The Response of Drug Expenditure to Non-Linear Contract Design: Evidence from Medicare Part D*

Liran Einav, Amy Finkelstein, and Paul Schrimpf†

August 2013

Abstract. We study the demand response to non-linear price schedules using data on insurance contracts and prescription drug purchases in Medicare Part D. Consistent with a static response of drug use to price, we document bunching of annual drug spending as individuals enter the famous "donut hole," where insurance becomes discontinuously much less generous on the margin. Consistent with a dynamic response to price, we document a response of drug use to the future out-of-pocket price by using variation in beneficiary birth month which generates variation in contract duration during the first year of eligibility. Motivated by these two facts, we develop and estimate a dynamic model of drug use during the coverage year that allows us to quantify and explore the effects of alternative contract designs on drug expenditures. For example, our estimates suggest that "filling" the donut hole, as required under the Affordable Care Act, will increase annual drug spending by $180 per beneficiary, or about 10%. Moreover, almost half of this increase is "anticipatory," coming from beneficiaries whose spending prior to the policy change would leave them short of reaching the donut hole. We also describe the nature of the utilization response and its heterogeneity across individuals and types of drugs.

JEL classification numbers: D12, G22.
Keywords: Medicare, Moral hazard, Contract Design, Health insurance, Health care.

*We thank Jason Abaluck, Raj Chetty, John Friedman, Nathan Hendren, Maria Polyakova, and many seminar participants for helpful discussions. Ray Kluender provided extraordinary research assistance. We gratefully acknowledge support from the NIA (R01 AG032449).

†Einav: Department of Economics, Stanford University, and NBER, leinav@stanford.edu; Finkelstein: Department of Economics, MIT, and NBER, afink@mit.edu; Schrimpf: Department of Economics, University of British Columbia, schrimpf@mail.ubc.ca.

1 Introduction

A classic empirical exercise is to study how demand responds to price. Many settings, from cell phones to electricity to health insurance, give rise to non-linear pricing schedules. These offer both challenges and opportunities for empirical estimation, while at the same time raising interesting conceptual questions regarding the nature of the demand response. In this paper, we study the demand response to non-linear contracts, and its implications for the impact of counterfactual contract design, in a particular context: the Medicare Part D prescription drug benefit.

The 2006 introduction of Medicare Part D was by far the most important benefit expansion in Medicare's nearly half-century of existence. As of November 2012, 32 million people (about 60% of Medicare beneficiaries) were enrolled in Part D, with expenditures projected to be $60 billion in 2013, or about 11% of total Medicare spending (Kaiser Family Foundation 2012a, 2012b). As rising drug spending places growing pressure on the federal budget, a natural question concerns the expenditure implications of alternative contract designs for Part D plans. For example, under the 2010 Affordable Care Act, drug benefits are slated to be further expanded by the requirement that the standard (i.e. minimal) benefit plan provide coverage in the famed "donut hole" by 2020 (Kaiser Family Foundation 2010). The expenditure implications of this change, and of alternative potential contract designs, are therefore, not surprisingly, the subject of considerable interest and attention.

We analyze the response of drug expenditures to insurance contract design. We use detailed micro data on insurance contracts and prescription drug purchases from a 20% random sample of Medicare Part D beneficiaries from 2007 to 2009. Our approach is motivated by the highly non-linear nature of the Part D contracts.

The nature of the contracts is illustrated by the government-defined standard benefit design, shown in Figure 1 for 2008. In this contract, the individual initially pays for all expenses out of pocket, until she has spent $275, at which point she pays only 25% of subsequent drug expenditures until her total drug spending reaches $2,510. At this point the individual enters the famed "donut hole," or the "gap," within which she must once again pay for all expenses out of pocket, until total drug expenditures reach $5,726, the amount at which catastrophic coverage sets in and the marginal out-of-pocket price of additional spending drops substantially, to about 7%. Individuals may buy plans that are actuarially equivalent to, or have more coverage than, the standard plan, so that the exact contract design varies across individuals. Nonetheless, a common feature of these plans is the existence of substantial non-linearities that are similar to the standard coverage we have just described. For example, in our baseline sample, a beneficiary faces an average price increase of almost 60 cents for every dollar of total spending as she enters the coverage gap.

Motivated by these contract features, we begin in Section 2 by presenting a simple, dynamic model of an optimizing agent's prescription drug utilization decisions given a specific, non-linear contract design. The model illustrates the key economic objects that determine the expenditure response to the contract. The first is the distribution of health-related events, which determine the set of potential prescription drug expenditures. The second is the "primitive" price elasticity that

captures the individual's willingness to trade off health and income. The third object is the extent to which individuals respond to the dynamic incentives associated with the non-linear contract.

