-2004 or 1998-2008 period (I use the original bandwidths to avoid Type I errors associated with specification searches). The evidence in Table 2.4 is consistent with a level shift of the age profile of insulin usage following eligibility for health insurance via Medicare with no difference between the short-run and long-run effects of coverage.

Table 2.3: OLS and FE Estimates of AR(1) Persistence Parameter, Insulin Usage

	1998-2004		1998-2008	
	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)
Y_{it-1}	0.86***	-0.05	0.87***	0.23***
	(66.85)	(-1.62)	(97.56)	(12.50)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications exclude Medicaid recipients and smokers. Columns (1) and (3) record estimates of the effect of lagged insulin usage on current insulin usage for OLS regressions using the 1998-2004 and 1998-2008 samples, respectively, while Columns (2) and (4) record estimates of the effect of lagged insulin usage on current insulin usage for Fixed-Effects regressions using the 1998-2004 and 1998-2008 samples, respectively,

It seems likely that human capital investment decisions are positively serially correlated over time, since investment today raises the return on investment tomorrow in standard models of human capital accumulation such as the Grossman model (cf. Grossman (1972)). This would lead a researcher examining the effect of a discontinuous regime change on human capital investment decisions to expect, as per the Monte Carlo results above, that the long-run effect of the change exceeds any short-run effect that may be detected at the cutoff. If an agent invests less in her health, then the return to investing in her health is lower in the subsequent period, and hence lower at increasing increments at each subsequent age relative to the counterfactual. The logic of the Grossman model would therefore lead us to expect that a discontinuous decrease in a health investment behaviour such as insulin usage at age 65 should lead to a snowball effect whereby the long-run decrease in the rate of insulin usage is even larger than can be found at the cutoff. This is moreover consistent with the tentative evidence from OLS and Fixed-Effects estimators of the extent of serial correlation in insulin usage decisions that suggests that they are positively serially correlated. I am nonetheless unable to find evidence of a discontinuous change in the slope of the age profile of insulin usage at age 65, despite the strong theoretical reasons to believe it exists and the evidence from simulations that suggests it would be detected by the method used here.

Table 2.4: Tests for Differences in Slopes via Coefficients on Local Polynomial Interacted With Eligibility Indicator: Insulin Usage Pre- and Post-Medicare

	1998-2004		1998-2008	
	(1)	(2)	(3)	(4)
$R_{it} \times 1[R_{it} \geq \bar{R}]$	0.00*	0.00	0.00	-0.00
	(1.97)	(0.07)	(1.29)	(-0.04)
$R_{it}^2 \times 1[R_{it} \geq \bar{R}]$		0.00		0.00
		(0.06)		(1.56)
Bandwidth	83.33	83.33	66.23	66.23

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications use either local linear or local quadratic regression with age in months as the running variable and a Uniform kernel, exclude Medicaid recipients and smokers, and include time dummies, education dummies, work status, marital status, and an indicator for poor health as covariates. Column (1) reports results from using local linear regression for the survey waves covering 1998-2004, Column (2) reports results from using local quadratic regression for the survey waves covering 1998-2004, Column (3) reports results from using local linear regression for the survey waves covering 1998-2008, and Column (4) reports results from using local quadratic regression for the survey waves covering 1998-2008.

2.5 Conclusion

Every dynamic model in economics implies that the sequence of past decisions matters for current decisions, and that anticipated future decisions affect current decisions. This chapter has contributed to ongoing efforts to unite economic models with identifying variation in the data. It allows for the identification of differences between short-run and long-run effects of a discontinuous change via a relatively mild assumption regarding the persistence of changes to the trajectory of an outcome of interest. There are at least two other potential applications where the running variable is not linked deterministically to time. One of these is a spatial regression-discontinuity design, where the running variable is ordered by distance from a geographical boundary. For example,

Michalopoulos and Papaioannou (2013) use a spatial regression-discontinuity design to examine the impact of arbitrarily drawn colonial-era national boundaries on economic activity. The method proposed here would complement the method derived in Dong and Lewbel (2015): while the latter would yield the marginal effect of a small change in the location of a boundary, the former, proposed in this chapter, would identify whether the effect observed at the boundary accumulates or dissipates away from that boundary. Another application would involve identifying the existence of spillover effects across individuals who qualify for a program, as in Fu and Gregory (2019), who combine a regression-discontinuity design with a structural model to identify the effects of neighbourhood externalities in government grants for rebuilding in New Orleans after Hurricane Katrina. Though finding that the treatment effect derivative is nonzero at the cutoff is likely not sufficient to identify externalities from exposure to a policy, it would provide supporting evidence that the outcomes of units further away from the cutoff are affected by the outcomes of those just above the cutoff. The appropriateness of the assumptions necessary in both of these cases can be expected to vary from application to application and hence their use will be at the discretion of the researcher.

References

Alpert, A. (2016). The Anticipatory Effects of Medicare Part D on Drug Utilization. *Journal of Health Economics*, 49:28–45.

Anderson, T. W. and Hsiao, C. (1982). Formulation and Estimation of Dynamic Models Using Panel Data. *Journal of Econometrics*, 18(1):47–82.

Angrist, J. D. and Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58(2):277–297.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*. John Wiley & Sons.

Battistin, E., Brugiavini, A., Rettore, E., and Weber, G. (2009). The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach. *American Economic Review*, 99(5):2209–26.

Blundell, R. and Bond, S. (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87(1):115–143.

Brown, J. B., Nichols, G. A., Glauber, H. S., and Bakst, A. (1999). Ten-Year Follow-Up of Antidiabetic Drug Use, Nonadherence, and Mortality in a Defined Population with Type 2 Diabetes Mellitus. *Clinical Therapeutics*, 21(6):1045–1057.

Card, D., Dobkin, C., and Maestas, N. (2008). The Impact of Nearly Universal Insurance Coverage on Health Care Utilization and Health: Evidence from Medicare. *American Economic Review*, 98(5):2242–2258.

Card, D., Dobkin, C., and Maestas, N. (2009). Does Medicare Save Lives? *Quarterly Journal of Economics*, 124(2):597–636.

Chen, S. and Van der Klaauw, W. (2008). The Work Disincentive Effects of the Disability Insurance Program in the 1990s. *Journal of Econometrics*, 142(2):757–784.

Chien, S., Campbell, N., Hayden, O., et al. (2013). *RAND HRS Data Documentation, Version M*. Santa Monica, CA: RAND Center for the Study of Aging.

De Preux, L. (2011). Anticipatory Ex Ante Moral Hazard and the Effect of Medicare on Prevention. *Health Economics*, 20(9):1056–1072.

Dong, Y. and Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *Review of Economics and Statistics*, 97(5):1081–1092.

Flavin, M. and Nakagawa, S. (2008). A Model of Housing in the Presence of Adjustment Costs: A Structural Interpretation of Habit Persistence. *American Economic Review*, 98(1):474–95.

Fu, C. and Gregory, J. (2019). Estimation of an Equilibrium Model With Externalities: Post-Disaster Neighborhood Rebuilding. *Econometrica*, 87(2):387–421.

Gelman, A. and Imbens, G. (2018). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, pages 1–10.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80(2):223–255.

Grossman, M. (2000). The Human Capital Model. In *Handbook of Health Economics*, volume 1, pages 347–408. Elsevier.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209.

Hall, R. E. (1988). Intertemporal Substitution in Consumption. *Journal of Political Economy*, 96(2):339–357.

Hyslop, D. R. (1999). State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women. *Econometrica*, 67(6):1255–1294.

Imai, S. and Keane, M. P. (2004). Intertemporal Labor Supply and Human Capital Accumulation. *International Economic Review*, 45(2):601–641.

Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies*, 79(3):933–959.

Keane, M. P. (2011). Labor Supply and Taxes: A Survey. *Journal of Economic Literature*, 49(4):961–1075.

Keane, M. P. (2016). Life-Cycle Labour Supply with Human Capital: Econometric and Behavioural Implications. *Economic Journal*, 126(592):546–577.

Keane, M. P. and Sauer, R. M. (2009). Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior. *Econometrica*, 77(3):975–991.

Kolesár, M. and Rothe, C. (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, 108(8):2277–2304.

Lalive, R. (2007). Unemployment Benefits, Unemployment Duration, and Post-Unemployment Jobs: A Regression Discontinuity Approach. *American Economic Review*, 97(2):108–112.

Lalive, R. (2008). How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach. *Journal of Econometrics*, 142(2):785–806.

Lee, D. S. and Card, D. (2008). Regression Discontinuity Inference with Specification Error. *Journal of Econometrics*, 142(2):655–674.

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

Michalopoulos, S. and Papaioannou, E. (2013). National Institutions and Subnational Development in Africa. *The Quarterly Journal of Economics*, 129(1):151–213.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, 49:1417–1426.

Rajagopalan, R., Joyce, A., Ollendorf, D., and Murray, F. (2003). PDB5: Medication Compliance in Type 2 Diabetes Subjects: Retrospective Data Analysis. *Value in Health*, 6(3):328.

Wooldridge, J. M. (2005). Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity. *Journal of Applied Econometrics*, 20(1):39–54.

Appendix: Parametric Identification of an AR(1) Autoregressive Parameter with a Regression-Discontinuity Design

The standard dynamic panel data model consists of an AR(1) process with a persistence parameter that is common across individuals indexed by $i \in \{1, ..., N\}$ for periods $t \in \{1, ..., T\}$,ρ, unobserved

individual-specific fixed effects η_i and unobserved individual-specific idiosyncratic shocks v_{it}, such that we have, for some variable of interest Y_{it},

$$Y_{it} = \alpha + \rho Y_{it-1} + \eta_i + v_{it}, \tag{2.5}$$

For example, if Y_{it} is consumption, then one justification for including Y_{it-1} on the right-hand side is that the Euler equation derived from individuals' first-order conditions from the life cycle consumption problem, $E_t[U'(C_{it+1})] = [\beta(1+r)]^{-1}U'(C_{it})$, implies that there is a direct link between present and future planned consumption (as well as the stronger, testable implication, if utility is CRRA and Y_{it} is the natural log of consumption, that $\rho = 1$ in Equation 2.5).

An immediate problem known since Nickell (1981) is that all of the OLS, fixed-effects, and first-differences estimators of ρ are biased. In the case of OLS, since Y_{it-1} contains η_i by construction, Y_{it-1} is endogenous. Even if η_i is removed via fixed-effects or first-differencing, $\triangle v_{it}$ contains v_{it-1} which is contained in Y_{it-1} and hence $\triangle Y_{it-1}$, and the fixed-effects transformation $\ddot{Y}_{it-1}$ contains all of the past and present values of v_{it}. Nickell (1981) showed that the resulting bias of the OLS estimator is positive for positive values of ρ and that of the fixed-effects estimator is negative. Hence the true value of ρ is contained in the interval in between $\widehat{\rho}_{OLS}$ and $\widehat{\rho}_{FE}$, and any estimator that produces estimates outside of that interval is known to be inconsistent.

It was first noticed by Anderson and Hsiao (1982) that even with data on Y_{it} alone, if $T \geq 3$ then there are valid instruments available for the equation in first differences under the assumption that v_{it} is independently and identically distributed across time. This is because if we take first differences, η_i is no longer present in the resulting first-differenced specification. It then follows that Y_{it-2} is a valid instrument for $\triangle Y_{it-1}$ in the first-differenced equation. (Instrument relevance trivially follows from Y_{it-2} being contained in $\triangle Y_{it-1}$). This is because although Y_{it-2} contains η_i, η_i is not present in the second-stage equation, and so all that remains to be checked is that $E[\triangle v_{it}|Y_{it-2}] = 0$. This follows from the assumption that v_{it} is i.i.d., since Y_{it-2} contains $v_{it-2}, v_{it-3}, ...$ and so forth, but neither of v_{it} or v_{it-1}, which are the only terms present in the second-stage equation. The subsequent history of dynamic panel data estimators, with the most

widely used methods being due to Arellano and Bond (1991) and Blundell and Bond (1998), is one of finding similar moment conditions and combining them in a Generalised Method of Moments (GMM) framework.

By contrast, in this chapter, I use the variation generated by the discontinuity at the cutoff to identify ρ as well as other parameters of interest. Intuitively, I am able to do this because I make one more identifying assumption than does the standard regression-discontinuity framework, and as a result am able to identify one more parameter. This is a break with the prior literature cited above, in that the moment conditions are typically motivated by the structure of the econometric model or economic theory rather than institutional "design-based" variation.

If we augment the standard panel regression-discontinuity design with a lagged dependent variable to account for the interdependence of decisions made at different times, the estimating equation becomes

$$Y_{it} = \rho Y_{it-1} + \beta_0 + \beta_1 D_{it} + f(R_{it}, \gamma_0) + g(R_{it}, \gamma_1) \times 1[R_{it} \geq \bar{R}] + \delta X_{it} + \zeta t + \eta_i + v_{it}. \quad (2.6)$$

Let, for any random variable W, $lim_{R \to \bar{R}} E[W|R = R + \varepsilon] = W^+$ and $lim_{R \to \bar{R}} E[W|R = R - \varepsilon] = W^-$. We obtain

$$(Y_t^+ - Y_t^-)$$

$$= \beta_1 (D_t^+ - D_t^-) + \rho (Y_{t-1}^+ - Y_{t-1}^-)$$

$$= \beta_1 (D_t^+ - D_t^-) + \rho \beta_1 (D_{t-1}^+ - D_{t-1}^-) + \rho^2 (Y_{t-2}^+ - Y_{t-2}^-)$$

$$= ...$$

$$= \sum_{s=0}^{t} \rho^s \beta_1 (D_{t-s}^+ - D_{t-s}^-) + \rho^t (Y_0^+ - Y_0^-).$$

From the above, it appears at first glance that we may require the stringent condition

$$(Y_0^+ - Y_0^-) = 0$$

but here the fact that the running variable is age, which advances deterministically with time, allows for weaker identifying assumptions. Note that the fact that the running variable, though continuous, differs between any two periods by an integer increment, implies that for any $s \in \mathbb{Z}^+$, there is some $m \in \mathbb{Z}^+$ such that

$$R_{it-s} = R_{it} - sm, \tag{2.7}$$

If age is measured in years, and the data is collected taken annually, we can take $m = 1$, which just says that age in $t - s$ is, mechanically, age in period t minus s. For age measured in months, and data collected every two years, we have $m = 24$. Note that this doesn't require that the running variable R_{it} is itself an integer, but that it can be divided into a set of ordered intervals according to the integers. Equation (3) can be modified to accommodate any combination of periodicity of the data and increment of the running variable between waves of the data. Now in practice, regression-discontinuity designs rely on restricting attention to observations within some bandwidth h of the cutoff (cf. Hahn et al. (2001)). This restricts attention to all data points such that $|\frac{R_{it}-\bar{R}}{h}| < 1$. This, together with equation (6), implies that, for a given bandwidth h, there exists some k such that

$$k = inf_{k \in \mathbb{R}^+}\{R_{it-k} : R_{it-k} < \bar{R} | 0 < \tfrac{R_{it}-\bar{R}}{h} < 1\},$$

which is formalised in Lemma 1 below. For both Lemma 1 and Proposition 2, I assume without loss of generality that $m = 1$ in Equation (3) above.

Lemma 2. Let $R_{it} \in \mathbb{R}$, indexed by $i \in \{1, ..., N\}$ and $t \in \{1, ..., T\}$. Let h be the bandwidth such that for a given period t, $|R_{it} - \bar{R}| < h$. If $R_{it-s} = R_{it} - s$ for all t and $s \in \mathbb{Z}^+$, (i) there is a set $\Omega := \{q : 0 \leq \frac{R_{it}-\bar{R}}{h} < 1 \implies R_{it-q} - \bar{R} < 0\}$, and (ii) $\exists k := inf\Omega \in \mathbb{Z}^+$.

Proof: Since by definition $|R_{it} - \bar{R}| < h \forall R_{it}$, $\exists R_{it}^{max}$ s.t. $0 \leq R_{it}^{max} - \bar{R} < h$ but $R_{it}^{max} + 1 - \bar{R} \geq h$. Similarly, $\exists R_{it}^{min}$ s.t. $R_{it}^{min} - \bar{R} > 0$ but $R_{it}^{min} - 1 - \bar{R} < 0$. Take $q \in \mathbb{Z}^+$ s.t. $q \geq 1 + (R_{it}^{max} - \bar{R})$. Then

$R_{it-q}^{max} = R_{it}^{max} - q = \bar{R} - 1 < \bar{R}$

$R_{it-q}^{max} - 1 = R_{it}^{max} - q - 1 = \bar{R} - 2 < \bar{R}$

...

$R_{it-q}^{min} = R_{it}^{min} - q < R_{it}^{min} - 1 < \bar{R}$

So $\Omega = \{q : q \geq 1 + (R_{it}^{max} - \bar{R})\}$ and $k = inf\Omega = 1 + (R_{it}^{max} - \bar{R}) \in \mathbb{Z}^+ \blacksquare$

Note that the running variable R_{it} can be continuous or discrete, but changes deterministically with the integer index t so that the difference between values of R_{it} for the same individual in between two periods is an integer. The above Lemma

Proposition 3. Let k be defined as in Lemma 1. Then if (ignoring covariates and common time trends for simplicity):

(i) $Y_{it} = \rho Y_{it-1} + \beta_1 D_{it} + f(R_{it}, \gamma_0) + g(R_{it}, \gamma_1) \times 1[R_{it} \geq \bar{R}] + \eta_i + v_{it}$,

(ii)

$$(Y_{t-k}^+ - Y_{t-k}^-) = 0,$$

with v_{it} i.i.d. and

(iii) $E[\eta_i + v_{it}|R_{is}]$ and $f(R_{it}, \gamma_0), g(R_{it}, \gamma_1)$ are continuous in R_{is}, R_{it} for all s, t,

then the numerator for the fuzzy RD estimand becomes

$$Y_t^+ - Y_t^-$$

$$= \beta_1(D_t^+ - D_t^-) + \rho\beta_1(D_{t-1}^+ - D_{t-1}^-) + ... + \rho^{k-1}\beta_1(D_{t-k+1}^+ - D_{t-k+1}^-)$$

$$= \beta_1 \sum_{s=0}^{k-1} \rho^s (D_{t-s}^+ - D_{t-s}^-).$$

Proof: Note that

$$Y_t^+ - Y_t^-$$

$$= \beta_1(D_t^+ - D_t^-) + \rho(Y_{t-1}^+ - Y_{t-1}^-)$$

$$= \beta_1(D_t^+ - D_t^-) + \rho\beta_1(D_{t-1}^+ - D_{t-1}^-) + \rho^2\beta_1(D_{t-2}^+ - D_{t-2}^-)$$

$$+ ... + \rho^{k-1}\beta_1(D_{t-k+1}^+ - D_{t-k+1}^-) + \rho^k(Y_{t-k}^+ - Y_{t-k}^-)$$

$$= \beta_1 \sum_{s=0}^{k-1} \rho^s(D_{t-s}^+ - D_{t-s}^-) + \rho^k(Y_{t-k}^+ - Y_{t-k}^-)$$

$$= \beta_1 \sum_{s=0}^{k-1} \rho^s(D_{t-s}^+ - D_{t-s}^-),$$

where line two follows from a combination of (i) and (iii), the lines preceding the last line follow from repeated application of (i) and (iii), and the last line follows from (ii)■.

That is, given a set of individuals in period t whose ages are above the threshold $\bar{R}$, but lie within the interval $[\bar{R}, \bar{R} + h]$ defined by the chosen bandwidth h, there is some greatest lower bound k for the set of positive integers that, if subtracted from all the ages between $\bar{R}$ and $\bar{R} + h$, would produce a set of ages that are less than $\bar{R}$. Equivalently, k picks out the period $t - k$ in which all the individuals above the cutoff - but within the chosen distance from the cutoff h - were below the cutoff. The reason this is useful is that in any period $t - q$ such that $q \geq k$, $D_{it-q} = 0$ for all the individuals within h of the cutoff by definition. That is, all of the individuals in the treatment group were in the control group for periods $t - k, t - k - 1, ..., 0$. Therefore if we are only interested in the effect of D_{it}, β_1, we only need the extra restriction that

$$(Y_{t-k}^+ - Y_{t-k}^-) = 0,$$

with k defined as above. This is a regression-discontinuity analog of the usual restrictions on initial conditions necessary in dynamic panel data models (for which see Baltagi (2008). It requires that there are no other discontinuities that would result from an age-based cutoff at an age k survey periods in the past. Otherwise, we will confound the persistent effect of this past discontinuity with the other effects on the behavior of the treatment group. For example, in my context, individuals who are 66, with a survey taken every other year, with $k = 2$, must not have undergone a discontinuous jump in behavior at age 62 - which may well happen with U.S. data

since this is the first age at which Americans can claim Social Security benefits. Fortunately, this can be tested as per the usual checks that "placebo" cutoffs do not have the same effect as the cutoff of interest. Identification of the persistence parameter ρ then comes from comparing individuals who switched status last period, so that $R_{it} \geq \bar{R}$ but $R_{it-1} < \bar{R}$, who exhibit no persistent effect of treatment since they were not in the treatment group in the last period, with those further away from the cutoff. To illustrate this, let $\triangle D_{it} = 1[R_{it} \geq \bar{R}] - 1[R_{it-1} \geq \bar{R}]$ pick out this group of individuals (since there are no cases where $R_{it-1} \geq \bar{R}$ and $R_{it} < \bar{R}$ by design), $D_{it-1} = 1[R_{it-1} \geq \bar{R}]$, and suppose that $k = 2$. We can rewrite the regression-discontinuity estimand as

$$\beta_1(D_t^+ - D_t^-) + \rho\beta_1(D_{t-1}^+ - D_{t-1}^-) + \rho^2(Y_{t-2}^+ - Y_{t-2}^-)$$

$$= \beta_1 + \rho\beta_1(D_{t-1}^+ - D_{t-1}^-)$$

by the definitions of k and h.

Now if we rewrite the estimating equation to include D_{it-1}, so that we can partial out $\beta\rho$ from β, we obtain

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \beta_1\rho D_{it-1} + f(R_{it}, \gamma_0) + g(R_{it}, \gamma_1) + \delta X_{it} + \zeta t + \eta_i + v_{it} \tag{2.8}$$

$$= \beta_0 + \beta_1 \triangle D_{it} + \beta_1(1+\rho) D_{it-1} + f(R_{it}, \gamma_0) + g(R_{it}, \gamma_1) + \delta X_{it} + \zeta t + \eta_i + v_{it}, \tag{2.9}$$

The advantage of rewriting equation (7) as equation (8) is that it highlights the role of intra-treatment group comparisons in identifying ρ. β is identified from the difference between the group that switched status this period and those below the cutoff in the same period, and is the contemporaneous impact on individuals' choices above the cutoff. Individuals who had already switched by the last period undergo a contemporaneous effect β_1 and an effect carried over from when they switched status $\rho\beta_1$, which together is $\beta_1(1+\rho)$.

The ability to obtain a consistent estimate of ρ comes from not having to rely on more traditional exogeneity assumptions on account of the regression discontinuity design. The usual assumption that $E[\eta_i + v_{it}|y_{it-1}] = 0$ will fail since $E[y_{it-1}]$ contains η_i by construction, and first-differencing

does not do better since $E[\triangle v_{it}|\triangle y_{it-1}] = 0$ will be violated for the same reason. Identification in the regression discontinuity design only depends on the continuity of these conditional expectation functions at the cutoff, with the only additional restriction for this case being no discontinuity in the period $t-k$ in which all individuals in the treatment group are in the control group. This is less demanding than the traditional exogeneity assumptions necessary for standard dynamic panel data estimators.

Here, instead of a restriction on the distribution of the initial conditions (as in Wooldridge (2005)) or the even stronger requirement that they are exogenous, we require that there is no discontinuity in the period $t-k$ in which all the individuals treated in period t are below the cutoff. Fortunately, in my scenario this involves a test for a discontinuity at or near the Social Security claiming age of 62, which would be a necessary robustness check even if I were not aiming to identify ρ. In practice identifying ρ in the panel RDD simply requires more robustness checks for placebo cutoffs at earlier ages of the kind already usually reported in standard RDDs.

There is a simple intuition for why we can identify the persistence parameter ρ using an RDD with a bandwidth that includes more years than elapse between survey years. Some individuals who are included to the right of the cutoff will already have switched status in the last survey year. As a result, the net effect on their behavior is a combination of the effect from the last survey year persisting and the effect in the current survey year. If we can pin down the effect of having changed status between the last survey and the current year by focusing on the subset of individuals who changed, we can back out the effect in the current year and hence the persistent component of previous treatments by subtraction.

In this scenario, estimating ρ becomes a special case of instrumental variables estimation with heterogeneity in the second-stage coefficients. The first effect is the contemporaneous effect of Medicare coverage, β_1. Individuals sufficiently far from the cutoff will also exhibit a spillover effect from having switched status in the previous period. Comparing individuals who only switched this period and so only exhibit the contemporaneous effect with their peers who exhibit both the contemporaneous and persistent treatment effects allows us to recover the extent of persistence ρ[2].

[2] A structural model is used to identify spillover effects due to spatial externalities in an RDD in Fu and Gregory

This is analogous to the scenario in the previous subsection where we can pick out individuals who were previously uninsured and enrol in traditional Medicare and compare them with individuals who were covered prior to age 65 and purchase supplemental insurance. Only one additional identifying assumption - the restriction on the initial conditions - is necessary to identify ρ.

Note that the only part of the above identification arguments that requires panel data is information on outcomes in previous periods. In principle, this information could be recorded in a single cross section - at no point in this subsection have the identification arguments had to rely on estimating equations in first differences or fixed-effects estimation, which are only possible with repeated observations of the same individuals. Two drawbacks of cross-sectional data will prevent identification according to the above arguments from succeeding in practice. First, few cross-sectional data sets contain detailed histories of past outcomes that are required for the approach in this subsection. Second, differential mortality in a cross-section will select for individuals with longer *ex ante* expected lifespans and hence higher returns from using prevention. This will lead to spurious upward drift in the probability of using prevention with age due to those with weaker incentives for usage being selected out of the sample. In a balanced panel, this concern no longer holds since we use the realised life-cycle profiles of usage to estimate the model. In a cross-section differential mortality will bias the estimator of ρ downward if β_1 is negative since the effect will only be observed to persist for a short period of time before exhibiting mean reversion. This in turn will lead to an underestimate of the longer-term effect at the cutoff for life-cycle decision-making. This problem can be expected to be especially acute when using the Uniform kernel, since it weights observations within the chosen bandwidth equally (though the order of the polynomial used for the local polynomial regressions either side of the cutoff will weight observations differently (Gelman and Imbens, 2018). By contrast, the Edge (Triangular) kernel gives more weight to observations closer to the cutoff, and so will suffer from this problem to a lesser extent. In a panel the Triangular kernel may tilt the scales too aggressively in favor of larger values of ρ, since it will give less weight to observations further from the cutoff. For this reason I exclusively use the Uniform kernel for the

(2019). The analysis in this subsection could in principle be applied to similar cases where an individual's actions are linked to those of other individuals due to social networks or geographical location, rather than to their own past actions.

panel RDDs in this chapter.

Static and Dynamic Treatment Effects

In a fuzzy cross-sectional RDD, the effect of interest is often defined as the Local Average Treatment Effect given by

$$\frac{Y^+ - Y^-}{D^+ - D^-},$$

(cf. Hahn et al. (2001)). Conveniently, if we take these limits for equation (1), we obtain

$Y^+ - Y^- = \beta_1(D^+ - D^-) + E[f(R^+,\gamma) - f(R^-,\gamma)] + E[\eta_i|R = R^+] - E[\eta_i|R = R^-] = \beta_1(D^+ - D^-)$ by continuity of $f(.)$ and η in R at $\bar{R}$, and so $\beta_1 = \frac{Y^+ - Y^-}{D^+ - D^-}$, the treatment effect of interest. In the sharp design, $D^+ = 1$ and $D^- = 0$ so this simplifies to $\beta_1 = Y^+ - Y^-$. With dynamic decision-making and the assumption from the last section that decisions from before period $t - k + 1$ do not matter for the RD estimand, we have

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \rho\beta_1 D_{it-1} + ... + \rho^{k-1}\beta_1 D_{it-k+1} + \sum_{j=0}^{k-1} \rho^j \eta_i + \sum_{j=0}^{k-1} \rho^j v_{it-j}, \tag{2.10}$$

so that while the static effect is given by β, which would be the only effect of treatment if $\rho = 0$, the cumulative effect from the previous $k-1$ periods is given by $\beta(\frac{1-\rho^k}{1-\rho})$. The individual coefficients are interpretable as net effects of treatment given that there was a discontinuity in treatment status in the previous period. Taking the limits as in the static case above for $k = 2$, we obtain

$$Y_t^+ - Y_t^- = \beta_1(D_t^+ - D_t^-) + \rho\beta_1(D_{t-1}^+ - D_{t-1}^-) \implies \beta_1 = \frac{Y_t^+ - Y_t^-}{(D_t^+ - D_t^- + \rho(D_{t-1}^+ - D_{t-1}^-))},$$

which shows that β_1 is a downward-biased estimator of the average effect for all individuals "near" the cutoff, if individuals "near" the cutoff have heterogeneous effects of being treated due to spillovers from previous periods. Suppose that the one-period spillover is the only additional treatment effect. Then the cumulative effect is

$$\frac{Y_t^+ - Y_t^-}{D_t^+ - D_t^-} + \frac{Y_t^+ - Y_t^-}{Y_{t-1}^+ - Y_{t-1}^-}\frac{Y_{t-1}^+ - Y_{t-1}^-}{D_{t-1}^+ - D_{t-1}^-} = \frac{Y_t^+ - Y_t^-}{D_t^+ - D_t^-} + \frac{Y_t^+ - Y_t^-}{D_{t-1}^+ - D_{t-1}^-},$$

now notice, moreover, that if we had just written the AR(1) process for Y_{it}, ignoring any effects of treatment status, and taken the limits as above, we would obtain

$$Y_t^+ - Y_t^- = \rho(Y_{t-1}^+ - Y_{t-1}^-) \implies \rho = \frac{Y_t^+ - Y_t^-}{Y_{t-1}^+ - Y_{t-1}^-},$$

so that the cumulative effect of interest is given by

$$\frac{Y_t^+ - Y_t^-}{D_t^+ - D_t^-} + \frac{Y_t^+ - Y_t^-}{D_{t-1}^+ - D_{t-1}^-} = \frac{Y_t^+ - Y_t^-}{D_t^+ - D_t^-} + \frac{Y_t^+ - Y_t^-}{Y_{t-1}^+ - Y_{t-1}^-}\frac{Y_{t-1}^+ - Y_{t-1}^-}{D_{t-1}^+ - D_{t-1}^-} = \beta_1 + \rho\beta_1, \tag{2.11}$$

given that β_1 captures the contemporaneous, static effect of treatment, ignoring spillovers from previous periods. It is trivial to extend this to the case where $k > 2$, where the sum of the coefficients on the treatment indicators $D_{it}, ..., D_{it-k+1}$ captures the dynamic effect of treatment for a group of individuals with discontinuities in treatment status across periods $t, ..., t-k$.

If the coefficient ρ is of independent interest, then a confidence interval for ρ is available via the delta method after taking the ratio of the coefficients on the contemporaneous treatment indicator D_{it} and the first lag of the treatment indicator D_{it-1} as the test statistic. Denoting the coefficient on D_{it-1} as $\beta_2 = \rho\beta_1$, and the standard errors of the estimators β_1 and β_2 by Σ_{11} and Σ_{22}, the delta method yields a standard error for ρ, Σ_ρ, of

$$\Sigma_\rho = \beta_1^{-2}[\Sigma_{11} + (\frac{\beta_2}{\beta_1})^2\Sigma_{22}], \tag{2.12}$$

which is then consistently estimated by replacing $\beta_1, \beta_2, \Sigma_{11}$ and Σ_{22} by their sample counterparts. This is of course provided that $\beta_1 \neq 0$. If $\beta_1 = \beta_2 = 0$, then the limiting distribution of their ratio is a Cauchy distribution, regardless of the true value of ρ.

Estimation

In this subsection I examine the identification of dynamic treatment effects in an RDD. The exposition in this section closely follows that of Dong and Lewbel (2015). Given a bandwidth h in which the observations lie, local linear regression with a Uniform kernel is equivalent to dropping observations for which the running variable lies outside of the interval $[\bar{R}-h, \bar{R}+h]$ and estimating

the equation (ignoring covariates and common time trends for simplicity):

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 D_{it-1} + \gamma_0(R_{it} - \bar{R}) + \gamma_1(R_{it} - \bar{R}) \times 1[R_{it} \geq \bar{R}] + \gamma_2(R_{it} - \bar{R}) \times 1[R_{it-1} \geq \bar{R}] + \eta_i + v_{it}, \tag{2.13}$$

However, it is important to note that there is an immediate problem with collinearity between the running variable interacted with the treatment indicator $1[R_{it} - \bar{R}]$ and the separate treatment indicators D_{it} and D_{it-1} if the interval for estimation is restricted to the one under which identification of the persistence parameter can be achieved by the preceding arguments. This is easily seen algebraically: since $1[R_{it-1} - \bar{R}]$ is equal to one both if an individual is currently past the cutoff and if they were past the cutoff in the previous period, $1[R_{it-1} - \bar{R}] = 1[R_{it} - \bar{R}] + \triangle 1[R_{it} - \bar{R}]$. Since the conditions for identification restrict attention to a few points of the running variable R_{it}, attribution of the separate effect D_{it-1} to state dependence (that is, the dependence of current Y_{it} on last period's realised Y_{it-1}) depends on assuming a counterfactual with no "trend" in the running variable that would otherwise have been observed: that is, that the effect of passing the cutoff value $\bar{R}$ is restricted to the level shift β_1.

In the usual terminology of regression discontinuity designs, this amounts to an assumption that the Conditional Expectation Function (CEF) $E[Y_{it}|R_{it}]$ would be locally constant either side of the cutoff in the absence of state dependence. Whether this assumption is credible will depend on the particular application at hand. For example, suppose Y_{it} is consumption for individual i in period t, and R_{it} is their age in period t. If they simply consume all their current income in each period, and their income process is locally constant either side of the cutoff value of $\bar{R}$, then it would be reasonable to expect their consumption to be constant with age either side of the cutoff as well. If their income is rising or falling steeply after the cutoff age, then failing to account for this will lead us to attribute dynamic decision-making to agents who are not making forward-looking decisions and behaving in line with a static model. The conditions for identification in this scenario do not allow for the separate identification of age trends that are independent of forward-looking decisions and the strength of state dependence. In the particular application cited above, this implies that this framework is less useful for identifying the persistence parameter using data on younger agents,

who typically experience rapid wage growth, than for older agents, whose income is typically less variable from year to year.

Note that a linear approximation to the function $E[Y|\bar{R} \leq R < \bar{R} + \varepsilon, R_{it-1} < \bar{R}]$ is given by

$$\beta_0 + \beta_1 + (\gamma_0 + \gamma_1)(R_{it} - \bar{R}), \tag{2.14}$$

while a linear approximation to $E[Y|\bar{R} - \varepsilon \leq R < \bar{R}, R_{it-1} < \bar{R}] = E[Y|\bar{R} - \varepsilon \leq R < \bar{R}, R_{it-1} \geq \bar{R}]$ is given by

$$\beta_0 + \gamma_0(R_{it} - \bar{R}), \tag{2.15}$$

so that in a hypothetical sharp RDD, the treatment effect for individuals with $R_{it-1} < \bar{R}$ (those who switched between the current survey period and the last survey period) is given by $plim(\widehat{\beta}_1)$. Moreover, we have that a linear approximation to $E[Y|\bar{R} \leq R < \bar{R} + \varepsilon, R_{it-1} \geq \bar{R}]$ is given by

$$\beta_0 + \beta_1 + \beta_2 + (\gamma_0 + \gamma_1 + \gamma_2)(R_{it} - \bar{R}), \tag{2.16}$$

so that the treatment effect for individuals with $R_{it-1} \geq \bar{R}$ is given by $plim(\widehat{\beta}_1 + \widehat{\beta}_2)$. If there is no spillover effect from having already crossed the threshold in the previous period, then these treatment effects are the same, and so $plim(\widehat{\beta}_2) = \beta_2 = 0$. $\widehat{\beta}_1$ corresponds to the static treatment effect and $\widehat{\beta}_1 + \widehat{\beta}_2$ the dynamic treatment effect. In a fuzzy design, we have separate regressions for the treatment indicators D_{it}, D_{it-1} on indicators of assignment to treatment status $1[R_{it} \geq \bar{R}], 1[R_{it-1} \geq \bar{R}]$ (in this chapter's scenario, whether an individual is older than 65 in the subscripted period):

$$\begin{aligned} D_{it} = \pi_{0t} + \pi_{1t}1[R_{it} \geq \bar{R}] + \pi_{2t}1[R_{it-1} \geq \bar{R}] + \pi_{3t}(R_{it} - \bar{R}) + \pi_{4t}1[R_{it} \geq \bar{R}] \times (R_{it} - \bar{R}) \\ + \pi_{5t}1[R_{it-1} \geq \bar{R}] \times (R_{it} - \bar{R}) + w_{1it} \end{aligned}, \tag{2.17}$$

$$D_{it-1} = \pi_{0t-1} + \pi_{1t-1}1[R_{it} \geq \bar{R}] + \pi_{2t-1}1[R_{it-1} \geq \bar{R}] + \pi_{3t-1}(R_{it} - \bar{R}) + \pi_{4t-1}1[R_{it} \geq \bar{R}] \times (R_{it} - \bar{R}) + \pi_{5t-1}1[R_{it-1} \geq \bar{R}] \times (R_{it} - \bar{R}) + w_{2it} \quad (2.18)$$

The estimated coefficients on treatment status $\widehat{\pi}_{1t}, \widehat{\pi}_{2t-1}$ are then used as divisors for $\widehat{\beta_1}$ and $\widehat{\beta_2}$, so that the static effect $\beta = \frac{Y_t^+ - Y_t^-}{D_t^+ - D_t^-}$ is given by $\frac{\widehat{\beta_1}}{\widehat{\pi}_{1t}}$ and the dynamic effect $\beta + \rho\beta = \frac{Y_t^+ - Y_t^-}{D_t^+ - D_t^-} + \frac{Y_t^+ - Y_t^-}{Y_{t-1}^+ - Y_{t-1}^-} \frac{Y_{t-1}^+ - Y_{t-1}^-}{D_{t-1}^+ - D_{t-1}^-}$ by $\frac{\widehat{\beta_1}}{\widehat{\pi}_{1t}} + \frac{\widehat{\beta_2}}{\widehat{\pi}_{2t-1}}$. Note that R_{it} is used in the regression for D_{it-1} and not R_{it-1} since R_{it} is the running variable in the main estimating equation and not R_{it-1}.

Chapter 3

Resolution of Uncertainty, New Information, and Accuracy in Subjective Medical Expenditures Forecasts

Abstract

I use recently collected data from the Health and Retirement Study to examine how individuals' forecasts of their out-of-pocket medical expenditures respond to anticipated and unanticipated shocks, as well as the accuracy of these responses. I find that individuals' medical expenses forecasts are inaccurate for at least three reasons. First, they do not take the sequential resolution of uncertainty into account when making their forecasts: individuals who are 10-12 months away from qualifying for Medicare at age 65 do not make significantly different forecasts from individuals who are 1-3 months from qualifying. By contrast, being diagnosed with, and surviving, an episode of cancer in between waves has a large effect on individuals' subjective distributions of medical expenditures, in line with its effect on actual medical ex-

penditures. Second, 6.9% of respondents are unable to respond to the forecast questions with probabilities that yield increasing CDFs for their subjective medical expenses distributions. Third, individuals are ignorant of at least one institutional detail concerning their insurance status: spouses who experience divorce or widowhood are more likely to report no longer being covered by their spouse's employer's plan within 24 months or fewer, despite official regulations mandating that they be included for up to 36 months following marital dissolution.

3.1 Introduction

This chapter examines the determinants and accuracy of individuals' beliefs about their future health care costs. Medical expenses are a significant component of both individual's planned spending and their exposure to risk over the life cycle (De Nardi et al., 2010). In consequence, how individuals form their predictions regarding their future out-of-pocket health care costs has important implications for their saving behavior and ability to adequately insure themselves against risk via precautionary saving, health insurance, or the use of preventive medicine.

Using recently collected data from the Health and Retirement Study, I am able to provide the first direct evidence on how anticipated and unanticipated changes contribute to individuals' perceptions of their future medical expenses, as well as the accuracy of these perceptions. Previous studies have tended to assume that individuals have rational expectations, with the econometrician no better informed than the agents being modelled, and individuals updating according to the rule that the econometrician sets in her model. Data on individuals' subjective expectations of their future medical expenses allow this assumption to be examined directly.

I am able to recover the following findings. First, a substantial number of respondents are unable to produce coherent forecasts, making statements that imply decreasing cumulative distribution functions, in violation of the laws of probability. This first finding provides *prima facie* plausibility for the conjecture that there may be important inaccuracies in the remaining forecasts. It adds to the literature begun by (Tversky and Kahneman, 1983) on the "conjunction fallacy": individuals commonly ascribe higher probabilities to the conjunction of two events they deem intuitively plausible than to either event individually. In the context of this chapter, the event "the next

12 months' medical expenses exceed $3000" covers a subset of the cases where the event "the next 12 months' medical expenses exceed $1500" and so mathematically can't be more probable than the latter. Nonetheless, 6.9% of the subjective medical expenses forecasts elicited in the Health and Retirement Study make this mistake.

Second, individuals display widespread ignorance regarding at least one insurance regulation in the United States: the COBRA Act of 1985 entitles spouses covered by an employer's insurance plan to continue on that plan for up to 36 months following divorce or widowhood (Madrian, 1998). Despite this, more than half of the individuals who are either divorced or widowed prior to being surveyed report that they are no longer covered by their spouse's insurance, even though the time between surveys is less than 36 months in almost all cases. This also has a marked effect on their medical expenses forecasts, even when limiting the sample so that the forecast period cannot entail a change in insurance status as a result of marital dissolution. This contributes to a large recent literature on the relatively low take-up of social programs among those who are eligible for them, and lack of awareness of these programs as an explanation for this phenomenon. Notable papers in this literature include Aizer (2003), who finds that information campaigns significantly increase enrolment in Medicaid, and more recently Bhargava and Manoli (2015), who find that simple informational nudges are enough to produce large increases in the filing of tax returns, concluding that the difficulty individuals have with parsing the necessary information is a major contributor to the low take-up of welfare programs. A factor that allows me to attribute the perceived lack of access to insurance to lack of awareness is that the default is for those already on their spouse's employer's health insurance plan to stay on. In other scenarios in the literature where individuals are required to sign up to benefit from a given rule or program, it is difficult to disentangle whether take-up is low because enrolling in a given program is administratively burdensome or because they are unaware that they are eligible for the program in question.

Third, individuals are slow to incorporate the resolution of uncertainty regarding their future medical expenses. Forecasts made 9-12 months from eligibility for healthcare under Medicare at age 65 do not differ significantly from forecasts made 0-3 months from eligibility. This is despite the fact that the probability of experiencing an adverse event for which the individual in question is

not covered mechanically falls as they approach the age of eligibility. This echoes Elder (2013), who finds that individuals do not take into account that they are mechanically more likely to survive to 75 conditional on having already survived to 64 compared with having survived to 60. There is some evidence that these same individuals do change their forecasts upon qualifying for Medicare, but almost entirely for the lower end of the distribution. These subjective perceptions contrast with Barcellos and Jacobson (2015) who find that there is a 53% reduction in the 95^{th} percentile of the medical expenses distribution upon qualifying for Medicare coverage at age 65. That individuals reactions to Medicare mostly occur just as they qualify, rather than as a gradual adjustment in anticipation of qualifying, is consistent with the results on discontinuous changes in individuals' behaviour found in Chapter 1 (see that Chapter for more discussion of this point).

Fourth, individuals underreact to new information regarding their health status when it arrives. Following diagnoses of diabetes and cancer, individuals revise their expectations of future out-of-pocket medical expenditures to a degree much smaller than is typically found following these diagnoses.

The rest of this chapter is organised as follows. Section 3.2 reviews the pre-existing literatures on subjective expectations of future expenditures and/or shocks and the determinants of out-of-pocket medical spending. Section 3.3 describes the data and some stylised facts regarding the subjective medical expenses forecasts as well as actual medical expenses realised in the survey. Section 3.4 describes the empirical strategies I use to examine how individuals' beliefs respond to anticipated and unanticipated changes as well as the accuracy of their beliefs. Section 3.5 describes the results of the empirical analysis. Section 3.6 concludes.

3.2 Review of the Literature

This chapter lies at the intersection of two separate literatures: one on the formation and accuracy of subjective expectations elicited from individuals in surveys, and one on the distribution and dynamics of medical expenditures. In the first literature, eliciting individuals' subjective expectations has been previously done for earnings risk (Dominitz and Manski, 1997), mortality risk (Hurd and

McGarry, 1995, 2002), and the probability of retirement (Benitez-Silva and Dwyer, 2005). A key challenge in each case is getting individuals to work with probabilities, which is typically quite difficult for survey respondents. For example, Dominitz and Manski (1997) report having to prompt individuals to revise their answers if they implied cumulative distribution functions that were decreasing. In surveys of mortality expectations, there are spikes at 0, 50, and 100% for responses to questions regarding the likelihood of death (Gan et al., 2005). In the case of the response that there is a 50% chance of an event occurring, it is unclear whether individuals are expressing a belief that either event is equally likely or agnosticism (Hurd, 2009).

The difficulty of eliciting interpretable subjective forecasts motivates two sub-literatures: one on rounding of subjective expectations (Manski and Molinari, 2010, Giustinelli et al., 2018, 2019), and one on the accuracy or otherwise of subjective forecasts. Giustinelli et al. (2019), most importantly for this chapter, finds that treating reported subjective probabilities as precise probabilities rather than as representing an interval of values can lead to misleading conclusions. When comparing point predictions and intervals, Giustinelli et al. (2019) found that 30% of the respondents in their data reported point probabilities that were not contained in the intervals elicited after further probing by the interviewer. Throughout this chapter I treat the point probabilities reported by individuals as precise. The point estimates I recover in this chapter may therefore be subject to bias due to individuals' reports of specific probabilities not reflecting the uncertainty of their beliefs and the true range of scenarios they view as plausible. Giustinelli et al. (2018) examine whether the rounding of reponses to 0, 50, or 100 reflects an underlying interval of probabilities representing individuals' uncertainty regarding the future. They develop a framework for representing the econometrician's uncertainty regarding the true belief underlying reports that are rounded to 0/50/100, and apply it to a subset of the expectations questions elicited in the Health and Retirement Study (HRS). This leads to weaker inferences that are presented in the empirical section of this chapter. The objects of interest are typically set-identified rather than point-identified. Though not addressed in this chapter, this suggests a potentially fruitful way of extending its empirical results: accounting for uncertainty in the responses due to rounding and then producing identified sets rather than point estimates of parameters.

Moreover, there are papers that assess the accuracy of individual forecasts. This is easiest when the forecast made is of events that are common to all forecasters, such as inflation (Inoue et al., 2009) or economic expansions and contractions (Souleles, 2004). In those cases, individuals' forecasts can be compared with their deviation from the common shock. It is less straightforward to assess the accuracy of forecasts when individuals are forecasting events with strong idiosyncratic components and private information, such as their expenditures (Jappelli and Pistaferri, 2000), earnings (Dominitz and Manski, 1997), dates of retirement (Benitez-Silva and Dwyer, 2005), or lifespan (Hurd and McGarry, 2002, Elder, 2013). This is because it is difficult, in the presence of private information, to determine what the "right" forecast should be when only considering the information accessible to the econometrician. Whether individuals' forecasts are found to be accurate or "rational" in the sense of rational expectations appears to depend on the event being forecast: Benitez-Silva and Dwyer (2005) fail to reject the null hypothesis that individuals have rational expectations regarding their date of retirement, while Elder (2013) finds systematic biases in individuals' forecasts of their probabilities of surviving to age 75.

The last literature to be discussed in this section is on the effect of medical expenses risk on individuals' decisions. Forecasts of medical expenses are most similar to individuals' forecasts of other processes with substantial idiosyncratic components, such as income, retirement, or mortality. French and Jones (2004) developed an explicit model of medical expenses processes that was later incorporated into structural models of life-cycle saving and labor supply by De Nardi et al. (2010), French and Jones (2011), and De Nardi et al. (2016a). These papers assume that individuals' subjective beliefs coincide with the econometrician's model of the medical expenses process. In this chapter I find that individuals' subjective medical expenses beliefs do not necessarily reflect many factors that influence their future medical expenditures.

3.3 Data

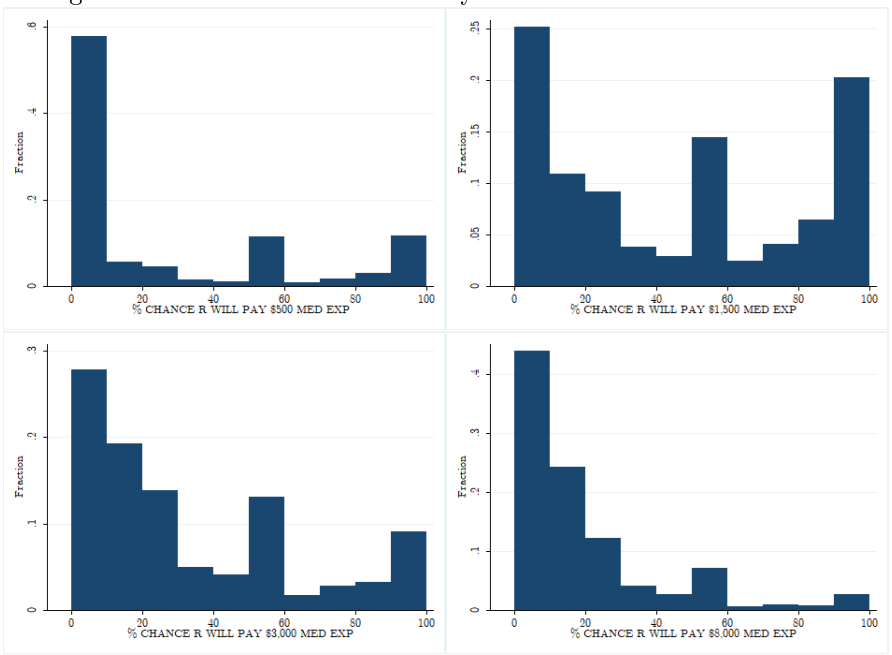
The data come from the 2010, 2012 and 2014 waves of the Health and Retirement Study (HRS), supplemented with a public use, cleaned version of the same data by a team of researchers at the

RAND corporation (informally referred to as the RAND HRS). This survey began in 1992 and has returned to survey the same individuals, a cohort of respondents aged 51-61 in 1992, every two years. The sample frame includes all members of the households in which these individuals live, so that the age range within the sample is 18 to 102. The key variables of interest are derived from questions that elicit individuals' subjective CDFs of future medical expenses. The first question in this series is

Please think about what you might spend out-of-pocket for your own medical expenses over the next year, including expenses such as doctor and dentist expenses, hospitals, nursing homes, prescription drugs and any others. Please include expenses that you would pay yourself (or a family member for you), but do not include what is covered by insurance. On a scale from 0 to 100 (where 0 means absolutely no chance and 100 means absolutely certain), what are the chances that you will spend more than $1,500 during the coming year?

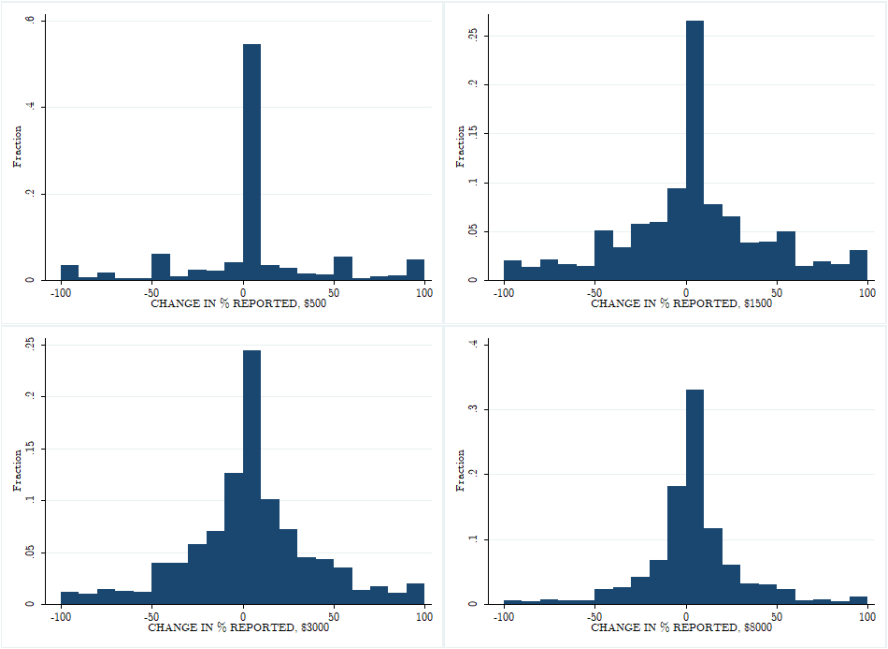
If a respondent answers "0", the question is repeated for a threshold of $500 so as to have a better chance of eliciting two points on the CDF. Otherwise, the question is repeated for two higher thresholds, $3000 and $8000. The rates of nonresponse are high, but mitigated by the large sample size. Across the 2010, 2012 and 2014 waves of the HRS, there are 61 335 person-year observations, with 39 702 individuals responding to both the $1500 and the $3000 threshold questions, and a further 10 755 person-year observations with responses to the lower $500 threshold.

Figure 3.1: Distributions of Subjective Probabilities of Incurring Out-of-Pocket Medical Expenditures in Following 12 Months Pooled Across Survey Years



Notes: Figures display fraction of respondents to the survey question "Please think about what you might spend out-of-pocket for your own medical expenses over the next year, including expenses such as doctor and dentist expenses, hospitals, nursing homes, prescription drugs and any others. Please include expenses that you would pay yourself (or a family member for you), but do not include what is covered by insurance. On a scale from 0 to 100 (where 0 means absolutely no chance and 100 means absolutely certain), what are the chances that you will spend more than $[X] during the coming year?", where [X] is one of 500, 1500, 3000 or 8000. Observations are pooled across the 2010, 2012 and 2014 waves of the Health and Retirement Study (HRS). There are 10 775 person-year observations of responses to the $500-threshold question across all waves, with 51 176 person-year observations for the $1500-threshold question, 39 702 for the $3000-threshold question, and 32 226 for the $8000-threshold question.

Figure 3.2: Distributions of Wave-to-Wave Changes in Subjective Probabilities of Incurring Out-of-Pocket Medical Expenditures in Following 12 Months, Pooled



Notes: Figures display difference in between survey waves in percentage points in answers to the survey question "Please think about what you might spend out-of-pocket for your own medical expenses over the next year, including expenses such as doctor and dentist expenses, hospitals, nursing homes, prescription drugs and any others. Please include expenses that you would pay yourself (or a family member for you), but do not include what is covered by insurance. On a scale from 0 to 100 (where 0 means absolutely no chance and 100 means absolutely certain), what are the chances that you will spend more than $[X] during the coming year?", where [X] is one of 500, 1500, 3000 or 8000. Observations are pooled across the 2012 and 2014 waves of the Health and Retirement Study (HRS).

Two limitations of the forecast data will limit the scope of the inferences that can be made later on in this chapter. First, as documented below, the distribution of forecasts will not be able to capture the fact that the modal figure for out-of-pocket medical spending over 24 months is $0.[1] The second is that if all the forecasts are the same, then we cannot make any inferences about any other moments of the distribution. This is true for 17.5% of those who respond to the subjective

[1] French and Jones (2004) and Keane and Stavrunova (2011) adopt different strategies for modelling medical expenditures distributions with a significant number of zero expenditures. The former paper simply recodes all observations of less than $250 to $250, while the latter specifically develops the "mixture of Tobits" model to account for this feature of the distribution.

medical expenses forecasts questions in the HRS. For these reasons, the majority of this chapter will examine the effects of the exogenous events studied on the raw percentages reported to the four subjective medical expenses questions, instead of working with an imputed subjective medical expenses distribution.

Histograms of the four questions, pooled across survey years (Figure 3.1), shows that they display similar heaping anomalies to other subjective expectations questions, with spikes at 0, 50 and 100% (Hurd, 2009). In Section 3.4 of this chapter, I examine the consequences of exogenous events on both the average percentage that individuals report. When imputing changes to the moments of the subjective distributions, I assume that individuals are located at one of these modal values before an exogenous change of interest occurs (most often at 50%).

Figure 3.2 shows the distribution of changes in the forecasts between waves. The distributions are what would be expected if no new information arrives between waves on average and there is no non-random drift in individuals' subjective expectations, which is consistent with rational expectations (compare Benitez-Silva and Dwyer (2005)). This is illustrated by the large spikes at 0 in the changes in between waves for all four questions. Since the main regressions in this chapter will have the change in between waves as the dependent variable, large systematic deviations from the modal outcome of no change should be relatively straightforward to detect.

The persistence of some of the explanatory variables between periods (such as the indicators for whether an individual lies within the observation period of interest) presents a problem for inference in the presence of serial correlation in the error term (Bertrand et al., 2004). Fortunately, it is well known that clustering the standard errors within the cross-sectional unit of observation (here, the individual) allows the usual hypothesis tests to have the correct size. Since clustering reduces to standard Eicker-Huber-White robust standard errors in the absence of clustering in the error term (Cameron and Miller, 2015), this will also automatically correct for the heteroscedasticity induced by the heaping present in the dependent variable.

Tables 3.1, 3.7, 3.9 and 3.11 examine the consequences of cross-sectional demographic differences on the forecasts that individuals make. The results are intuitive: whites, individuals with more education, and richer individuals are more likely to report higher probabilities of spending at least

$500 or at least $1500, but are more likely to report lower probabilities of spending at least $8000 in the twelve months following the survey. This likely reflects the better options available to such individuals for purchasing planned medical expenditures such as preventive care and doctor's appointments,who are in turn less exposed to the risk of incurring large medical expenditures. (Note however that these results are obtained while controlling for insurance indicators on the right hand side simultaneously, so access to better insurance options provides an incomplete account for the differences according to demographics. In particular, racial and educational gaps in medical expenditures forecasts persist even after controlling for insurance status and the log of household income, indicating that there is a substantial gap in both planned expenditures and medical expenditures risks that is not explained by these factors).

Tables 3.2, 3.8, 3.10 and 3.12 report the effect of the most common forms of insurance held by Americans on their medical expenses forecasts, realized medical expenses, and use of health services conditional on their demographic characteristics. The direction of causation is confounded in most cases by adverse or advantageous selection (see Fang et al. (2008) for evidence on the latter), as well as moral hazard (Einav and Finkelstein, 2011). The most consistent pattern is that being covered by Medicaid is strongly positively correlated with the utilisation of medical services (Table 3.12) while being strongly negatively correlated with both forecasts of future medical expenses and both measures of average past expenditures and the risk of incurring large medical expenditures (as measured by the 95^{th} percentile. This seems to warrant the tentative conclusion that the low-income low-asset individuals eligible for Medicaid coverage enjoy significant protection from medical expenditures risk as a consequence of their enrolment.

Overall, the patterns found for the subjective medical expenditures forecasts appear to be the same as those found for realised medical expenditures and use of health services (Appendix A). Richer households' members report 0.6 more doctor's visits for every additional percentage point increase in household income (Column 5 of Table 3.11 in Appendix A), while also reporting fewer hospitalisations (Columns 1-3 of Table 3.11). Accordingly, their medical expenditures are higher on average, but while they forecast higher probabilities of incurring at least $500 or $1500 over the subsequent 12-month period, they forecast *lower* probabilities of incurring at least $8000. This

Table 3.1: Correlates of 12-Month Forecasts of Future Medical Expenditures: Demographic Characteristics

	(1) $Pr(>= 500)$	(2) $Pr(>= 1500)$	(3) $Pr(>= 3000)$	(4) $Pr(>= 8000)$
White	3.96**	4.21***	-0.15	-3.43***
	(2.58)	(5.85)	(-0.23)	(-6.51)
Male	-2.17	0.52	-0.71	-0.93*
	(-1.46)	(0.79)	(-1.24)	(-2.16)
High School Grad.	2.50	1.01	-2.88**	-4.60***
	(1.29)	(1.07)	(-3.23)	(-5.85)
College Grad.	7.87**	4.25***	-1.59	-4.85***
	(3.15)	(3.91)	(-1.58)	(-5.76)
Poor Health	4.95*	11.08***	9.78***	5.01***
	(2.41)	(13.41)	(12.16)	(7.46)
Age	0.20	1.03**	0.37	0.11
	(0.27)	(2.66)	(1.04)	(0.38)
Age^2	0.00	-0.01*	-0.00	0.00
	(0.31)	(-2.47)	(-0.60)	(0.19)
ln(HHInc)	1.72*	1.96***	0.61	-0.93***
	(2.03)	(5.46)	(1.82)	(-3.50)
N	2757	18864	16000	13658

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for sources of health insurance.

likely reflects higher outlays on preventive medicine and better access to health care, together with the lower risk of incurring large medical expenses that this engenders. Note that insurance status is controlled for in these regressions, so that there is a statistically significant independent effect of household income on these outcomes even after accounting for the insurance coverage to which households have access.

In order to translate differences in subjective medical expenditures forecasts into changes in

Table 3.2: Correlates of 12-Month Forecasts of Future Medical Expenditures: Insurance Dummies

	(1) $Pr(>= 500)$	(2) $Pr(>= 1500)$	(3) $Pr(>= 3000)$	(4) $Pr(>= 8000)$
Medicaid	-7.35*	-10.87***	-4.46*	-0.37
	(-2.53)	(-6.03)	(-2.34)	(-0.20)
Medicare	-4.85	0.44	-1.63	-1.19
	(-1.55)	(0.35)	(-1.44)	(-1.38)
Own Employer	3.70*	2.67***	-1.02	-2.27***
	(2.13)	(3.63)	(-1.52)	(-4.22)
Spouse's Employer	1.91	-0.91	-2.42**	-1.67*
	(0.73)	(-0.89)	(-2.72)	(-2.53)
Union	-4.83**	-5.94***	-2.46**	0.17
	(-2.58)	(-6.79)	(-3.24)	(0.29)
VA/CHAMPUS	-8.45**	-11.13***	-6.75***	-4.06***
	(-3.01)	(-6.83)	(-4.61)	(-3.67)
N	2757	18864	16000	13658

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for source of insurance, sex, race, education (via indicator variables for whether an individual has graduated high school (incl. GED holders) or graduated from a four-year college), a quadratic in age, and the natural log of household income.

moments of the medical expenses distribution such as the mean or the variance, we need to make some parametric assumption regarding the distribution of medical expenditures. The reason for this is that since the question in the HRS picks out different quantiles of the forecast distribution for each individual, we cannot in general make interpersonal comparisons of medical expenditures forecasts or aggregate across forecasts unless individuals face the same underlying data-generating process. Fortunately, each individual's responses pick out at least two quantiles and so, given that previous studies have found that the lognormal distribution approximates the cross-sectional distribution of medical expenses reasonably well (French and Jones, 2004) and that this requires

only two quantiles to be fit to the data, I fit a lognormal distribution with the quantiles implied by individuals' subjective forecasts.

3.4 Empirical Strategy

In this section, I describe the methods I use in the following section to (a) examine how individuals' forecasts of their out-of-pocket medical expenditures respond to anticipated and unanticipated changes, and (b) assess the accuracy of their forecasts. A key point is that I cannot use the HRS data to compare predicted and realised medical expenditures one-for-one, as the predictions refer to a different time period from the one for which the realisations are recorded. I nonetheless can compare whether the changes in the moments of the predicted medical expenditures distribution are similar to those recorded elsewhere in the literature, and whether the size of the relative contributions of different events to the moments of the predicted medical expenses distribution matches their relative contributions *ex post*.

3.4.1 Event-Study Design for Anticipated and Unanticipated Shocks

The main regressions of interest have $\triangle P_{it}^{\tau} \in [-100, 100]$ as the dependent variable, where $P_{it}^{\tau} \in [0, 100]$ denotes the individual i's report of the probability (in percent) that their following twelve months' out-of-pocket medical expenses will exceed $\$\tau$ when surveyed in period t, for $\tau \in \{500, 1500, 3000, 8000\}$. To track the effect of unanticipated events, I bin the forecasts by individuals' age in months, starting with individuals who are 12 months away from qualifying for Medicare (and hence make forecasts for a 12-month period with no Medicare coverage throughout) and end with individuals who are newly qualified for Medicare coverage (and hence make forecasts for a 12-month period during which they are covered by Medicare for the entirety). As a placebo test, I examine whether forecasts are revised to a similar extent in the months following qualification for Medicare. I aggregate the months into three-month periods, so that the resulting dummy variables are indicators for the quarters preceding and following the month in which individuals turn age 65. I will write $Q_{itk} = 1$ for an individual who has had k quarters of eligibility for Medicare.

For example, $Q_{it,-5}$ is 1 if i is between 12 and 15 months (inclusive) from Medicare eligibility, and 0 otherwise. I estimate coefficients for five quarters before and after the quarter in which individuals qualify for Medicare, with a separate intercept α_1 to absorb differences from individuals who are more than five quarters away from becoming eligible, so that all the coefficients are relative to the average for those who are more than five quarters away from eligibility. The main regressions of interest take the form

$$\triangle P_{it}^{\tau} = \alpha_0 + \alpha_1 1[Q_{itj} = 1, j \geq 6] + \sum_{k=-5}^{5} \beta_k Q_{itk} + \delta X_{it} + \sum_{s=2}^{3} \theta_s 1[t = s] + \sum_{s=0}^{1} \phi_s' Z_{it-s} + v_{it}, \quad (3.1)$$

where X_{it} is a vector of demographic and socioeconomic characteristics, including insurance coverage dummies (coverage via the individual's or their spouse's employer, union, Veteran's Affairs, Medicaid, or Medicare), Q_{itk} are a set of indicators for the number of quarters individual i is from Medicare eligibility in survey year t, Z_{it} is a vector of dummy variables that record the onset of cancer or diabetes, or the dissolution of individual i's marriage due to separation, divorce, or widowhood in between periods $t-1$ and t, together with an interaction between the marital dissolution indicator and the indicator for whether an individual is covered by their spouse's employer's insurance plan (contained in X_{it}), and v_{it} is unobserved heterogeneity. The revisions to the forecasts $\triangle P_{it}^{\tau}$ should not be affected by events that convey no new information, or that do not materially affect individuals' medical expenses in expectation. In consequence the coefficient β_0 on the variable Q_{i0}, for example, measures the change in an individual's medical expenses forecast in the quarter they qualify for Medicare both relative to the previous period and relative to the other quarters around the age of eligibility. Since individuals are mechanically more likely to have more of the twelve months following a given survey period covered by Medicare the closer they are to eligibility (as well as relative to the previous survey period, when they were on average two years younger), all of the coefficients $\beta_{-4}, ..., \beta_{-1}$ should be negative due to the sequential resolution of uncertainty regarding the subsequent twelve months of out-of-pocket medical expenditures. Since the Z_{it} record changes that occur in between waves, the response of $\triangle M_{it}^{E,W}$ to the Z_{it} reflects the

incorporation of new information due to *ex ante* unanticipated events. I use only the events that occur in between periods $t-1$ and t and in between $t-2$ and $t-1$ since this drops no further periods from which observations can be drawn than does using the first difference of the forecasts as the dependent variable.

The main problem with any interpretation of the insurance indicators by themselves is that adverse selection and moral hazard will induce a positive correlation between spending on health care and insurance coverage (Einav and Finkelstein, 2011) while advantageous selection will induce a negative correlation (Fang et al., 2008). In consequence, if it appears that individuals' out-of-pocket cost forecasts are higher when insured, then this could be evidence of adverse selection rather than a behavioral bias - higher cost individuals are both more likely to opt into insurance and be diagnosed. Conversely, advantageous selection would cause forecasts of out-of-pocket costs to be lower when they are insured. Similarly, though individuals' insurance is likely to affect their reaction to unanticipated events, I report only the interaction between marital dissolution and whether individuals report being covered by their spouse's employer as evidence for the main mechanism by which divorce/separation/widowhood affect medical expenses forecasts. The coefficients on the unanticipated events indicators are then to be interpreted as a weighted average of the effect by insurance type.

Another caveat for interpreting the results regarding unanticipated diagnoses is that access to health insurance increases the probability that a latent condition is diagnosed, since the variables for diagnosis of diabetes and cancer in the HRS data are self-reported. In consequence, the estimated effects will not just be weighted averages with weights as a function of the share of individuals in the population with a particular insurance type. The weights will therefore also be a function of the ease and frequency with which individuals are screened for cancer and diabetes given their insurance (or lack thereof).

For completeness, I also include regressions with R_{it}, the level of realised out-of-pocket medical expenditures over the two years prior to period t for individual i, as the dependent variable. As discussed in the next subsection, these estimates will not be directly comparable to the forecasts made by individuals. They can nonetheless serve as a check on the accuracy of these forecasts

via a comparison of the relative sizes of the effects of different exogenous changes. I elaborate on this and other methods I use for assessing the accuracy of medical expenses forecasts in the next subsection. The equation for realisations nonetheless changes the timing of the Q_{itk} variables to reflect the fact that individuals with $Q_{it8} = 1$ qualified for Medicare two years prior, so that the estimating equation for the realisations in levels is

$$R_{it} = \alpha_0 + \alpha_1 1[Q_{itj} = 1, j \geq 10] + \sum_{k=-1}^{9} \beta_k Q_{itk} + \delta X_{it} + \sum_{s=2}^{3} \theta_s 1[t = s] + \sum_{s=0}^{1} \phi_s' Z_{it-s} + v_{it}, \quad (3.2)$$

Estimates from Equation 3.2 are presented in Appendix B. In the next subsection I discuss the barriers to directly comparing estimates Equation 3.2 with estimates from Equation 3.1 using the HRS data in the next subsection.

3.4.2 Assessing the Accuracy of Forecasts

The ideal data for assessing the accuracy of forecasts would have individuals' subjective cumulative distribution functions (CDFs) for the exact period in between survey waves, with the subsequent waves recording the actual outcomes realised over the same period for which the forecasts were made. Then we would be able to take the difference between the forecast and the outcome for each individual, and examine the distribution of errors across individuals.

Since the forecasts in the HRS data are made for the twelve months following the survey, and out-of-pocket medical expenditures are elicited for the preceding two years, the two measures found in the HRS are not directly comparable except under very strong assumptions. (There are only 46 person-year observations in which the delay in between consecutive survey years is 12 months or fewer. If these cases were more numerous, it would be possible to compare the forecast in the previous survey wave with the difference in realised two-year expenditures in between waves, as the latter could only capture differences that must have arisen within the twelve-month period for which the previous survey wave's forecast was made). The single exception to this is where events' effect on forecasts has the opposite sign to that of their effect on realised outcomes. In that case, individuals

are either mistaken regarding the first-order effect of a change on their circumstances or its persistent effects through time (i.e. if individuals' forecasts exhibit negative serial correlation following an innovation, whereas realisations are positively serially correlated following an innovation, then this in itself is inaccurate even if we cannot make precise statements regarding the magnitude of the error).

Instead, I adopt two strategies to assess the accuracy of the forecasts. Both rely on the auxiliary step of imputing implied changes to the mean and variance of a lognormal distribution resulting from the changes in percentiles that are recorded in the main results. The method for imputing this distribution is described in Appendix A, and is justified with reference to the common finding that medical expenditures distributions are well approximated by a lognormal distribution with a spike at 0 (French and Jones, 2004).

The first strategy relies on comparing the predicted changes to means and variances estimated in the main regressions to the estimates of actual changes for the same events from other studies. The key papers I will rely on are Barcellos and Jacobson (2015) for their estimates of the discontinuous changes in the out-of-pocket medical expenses distribution at the age of Medicare eligibility, Narang and Nicholas (2017) and Shrestha et al. (2019) for estimates of the effects of cancer and diabetes respectively on out-of-pocket medical expenditures.

The second strategy relies on comparing the relative magnitudes of the changes to individuals' predicted means and variance across events. For example, if an individual predicts that the dissolution of their marriage via separation, divorce or widowhood is likely to result in a similar shift in the mean of their medical expenditures to that following a cancer diagnosis, and the two have disparate effects on average realised medical expenditures, this will be taken as evidence of inaccuracy in updating their beliefs. Estimates provided in this chapter for changes to the realised distribution of medical expenses following unanticipated events will be useful for these comparisons alone. While they cannot be compared directly with the forecasts themselves, average changes in between waves relative to other innovations are comparable: since on average all these events occur at a similar remove from the survey period, if none of them are more likely to occur closer to the preceding or contemporaneous interview date, then their relative influence on average out-of-pocket

expenditures should occur over comparable periods of time. In consequence, individuals' forecasts' accuracy regarding the relative effects of events on the moments of the out-of-pocket expenditures distribution can be assessed by comparing them to the actual relative effects of these events on realised means and variances.

3.4.3 Translating Subjective Medical Expenses Forecasts into Changes in Subjective Means and Variances

The following result holds for log-normal random variables (Aitchison and Brown, 1957):

Theorem Let $Z \sim N(0,1)$. Then if X is log-normally distributed, such that $Y = log(X) \sim N(\mu, \sigma^2)$, then the following relationship holds between ψ_q and v_q, the quantiles of order q of X and Z:

$$\psi_q = exp(\mu + v_q\sigma)$$

Proof: See Aitchison and Brown (1957). ■

from this result it follows that, for a response that identifies the probability of exceeding a threshold z as p, that $z = \psi_{1-p} = exp(\mu + v_{1-p}\sigma)$; hence given two thresholds, z_{1-p_0} and z_{1-p_1} (e.g. \$1500 and \$3000 in the HRS data), we can solve two equations for the two unknowns, μ and σ, viz.

$$z_{1-p_0} = exp(\mu + v_{1-p_0}\sigma)$$
$$z_{1-p_1} = exp(\mu + v_{1-p_1}\sigma)$$

since v_{1-p_0} and v_{1-p_1} are the known $1 - p_0^{th}$ and $1 - p_1^{th}$ quantiles of the standard Normal distribution. This implies the system of linear equations

$$ln(z_{1-p_0}) = \mu + v_{1-p_0}\sigma$$

$ln(z_{1-p_1}) = \mu + v_{1-p_1}\sigma$

which has the general solutions $\widehat{\sigma} = ln\left(\frac{z_{1-p_0}}{z_{1-p_1}}\right)/(v_{1-p_0} - v_{1-p_1})$ and $\widehat{\mu} = ln(z_{1-p_1}) - \frac{v_{1-p_1}}{(v_{1-p_0} - v_{1-p_1})} ln\left(\frac{z_{1-p_0}}{z_{1-p_1}}\right)$.

For example, suppose an individual forecasts a 97.5% chance of their medical expenses exceeding \$1500 and a 2.5% chance of their expenses exceeding \$3000. We know that the 2.5th percentile of the standard Normal distribution is -1.96, and that its 97.5th percentile is 1.96. So we obtain

$1500 = exp(\mu - 1.96\sigma)$,

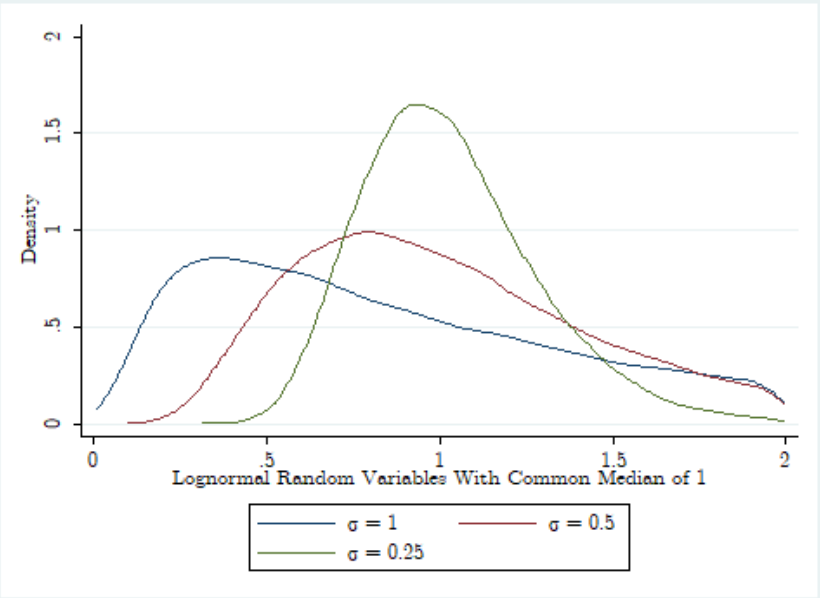
$3000 = exp(\mu + 1.96\sigma)$,

which yields $\mu = 7.66$ and $\sigma = 0.177$. To obtain the implicit forecast of the mean of out-of-pocket medical expenditures, we can simply take $exp(7.66 + \frac{0.177^2}{2}) = \$2155.26$. To obtain the 95^{th} percentile, we can take $\psi_{0.95} = exp(\mu + v_{0.95}\sigma) = \psi_{0.95} = exp(7.66 + 1.645 * 0.177) = \$2838.88$. As expected, the 95^{th} percentile lies below the 97.5^{th} percentile, and the distribution is fairly tightly centered. The variance is obtained from the formula $(exp(\sigma^2) - 1) \times exp(2\mu + \sigma^2)$.

For those who answered "0%" at the \$1500 threshold, I assume a rounding interpretation so as to use their subsequent answer at the \$500 threshold. (Those who answer "0%" at both the \$1500 and \$500 threshold are dropped from the analysis). So I take them to mean that they have a 99% chance of their expenses lying below \$1500.

Plugging in the estimates from the changes in forecasts in the main text, we can recover the implied effects on both mean expenditures and the variance of out-of-pocket medical expenditures by holding one quantile constant at a modal value for that quantile (say, 50% for \$1500 - implying that it is the median level of future expenditures) while changing the quantile to which another threshold pertains by a number of percentage points obtained in the empirical exercises. The main implications of doing so will be that the events examined in the main text - with the exception of receiving a cancer diagnosis - have modest effects on the implied subjective mean expenditures faced by individuals, but substantial effects on the implied subjective variances of

Figure 3.3: Probability Distribution Functions for Three Lognormal Distributions, Each with $\mu = 0$, and $\sigma = 1$, 0.5 and 0.25.



Notes: The three lognormal distributions show how holding one quantile constant and changing another quantile of a lognormal random variable affects the distribution. All three distributions in the figure have the same median, $e^{\mu} = 1$ and hence the common parameter $\mu = 0$. However their means and variances are then $e^{0.5\sigma^2}$ and $e^{\sigma^2}(e^{\sigma^2} - 1)$, respectively, both of which increase as σ increases due to the rightward skew of the lognormal distribution. Intuitively, the further away the mean is from the median, the more skewed and dispersed is the distribution. This illustrates how decreasing any quantile of the lognormal distribution while holding another quantile constant decreases both its mean and its variance, while increasing a quantile while holding another quantile constant does the opposite. This explains why the changes to the mean and the variance summarised in Table 3.6 are in the same direction - either both increase or both decrease.

future expenditures, indicating that individuals take the changes in their risk of large expenditures to be first-order rather than the changes in their expected expenditures.

Changing percentiles of a distribution necessarily affects the entire distribution - in the case of the lognormal distribution, the first two moments depend on both μ and σ^2, and so each changes if one of them changes. If the median of the lognormal distribution e^{μ} changes, then this necessarily implies a change in both the mean and the variance of the lognormal distribution. This is in contrast with a Normal distribution where location and scale changes are independent of one another. Since the third central moment is always positive for a lognormal distribution, it is always right-skewed; as a result, the only way for one of the quantiles to change while another quantile remains the same is to shift mass into or out of the long right tail, thereby either increasing or decreasing the variance of the distribution, even if its median remains the same (say). Figure 3.3 plots the probability distribution functions of three lognormal distributions with the same median of $e^{\mu} = 1$ but different values for every other quantile. This illustrates how it is impossible to change one quantile, while holding another quantile constant, without affecting the shape of the entire rest of the distribution.

In the following section, I examine the consequences of both anticipated and unanticipated changes in individuals' insurance and health status for their medical expenses forecasts.

3.5 Results

In this section, I examine the consequences for medical expenses forecasts of (1) anticipated events, primarily qualifying for health insurance via the Medicare program at age 65, and (2) unanticipated permanent changes to lifetime medical expenditures, such as the onset of diabetes, cancer, or marital dissolution. Before doing so, I document a stylised fact that provides evidence that subjective medical expenses forecasts may be systematically inaccurate.

3.5.1 Incoherent Subjective Forecasts of Out-of-Pocket Medical Expenditures

Table 3.3: Correlates of Reporting Forecast Probabilities that Imply Decreasing Cumulative Distribution Functions: Probit Estimates

	(1)		(2)
White	-0.30***	Medicaid	0.02
	(-4.44)		(0.13)
Male	-0.13*	Medicare	0.10
	(-2.19)		(0.74)
High School Grad.	-0.11	Own Employer	0.03
	(-1.44)		(0.37)
College Grad.	-0.39***	Spouse's Employer	0.13
	(-3.95)		(1.33)
Poor Health	0.01	Union	0.12
	(0.10)		(1.54)
Age	0.05	VA/CHAMPUS	0.03
	(1.32)		(0.18)
Age^2	-0.00	Cognition	-0.05***
	(-1.15)		(-5.97)
ln(HHInc)	-0.09**		-
	(-2.65)		-
N	4411		4411

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Results are from a single Probit regression with all the listed variables on the right hand side and an indicator variable for whether an individual's responses to the subjective medical expenses forecasts questions implied an increasing CDF for future medical expenses.

For 6.9% of the sample, the answers given violate the laws of probability: individuals will

assign higher probabilities to spending exceeding higher thresholds, which is impossible, since (for example) any event where an individual spends more than $3000 is also an event where they spend more than $1500. This is an instance of what Tversky and Kahneman (1983) called "The Conjunction Fallacy" (see above): subjective beliefs that entail the belief that, for some events A and B, $Pr(A \cap B) > max\{Pr(A),\ Pr(B)\}$, which is impossible. The rate at which this occurs is in line with other survey questions that elicit subjective CDFs. Dominitz and Manski (1997), using subjective forecasts of future income, find that only 22 of 437 respondents (5.03%) to their questions on the likelihood of different realisations of household income give answers that imply a nonincreasing CDF. The incidence of incoherence in responses to this question in the HRS appears to be twice as high as in the worst previous scenario recorded in the literature.

Table 3.3 reports the results of a probit regression of a dummy variable for whether an individual is a coherent forecaster. By far the largest effects come from the educational gradient: high school dropouts, high school graduates, and those who did not complete college are all more likely to violate the laws of probability in their forecasts than are college graduates. Even conditional on education, there is a separate effect of cognition, with the results from the cognitive test score testing individuals on their fluid working memory, long-term memory, and numeracy. Even conditional on these characteristics, poorer individuals and people of colour are more likely to make incoherent forecasts, which probably reflects differences in early life circumstances that persist into adulthood.

Overall, it appears that those who are more vulnerable in other settings - high school dropouts, minorities, and poorer individuals - are more likely to be unable to make coherent forecasts of their future medical expenses. Part of this effect may be due to the coverage provided by the state - those who bear a lower proportions of the costs of their treatment may be "rationally inattentive" in that the return to an accurate forecast of their future out-of-pocket medical expenses is lower, and hence invest less time and effort into making accurate forecasts (though I find little evidence of an additional correlation between enrolment in Medicaid and coherence). Unsurprisingly, lower cognitive ability makes incoherent forecasts more likely - though it is unclear what role fluid memory plays in making this so over and above numeracy. I leave it to future research to determine the role of memory in forming expectations regarding the future. These explanations both rely on

the assumption that cognitive resources are scarce and hence agents' behaviour is explicable as maximisation subject to their cognitive resource constraint Borghans et al. (2008). In principle, this could be modelled as a first stage before the economic decisions of the agent are made where the agent decides upon the optimal allocation of cognitive resources (cf. Matejka and McKay (2014)).

The problem is that with mathematically incoherent beliefs we cannot go further in deriving forecasts of the distribution of medical expenses. (Dominitz and Manski collected their data themselves and prompted individuals whose CDFs violated monotonicity to revise their answers). What can be done is to examine the extent to which the rest of the sample exhibits *systematic* deviations from rationality, as opposed to mere incoherence, which might be informative regarding their behaviour. For instance, agents who are overly optimistic regarding their future medical expenses will tend to under-save relative to their eventual needs. This motivates testing the accuracy of these forecasts in other aspects. The fact that many individuals struggle to work with subjective probabilities suggests that there may be inaccuracies in the forecasts of the remainder who do manage to report nondecreasing CDFs.

3.5.2 Revisions to Forecasts Due to Anticipated Receipt of Health Insurance

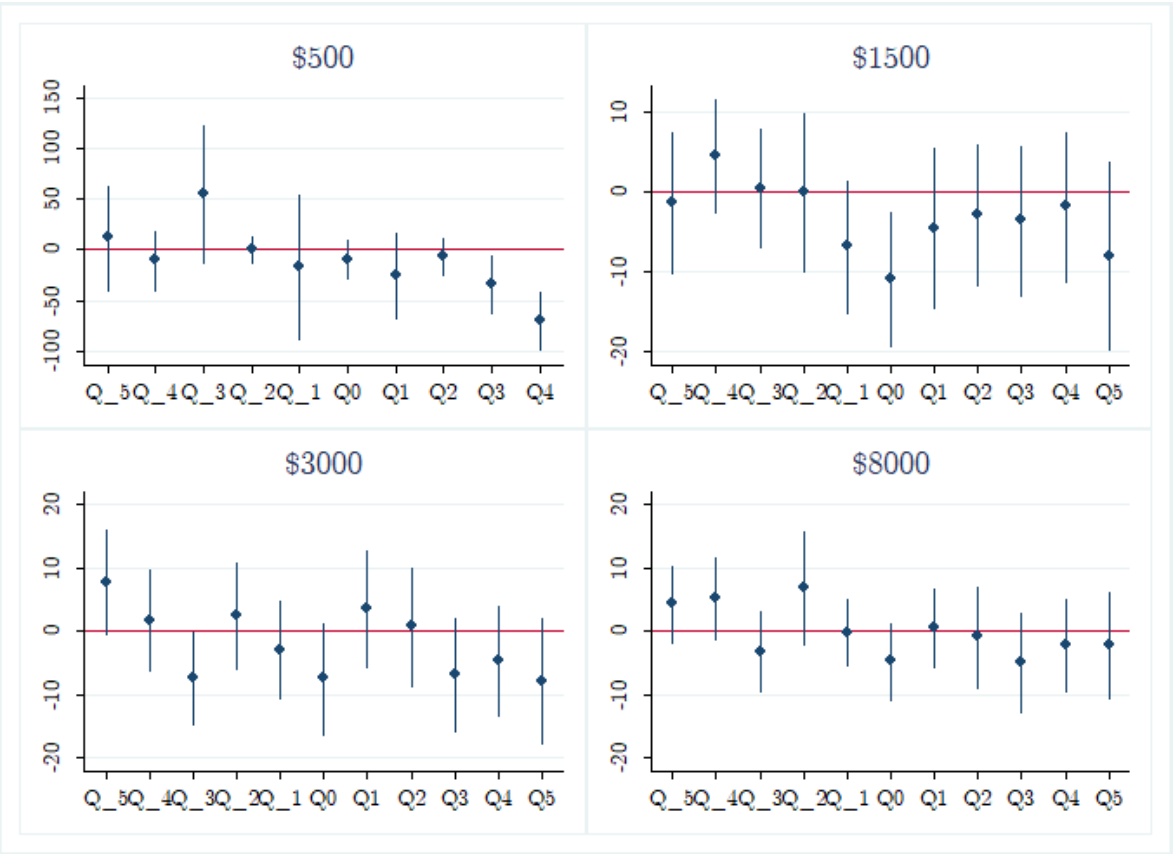
First, I examine the impact of the anticipation of and eligibility for Medicare coverage on the individuals' forecast probabilities for different thresholds of health care spending. Any revision to forecasts as a result of the anticipated change in insurance status, since it is already taken into account *ex ante*, will be due to the sequential resolution of uncertainty. Even if individuals have the correct distribution of probabilities over adverse health events that require them to pay out of pocket, they are likely to be unable to predict the exact timing of necessary medical expenditures. Regarding discretionary medical expenditures, individuals may choose to delay spending on prescription drugs until they are cheaper, for instance (as Medicare Part D is available to subsidise prescription drugs for over-65s over the entire sample period) but if individuals smooth their expenditures across time then this should not alter forecast medical expenditures. Hence any change in the forecast leading up to the age 65 threshold is attributable to the sequential resolution of

uncertainty conditional on the assumption that individuals can smooth consumption across periods. An observation that would imply that individuals do not smooth consumption across periods would be an expected increase in out-of-pocket expenditures upon qualifying for Medicare, since this would indicate individuals' medical expenditures are tracking their disposable income instead of being smoothed across periods.

Figure 3.4 records the consequences of qualifying for Medicare on subjective medical expenditures forecasts. There appear to be no statistically significant changes leading up to the quarter of Medicare eligibility, which is surprising since there should be significant differences both in planned expenditures and the risk of incurring large out-of-pocket medical expenditures between individuals who are eligible 10-12 months from the survey period and individuals who are 1-3 months away from qualifying for Medicare when surveyed. There is nonetheless a large and statistically significant drop of nearly 10 percentage points in the average probability of incurring at least $1500 in out-of-pocket expenditures in the quarter of qualifying for Medicare. This is nonetheless quite a modest change overall. In Appendix A, I show that if we assume a lognormal distribution for medical expenditures, this translates into an expected decrease of $111 and a relative reduction in the variance of medical expenses of 12%. This latter conclusion implies much less protection from medical expenses risk than is found by Barcellos and Jacobson (2015), who document a 53% reduction in the 95^{th} percentile of out-of-pocket medical expenses that occurs discontinuously upon qualifying for Medicare. Figure 3.5 in Appendix B, which shows a significantly larger decrease in medical expenditures (even after dividing by 2 to obtain annual mean expenditures) of approximately $400, along with the disparity with the results in Barcellos and Jacobson (2015), supports the conclusion that individuals underestimate how much protection from medical expenditures risk is provided to them by Medicare.

Two main lessons come out of the empirical exercises in this subsection. First, individuals are slow to account for the sequential resolution of uncertainty when forecasting their future out-of-pocket medical expenditures. Revisions to forecasts are too slow to be consistent with a fine-grained approach to the sequential resolution of uncertainty. This is consistent with the results regarding subjective mortality expectations found in Elder (2013), who found that individuals do not take

Figure 3.4: Changes to Forecast Probabilities at Different Lags and Leads from Qualifying for Medicare Coverage



Notes: The dependent variable in each panel is the subjective forecast of the probability, from 0% to 100% that the following twelve months' medical expenses will exceed the number on the x-axis. Each plotted coefficient represents the coefficient from a dummy variable indicating how many quarters from Medicare eligibility a respondent is, pooling together data from the 2010-2014 waves of the Health and Retirement Study (HRS). Q_0 corresponds to the first quarter of Medicare eligibility, i.e. the individual is in the month of qualifying at the time of the survey or one of the two months following. Error bars correspond to the 95% confidence intervals for each coefficient. All specifications include a set of dummy variables for demographic, socioeconomic and insurance status (excluding Medicare eligibility since it is multicollinear with the coefficients of interest), as well as survey year dummies to absorb period effects common to all respondents. Standard errors are clustered at the individual level.

into account the fact that they are likelier to survive to a given age a if they have already survived to $a-m$ than if they have survived to $a-n$, where $m < n$.

Second, individual forecasters do not seem to be skilled at estimating their risk of incurring large medical bills. This is supported by the disparity between existing estimates of the extent of medical expenditures risk provided by Medicare and the decrease in the variance upon qualifying for Medicare implied by the estimates above. As a result, they may undervalue insurance for its role in reducing the probability of costly tail events. Self-insurance decisions are also likely to be distorted accordingly, leading individuals to overcompensate for their perceived lack of insurance.

In the next subsection, I turn to the consequences of unanticipated exogenous events such as diagnoses of persistent illnesses, and marital dissolution through either widowhood or divorce for both subjective medical expenditures forecasts and realised out-of-pocket medical expenditures.

3.5.3 Revisions to Forecasts Due to Unanticipated Persistent Changes

Next I turn to the effect of unanticipated permanent changes to individuals' health care costs on their twelve-month forecasts. I focus on diagnoses of diabetes, cancer, and marital dissolution. Since the exact date of onset of each of these is unlikely to be anticipated in advance by individuals, they are not in individuals' information set *ex ante* and hence ought to produce innovations in individuals' medical expenses forecasts. The results in here record estimates of elements of the parameter vector ϕ from Equation 3.1.

As a placebo test for the effect of diagnosis of a persistent condition, I examine changes in marital status (divorce or widowhood) for individuals who rely on their spouse's employer for their health insurance. A key piece of legislation that makes this a placebo test is the COBRA Act of 1985 (Madrian, 1998), which mandates employers to maintain enrolment for spouses of employees who were covered by their health insurance policy prior to their divorce or the death of the employee for up to 36 months. As a result, while these are likely to be unexpected transitions two years in advance (and the HRS data is collected every two years), they should not result in any change in forecast medical expenses for the subsequent 12-month period. This is because even if the marital transition happened 24 months prior to the survey date, the spouse that would otherwise not have

coverage still has 12 months left of enrolment on their former spouse's employer's health insurance plan. Any changes to medical expenses forecasts following a marital transition would therefore be attributable to ignorance of the institutional environment (as documented by Harris and Keane (1998) and used as an explanation for the empirical results in Fang et al. (2008)). If medical expenses forecasts are revised upwards, in particular, then this suggests that part of individuals' response to unanticipated shocks is exaggerated due to incomplete accounting for the resources available with which to cope with adverse events.

The results are displayed in Table 3.4. Individual forecasts of future medical expenditures react strongly to diagnoses of cancer and marital dissolution, but there is little evidence of updating of forecasts following a diagnosis of diabetes. As discussed above, marital dissolution should have no effect on medical expenses forecasts, but has a large impact on subjective medical expenses forecasts. In fact (see below) assuming a lognormal distribution for out-of-pocket medical expenditures implies that individuals revise their expected expenditures upward to a smaller degree following a diagnosis of cancer (though not by a great deal in relative terms), and also expect the variance of their medical expenditures to increase by nearly three times as much in relative terms following either widowhood or divorce. Accompanying results in both Tables 3.4 and 3.5 support the conclusion that this is due to these individuals being unaware of the law that entitles them to stay on their former spouse's employer-based health insurance. When an interaction term between whether an individual reports being covered by their spouse's employer is included (Table 3.4), a Wald test cannot reject the null hypothesis that the effect of marital dissolution is zero for those individuals who report being covered by their spouse's employer. Moreover, Table 3.5 records a significant decrease of up to 5 percentage points in the probability that individuals report being covered by their spouse's employer - an effect that *increases* in magnitude when the delay in between survey waves is restricted to 24 months or less, a period in which individuals are mandated by law to be retained by their employer's health insurance plan. The lack of response to the diagnosis of diabetes is likely due to many individuals being able to manage their diabetes in the early stages of diagnosis without the use of prescription drugs or medical equipment. More puzzling is the *decrease* in medical expenses forecasts of the risk of exceeding \$1500 or \$3000 in out-of-pocket expenditures

Table 3.4: Changes in Medical Expenditures Forecast Probabilities Following Diagnosis of Cancer or Diabetes, or Marital Dissolution

	(1) $Pr(>= 500)$	(2) $Pr(>= 1500)$	(3) $Pr(>= 3000)$	(4) $Pr(>= 8000)$
Cancer in $(t-1, t)$	15.82	8.18*	7.11	6.13
	(1.11)	(2.19)	(1.87)	(1.82)
Diabetes in $(t-1, t)$	-2.20	4.50	3.95	1.90
	(-0.10)	(1.49)	(1.43)	(0.80)
Marital in $(t-1, t)$	22.45**	-1.20	-2.51	-2.66
	(2.64)	(-0.46)	(-0.94)	(-1.28)
Marital in $(t-1, t) \times$ Spousal	-22.93*	-1.80	-1.18	2.92
	(-2.05)	(-0.15)	(-0.13)	(0.56)
Cancer in $(t-2, t-1)$	-40.85*	5.55	-2.17	0.61
	(-2.25)	(1.40)	(-0.65)	(0.22)
Diabetes in $(t-2, t-1)$	-2.84	-5.72*	-6.31*	-3.63
	(-0.27)	(-2.27)	(-2.23)	(-1.48)
Marital in $(t-2, t-1)$	0.26	0.46	3.08	1.47
	(0.03)	(0.14)	(0.81)	(0.39)
Marital in $(t-2, t-1) \times$ Spousal	8.25	-14.25	-14.16	-4.61
	(0.52)	(-1.51)	(-0.73)	(-0.33)
N	444	8016	6217	4964

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level.

in the survey year following the one immediately after a diagnosis of diabetes. This is directly contrary to the evidence (Shrestha et al., 2019).

Some tentative conclusions can be drawn from the implied mean changes from the forecasts, together with the relative extent of the changes due to Individuals' forecasts for the 12 months following a cancer diagnosis indicate an increase in mean out-of-pocket expenditures of only $96 for what is on average their second year following diagnosis. By contrast, in the most recent study of

Table 3.5: Changes in Reporting of Being Covered By Spouse's Employer and Household Income Following Divorce and/or Widowhood

	(1) △Spouse's Employer	(2) (1), ≤ 24 Months Between Waves	(3) △ ln(HHInc)
Marital Dissolution	-0.05**	-0.06**	-0.50***
	(-2.93)	(-2.82)	(-8.21)
Medicaid	-0.03***	-0.02*	0.05
	(-3.48)	(-2.41)	(0.76)
Medicare	-0.05***	-0.04***	0.03
	(-5.44)	(-3.81)	(1.33)
Own Employer	-0.06***	-0.05***	0.05**
	(-10.87)	(-7.84)	(2.71)
Spouse's Employer	-	-	0.05*
			(2.56)
Union	0.01*	0.01	-0.01
	(2.57)	(1.55)	(-0.81)
VA/CHAMPUS	-0.03*	-0.02	0.04
	(-2.54)	(-1.68)	(1.23)
N	11750	8649	11764

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level..

out-of-pocket costs following cancer diagnosis, Narang and Nicholas (2017) use the same data (the HRS) calculate that medical expenditures on average in the first year following cancer diagnosis (averaging across likely periods of diagnosis in between survey waves) are approximately $2116 even among those with the most generous insurance coverage (Medicaid beneficiaries). One explanation could be that chemotherapy is undertaken relatively quickly following a cancer diagnosis, and this is responsible for the majority of out-of-pocket expenditures on cancer treatment.

The biggest puzzle regarding the unanticipated changes is the lack of response of the forecasts to diagnoses of diabetes, which are known to increase out-of-pocket expenditures at an increasing

rate the longer individuals live with a diagnosis (Shrestha et al., 2019). Though in the immediate wake of the diagnosis this could be attributed to individuals foregoing preventive care that their diagnosis necessitates, this ought to lead to a higher risk of out-of-pocket medical expenditures later on, which is not reflected in the forecasts.

Marital dissolution does not affect realised expenditures in the first survey wave after it occurs (Table 3.13 in Appendix B). By contrast, the relative impact on forecasts implies that individuals think they take on more risk following marital dissolution than following cancer diagnosis, and expect to spend more on average following that marital dissolution than a cancer diagnosis (Table 3.6 in Appendix A). One explanation for this latter result may be that most of the medical expenses risk is resolved shortly after the diagnosis of cancer and individuals discover what their treatment plan and options for coverage are, whereas if individuals think (mistakenly) that they have lost their spouse's employer's health insurance, then this is a more persistent source of medical expenses risk. Hence to the extent people are mistaken about the relative impacts of the two events, we cannot rule out that they would be accurately updating their information had they been correct about whether or not they had lost insurance from their spouse's employer.

3.5.4 Implied Changes to Means and Variances of Subjective Medical Expenses Distributions

Table 3.6 shows the implied effects on mean expenditures obtained by holding one quantile constant while changing another. Since each of the \$500 threshold answers was elicited from survey respondents who answered that they had a "0%" chance of incurring \$1500 or more of out-of-pocket medical expenditures in the 12 months following, I take \$1500 as the 99^{th} percentile of a distribution that faces a change in which quantile is \$500 for changes at the \$500 threshold. For changes to responses to the other subjective medical expenses forecast questions, I take the lower quantile as the median (the modal response for most questions) and the upper quantile as the 99^{th} percentile (corresponding to another modal response of "0%" for higher thresholds even when the answer to the \$1500-threshold question is higher than 0%). The results have been referred to above, and are consistent with the conclusions that individuals underestimate the extent to which Medicare

Table 3.6: Implied Changes in the Mean and Variance of Lognormally Distributed Medical Expenses From Changes in Main Text

	(1) △Mean ($)	(2) △Variance (%)
Qualifying for Medicare	-110.53	-11.60
Cancer Diagnosis	96.49	10.84
Marital Dissolution	100.07	29.08

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: All estimates come from changes to the mean and variance of a lognormal distribution of out-of-pocket medical expenditures. All changes are implied by the changes to the percentiles of the subjective medical expenditures distribution from the events listed on the left hand column of the table. Details of the method for translating changes in quantiles of a lognormal distribution to changes in its moments can be found in Appendix A.

protects them from medical expenditure risk and cancer exposes them to medical expenditures risk, while they overestimate the increase in medical expenditures risk posed by being separated, divorced or widowed (which should be zero according to the protections afforded under law in the United States).

The results depend on only a single quantile changing at once. This fits with the results obtained in the main text, which typically find only one quantile displaying a statistically significant change for a given exogenous event. In consequence, for individuals who have an implied 1% chance of incurring medical expenditures that exceed the next-highest threshold (automatically true for all respondents to the $500 question, who all answer "0%" to the $1500 threshold question), the changes to their subjective means are modest compared to changes in their subjective variances. This is also driven by the leftward skew of the lognormal distribution. The largest changes are for marital dissolution, which can be rationalised if individuals perceive that employer-provided

coverage provides more protection from medical expenses risk on average than does Medicare. This would explain why individuals mistakenly believing they have lost their spouse's employer-provided coverage has a larger effect on the variance of their out-of-pocket medical expenditures than does qualifying for health insurance via Medicare.

3.6 Conclusion

This chapter has provided the first examination of the accuracy of medical expenses forecasts and the separate effects of the sequential resolution of uncertainty and the revelation of new information on those forecasts. Future work will focus on examining whether the conclusions of previous empirical analyses that assumed rational expectations are significantly altered by accounting for individuals' subjective forecasts of their future medical expenses *ex ante*. The first finding of this chapter is that a large proportion of individuals are unable to make self-consistent forecasts of their future medical expenditures distributions. The second main finding - that individuals do not react to the sequential resolution of uncertainty - could be rationalised with Epstein-Zin preferences (Epstein and Zin, 1991), where individuals treat the sequential resolution of uncertainty differently to the overall uncertainty they face at the beginning of a dynamic decision problem. So far, though widely used in finance, Epstein-Zin preferences have not, to my knowledge, been used in studies of life-cycle consumption that incorporate medical expenditures risk. The third main finding of ignorance regarding entitlement to continued insurance coverage after the end of a marriage is easiest to incorporate into standard models, and can be used to examine how individuals' self-insurance, risk-taking and health investment behaviors differ from what they would have been had individuals taken their continued access to insurance into account.

References

Aitchison, J. and Brown, J. A. (1957). The Lognormal Distribution with Special Reference to its Uses in Economics.

Aizer, A. (2003). Low Take-Up in Medicaid: Does Outreach Matter and For Whom? *American Economic Review*, 93(2):238–241.

Barcellos, S. H. and Jacobson, M. (2015). The Effects of Medicare on Medical Expenditure Risk and Financial Strain. *American Economic Journal: Economic Policy*, 7(4):41–70.

Benitez-Silva, H. and Dwyer, D. S. (2005). The Rationality of Retirement Expectations and the Role of New Information. *Review of Economics and Statistics*, 87(3):587–592.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 119(1):249–275.

Bhargava, S. and Manoli, D. (2015). Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review*, 105(11):3489–3529.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Ter Weel, B. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, 43(4):972–1059.

Cameron, A. C. and Miller, D. L. (2015). A Practitionerâs Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372.

De Nardi, M., French, E., and Jones, J. B. (2010). Why Do the Elderly Save? The Role of Medical Expenses. *Journal of Political Economy*, 118(1):39–75.

De Nardi, M., French, E., and Jones, J. B. (2016a). Medicaid Insurance in Old Age. *The American Economic Review*, 106(11):3480–3520.

Dominitz, J. and Manski, C. F. (1997). Using Expectations Data to Study Subjective Income Expectations. *Journal of the American Statistical Association*, 92(439):855–867.

Einav, L. and Finkelstein, A. (2011). Selection in Insurance Markets: Theory and Empirics in Pictures. *Journal of Economic Perspectives*, 25(1):115–38.

Elder, T. E. (2013). The Predictive Validity of Subjective Mortality Expectations: Evidence from the Health and Retirement Study. *Demography*, 50(2):569–589.

Epstein, L. G. and Zin, S. E. (1991). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis. *Journal of Political Economy*, 99(2):263–286.

Fang, H., Keane, M. P., and Silverman, D. (2008). Sources of Advantageous Selection: Evidence from the Medigap Insurance Market. *Journal of Political Economy*, 116(2):303–350.

French, E. and Jones, J. B. (2004). On the Distribution and Dynamics of Health Care Costs. *Journal of Applied Econometrics*, 19(6):705–721.

French, E. and Jones, J. B. (2011). The Effects of Health Insurance and Self-Insurance on Retirement Behavior. *Econometrica*, 79(3):693–732.

Gan, L., Hurd, M. D., and McFadden, D. L. (2005). Individual Subjective Survival Curves. In *Analyses in the Economics of Aging*, pages 377–412. University of Chicago Press.

Giustinelli, P., Manski, C. F., and Molinari, F. (2018). Tail and Center Rounding of Probabilistic Expectations in the Health and Retirement Study. Technical report, National Bureau of Economic Research.

Giustinelli, P., Manski, C. F., and Molinari, F. (2019). Precise or Imprecise Probabilities? Evidence from Survey Response on Late-Onset Dementia. Technical report, National Bureau of Economic Research.

Harris, K. M. and Keane, M. P. (1998). A Model of Health Plan Choice:: Inferring Preferences and Perceptions from a Combination of Revealed Preference and Attitudinal Data. *Journal of Econometrics*, 89(1-2):131–157.

Hurd, M. D. (2009). Subjective Probabilities in Household Surveys. *Annual Review of Economics*, 1(1):543–562.

Hurd, M. D. and McGarry, K. (1995). Evaluation of the Subjective Probabilities of Survival in the Health and Retirement Study. *Journal of Human Resources*, pages S268–S292.

Hurd, M. D. and McGarry, K. (2002). The Predictive Validity of Subjective Probabilities of Survival. *Economic Journal*, 112(482):966–985.

Inoue, A., Kilian, L., and Kiraz, F. B. (2009). Do Actions Speak Louder Than Words? Household Expectations of Inflation Based on Micro Consumption Data. *Journal of Money, Credit and Banking*, 41(7):1331–1363.

Jappelli, T. and Pistaferri, L. (2000). Using Subjective Income Expectations to Test for Excess Sensitivity of Consumption to Predicted Income Growth. *European Economic Review*, 44(2):337–358.

Keane, M. and Stavrunova, O. (2011). A Smooth Mixture of Tobits Model for Healthcare Expenditure. *Health Economics*, 20(9):1126–1153.

Madrian, B. C. (1998). Health Insurance Portability: The Consequences of COBRA. *Regulation*, 21:27.

Manski, C. F. and Molinari, F. (2010). Rounding Probabilistic Expectations in Surveys. *Journal of Business & Economic Statistics*, 28(2):219–231.

Matejka, F. and McKay, A. (2014). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, 105(1):272–298.

Narang, A. K. and Nicholas, L. H. (2017). Out-of-Pocket Spending and Financial Burden Among Medicare Beneficiaries With Cancer. *JAMA Oncology*, 3(6):757–765.

Shrestha, S. S., Zhang, P., Hora, I. A., and Gregg, E. W. (2019). Trajectory of Excess Medical Expenditures 10 Years Before and After Diabetes Diagnosis Among US Adults Aged 25–64 Years, 2001–2013. *Diabetes Care*, 42(1):62–68.

Souleles, N. S. (2004). Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys. *Journal of Money, Credit and Banking*, pages 39–72.

Tversky, A. and Kahneman, D. (1983). Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4):293.

Appendix A: Descriptive Differences in Realised Medical Expenditures and Health Care Utilisation

Table 3.7: Correlates of 2-Year Realised Medical Expenditures: Demographic Characteristics, Threshold Forecasts × 2

	(1) ($OOP \geq 1000$)	(2) ($OOP \geq 3000$)	(3) ($OOP \geq 6000$)	(4) ($OOP \geq 16000$)
White	0.08***	0.06***	0.03***	0.01***
	(7.83)	(7.37)	(5.83)	(3.50)
Male	-0.07***	-0.04***	-0.02***	-0.00
	(-7.79)	(-4.71)	(-3.45)	(-0.05)
High School Grad.	0.05***	0.02*	0.01*	0.00
	(4.21)	(2.38)	(2.07)	(0.83)
College Grad.	0.09***	0.06***	0.03***	0.01
	(6.45)	(4.76)	(3.97)	(1.42)
Poor Health	0.16***	0.18***	0.11***	0.03***
	(15.40)	(17.02)	(13.62)	(7.08)
Age	0.03***	0.01*	-0.00	-0.00
	(5.34)	(2.30)	(-0.13)	(-0.58)
Age^2	-0.00***	-0.00	0.00	0.00
	(-3.65)	(-1.25)	(0.84)	(0.99)
ln(HHInc)	0.04***	0.03***	0.01***	0.00
	(8.15)	(6.56)	(4.26)	(1.87)
N	19801	19801	19801	19801

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for source of insurance. Each column's dependent variable is an indicator variable for whether the preceding two years' out-of-pocket medical expenditures exceeded the figure in the column header.

Table 3.8: Correlates of 2-Year Realised Medical Expenditures: Insurance Dummies, Threshold Forecasts × 2

	(1) ($OOP \geq 1000$)	(2) ($OOP \geq 3000$)	(3) ($OOP \geq 6000$)	(4) ($OOP \geq 16000$)
Medicaid	-0.20***	-0.09***	-0.04**	-0.02***
	(-9.05)	(-4.96)	(-3.14)	(-4.31)
Medicare	0.01	-0.01	-0.02	-0.01*
	(0.77)	(-0.56)	(-1.56)	(-2.05)
Own Employer	0.09***	0.05***	0.02**	-0.00
	(8.83)	(5.59)	(2.75)	(-1.26)
Spouse's Employer	0.04**	0.01	-0.00	-0.00
	(2.96)	(1.19)	(-0.28)	(-0.41)
Union	-0.07***	-0.05***	-0.02***	-0.00
	(-5.69)	(-5.49)	(-3.78)	(-1.07)
VA/CHAMPUS	-0.11***	-0.09***	-0.03**	-0.00
	(-4.90)	(-5.21)	(-2.86)	(-0.70)
N	19801	19801	19801	19801

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for sex, race, education (via indicator variables for whether an individual has graduated high school (incl. GED holders) or graduated from a four-year college), a quadratic in age, and the natural log of household income.

Table 3.9: Correlates of 2-Year Realised Medical Expenditures: Demographic Characteristics, Extensive Margin, Average $95th$ Percentile

	(1) $(OOP > 0)$	(2) OOP	(3) 95^{th} Percentile
White	0.06***	548.58***	1757.20***
	(9.23)	(6.17)	(4.57)
Male	-0.05***	-316.92***	-892.84**
	(-8.97)	(-3.95)	(-2.60)
High School Grad.	0.07***	130.74	113.03
	(7.47)	(1.02)	(0.23)
College Grad.	0.10***	418.89**	946.53
	(9.26)	(2.87)	(1.63)
Poor Health	0.02**	1792.00***	5212.65***
	(3.10)	(13.09)	(10.88)
Age	0.01	23.88	-60.24
	(1.69)	(0.43)	(-0.28)
Age^2	-0.00	0.25	1.69
	(-0.39)	(0.51)	(0.92)
ln(HHInc)	0.03***	278.67***	638.35*
	(8.66)	(5.22)	(2.49)
N	19801	19801	19801

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for source of insurance.

Table 3.10: Correlates of 2-Year Realised Medical Expenditures: Insurance Dummies, Extensive Margin, Average $95th$ Percentile

	(1) $(OOP > 0)$	(2) OOP	(3) 95^{th} Percentile
Medicaid	-0.20***	-1121.13***	-2584.14*
	(-8.16)	(-7.49)	(-2.35)
Medicare	0.01	-327.92	-1747.29*
	(1.01)	(-1.94)	(-2.32)
Own Employer	0.08***	305.99**	454.01
	(11.85)	(3.24)	(1.05)
Spouse's Employer	0.08***	94.09	-86.63
	(10.04)	(0.68)	(-0.15)
Union	-0.03***	-409.29***	-1330.47**
	(-3.87)	(-4.23)	(-2.91)
VA/CHAMPUS	-0.05***	-715.23***	-1355.94
	(-3.37)	(-4.11)	(-1.48)
N	19801	19801	19801

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for sex, race, education (via indicator variables for whether an individual has graduated high school (incl. GED holders) or graduated from a four-year college), a quadratic in age, and the natural log of household income.

Table 3.11: Correlates of Use of Medical Services: Demographic Characteristics

	(1) Hospital Stay?	(2) # Stays	(3) # Hospital Nights	(4) Doctor Visit?	(5) # Doctor Visits
White	0.02***	0.04***	0.00	0.02**	0.72*
	(3.44)	(3.86)	(0.02)	(2.73)	(2.19)
Male	0.01	0.01	0.21**	-0.07***	-1.50***
	(1.02)	(1.26)	(3.20)	(-12.05)	(-5.78)
High School Grad.	0.03***	0.06***	0.08	0.09***	1.76***
	(3.78)	(3.73)	(0.61)	(8.98)	(4.02)
College Grad.	0.00	0.03	0.04	0.11***	2.42***
	(0.20)	(1.46)	(0.28)	(10.68)	(5.43)
Poor Health	0.15***	0.32***	1.15***	0.03***	6.36***
	(16.57)	(13.69)	(9.03)	(4.30)	(11.72)
Age	-0.02***	-0.03***	-0.07*	0.00	-0.11
	(-4.33)	(-3.35)	(-2.13)	(0.66)	(-0.76)
Age^2	0.00***	0.00***	0.00*	0.00	0.00
	(4.85)	(3.67)	(2.38)	(0.42)	(1.66)
ln(HHInc)	0.01*	0.00	0.01	0.04***	0.60***
	(2.17)	(0.58)	(0.17)	(11.68)	(4.41)
N	19784	19782	19780	19794	19385

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for source of insurance.

Table 3.12: Correlates of Use of Medical Services: Insurance Dummies

	(1) Hospital Stay?	(2) # Stays	(3) # Hospital Nights	(4) Doctor Visit?	(5) # Doctor Visits
Medicaid	0.09***	0.17***	0.63	0.09***	5.65**
	(4.44)	(3.48)	(1.61)	(4.94)	(2.66)
Medicare	0.04***	0.08**	0.38**	0.03***	0.52
	(3.56)	(2.95)	(2.84)	(3.87)	(0.95)
Own Employer	0.02***	0.03*	0.16*	0.10***	2.04***
	(3.37)	(1.98)	(2.21)	(15.07)	(8.30)
Spouse's Employer	0.01	0.02	0.03	0.09***	2.04***
	(1.67)	(1.12)	(0.49)	(11.53)	(4.28)
Union	0.01	0.01	-0.01	0.01	1.09**
	(0.70)	(0.83)	(-0.14)	(1.41)	(2.90)
VA/CHAMPUS	0.05**	0.08**	0.11	0.08***	2.50***
	(2.85)	(2.63)	(0.88)	(8.29)	(4.58)
N	19784	19782	19780	19794	19385

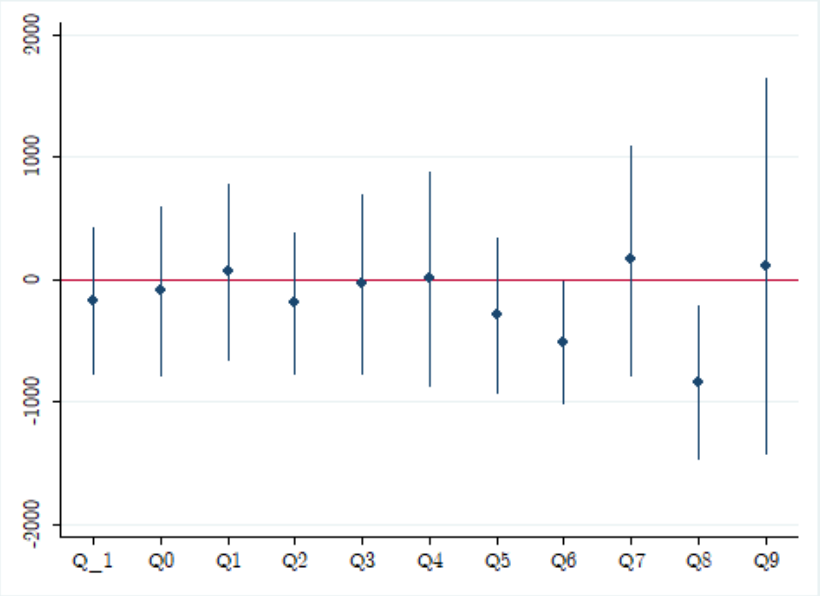
t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for sex, race, education (via indicator variables for whether an individual has graduated high school (incl. GED holders) or graduated from a four-year college), a quadratic in age and the natural log of household income.

Appendix B : Effect of Anticipated and Unanticipated Changes on 2-Year Out-of-Pocket Medical Expenditures

Figure 3.5: Changes to Realised Medical Expenditures over Preceding Two Years at Different Lags and Leads from Qualifying for Medicare Coverage



Notes: The dependent variable is the previous two years' out-of-pocket medical expenditures in U.S. dollars ($). Each plotted coefficient represents the coefficient from a dummy variable indicating how many quarters from Medicare eligibility a respondent is, pooling together data from the 2010-2014 waves of the Health and Retirement Study (HRS). Q_0 corresponds to the first quarter of Medicare eligibility, i.e. the individual is in the month of qualifying at the time of the survey or one of the two months following. Error bars correspond to the 95% confidence intervals for each coefficient. All specifications include a set of dummy variables for demographic, socioeconomic and insurance status (excluding Medicare eligibility since it is multicollinear with the coefficients of interest), as well as survey year dummies to absorb period effects common to all respondents. Standard errors are clustered at the individual level.

Table 3.13: Changes in the Distribution of Realised Out-of-Pocket Medical Expenditures Following Diagnosis of Cancer or Diabetes, or Marital Dissolution: Extensive Margin, Average Partial Effect & 95^{th} Percentile

	(1) OOP	(2) $(OOP > 0)$	(3) 95^{th} Percentile
Cancer in $(t-1, t)$	1942.46***	-0.02	6992.28***
	(3.92)	(-0.79)	(4.32)
Diabetes in $(t-1, t)$	-75.12	0.05*	-570.18
	(-0.38)	(2.54)	(-0.43)
Marital Dissolution in $(t-1, t)$	265.40	-0.01	1147.08
	(0.43)	(-0.32)	(0.84)
Marital Dissolution $\times$ Spousal in $(t-1, t)$	-848.33	0.03	-3214.91
	(-1.00)	(0.44)	(-1.70)
Cancer in $(t-2, t-1)$	1077.91*	-0.01	2845.01
	(2.32)	(-0.29)	(1.72)
Diabetes in $(t-2, t-1)$	123.44	0.02	727.18
	(0.52)	(1.18)	(0.63)
Marital Dissolution in $(t-2, t-1)$	-532.95**	-0.07	-1515.60
	(-2.88)	(-1.90)	(-1.02)
Marital Dissolution $\times$ Spousal in $(t-2, t-1)$	1091.72	-0.03	4224.25*
	(0.91)	(-0.18)	(1.97)
N	8585	8585	8585

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Standard errors are clustered at the individual level. All specifications control for period effects via wave indicator variables and include controls for sex, race, education (via indicator variables for whether an individual has graduated high school (incl. GED holders) or graduated from a four-year college), a quadratic in age and the natural log of household income.