After describing the institutional setting and the data in more detail in Section 3, we continue in Section 4 by presenting two pieces of descriptive evidence, which point to the presence and qualitative importance of the last two objects of the model. First, we document significant "excess mass," or "bunching" of annual spending levels around the kink in the budget set, at which beneficiaries enter the gap and the price increases substantially. This basic behavioral response is visually apparent in even the basic distribution of annual drug spending in any given year, as shown in Figure 2 for 2008. The exact amount at which the kink occurs changes from year to year as the government adjusts the parameters associated with the standard Part D benefit design. We show that the location of the bunching moves in lock steps with these changes. This illustrates the presence of a non-zero price response, or willingness to trade off health and income.

Second, we show that individuals respond to the dynamic incentives provided by the non-linear Part D contract, which is the other key economic element of the model. To do so, we take advantage of the fact that individuals newly eligible for Medicare can enroll in a Part D plan beginning the month that they turn 65. Since their initial coverage period will be the remaining length of the calendar year, the initial contract length among 65 year old beneficiaries will vary depending on their birth month. These institutional rules thus provide a setting in which individuals who enroll in the same plan but in different months face the same initial price for drugs but different future prices for a reason that is plausibly unrelated to prescription drug use. Using this design, we find that current prescription drug use responds to the future price arising from the non-linear contract.

In Section 5 we parameterize the model and estimate it using method of moments. The foregoing descriptive patterns are used, along with more standard moments of the spending distribution, in estimation. The estimated model fits the data quite well.

Section 6 presents the main results. We focus primarily on a variety of counterfactual policy simulations that examine the effect of alternative contract designs on spending. For example, we consider the requirement in the 2010 Affordable Care Act (ACA) that, effective in 2020, the standard (i.e. minimal) benefit plan eliminates the donut hole, providing the same 25% consumer cost-sharing from the deductible to the catastrophic limit (compared to the 100% consumer cost sharing in the gap in the original design). We estimate that this ACA policy of "filling the gap" will increase total drug spending by $180 per beneficiary (or about 10%), and will increase Medicare drug spending by substantially more (by $275 per beneficiary, or about 30%). By comparison, holding behavior constant, we estimate the "mechanical" consequence of filling the gap would be to increase average Medicare drug spending by only about $150, or just over half of our estimated effect.

Our results illustrate some of the subtle effects that non-linear contracts can produce, including changes in spending behavior for individuals that are far away from the gap, and how filling the gap can, somewhat counter-intuitively, provide less coverage on the margin to some individuals, causing them to *decrease* their spending. We illustrate some of these effects by exploring how counterfactual contract changes affect spending of individuals with different expected prescription drug spending.

For example, we estimate that almost half of our estimated $180 per-beneficiary increase in annual total drug spending from filling the gap comes from "anticipatory" responses by individuals whose annual spending prior to the policy change would have been below the gap.

In the last section of the paper, we explore in some more detail the source and nature of the spending response to the insurance contract. We do so by returning to the descriptive analysis of "bunching" in response to the kink. We show that at least some of this bunching is associated with a slowdown in the propensity to purchase drugs toward the end of the coverage year by individuals near the kink. This slowdown occurs across a variety of classes of drugs; it is somewhat more pronounced for chronic than acute drugs and substantially more pronounced for branded than generic drugs. We also show that some, but not all of this decline in purchasing at the end of the year likely reflects inter-temporal substitution whereby individuals around the gap defer filling some of their prescription drugs until January of the subsequent year, when coverage resets and the out-of-pocket price declines. Finally, we explore heterogeneity in the response to the kink across individuals, finding, for example, that healthier individuals are associated with greater bunching at the kink, and thus with presumably greater sensitivity to price.

Our paper is related to several distinct literatures. First, not surprisingly, there is a growing literature on the new Part D program. Much of this literature focuses on consumer's choices of plans (Heiss et al. 2010, 2012; Abaluck and Gruber 2011; Kling et al. 2012; Ketchem et al. 2012) although there are papers exploring other topics, such as the impact of the introduction of Part D on drug use (Yin et al. 2008; Duggan and Scott Morton 2010), and the impact of public subsidies on firm pricing behavior (Decarolis 2012).

Second, outside of the Part D context, there is an empirical literature examining how prescription drug spending responds to cost-sharing features of drug insurance. Chandra, Gruber, and McKnight (2010) provide one recent estimate, as well as a review of a handful of prior papers. In general, this literature has tended to provide "reduced form" estimates of the drug spending impact of plausibly exogenous variation in insurance contracts. In addition, the contracts studied usually differ in both the prescription drug cost-sharing and medical cost-sharing, complicating the isolation of the own-price effect of prescription drugs.[1]

Third, there is, of course, a vast and venerable empirical literature on the "moral hazard" (that is, spending) effects of medical insurance contracts that cover outpatient and inpatient care. We make no attempt to summarize this large literature here. We note, however, that most of it has aimed to characterize the spending effect of a health insurance contract with respect to "the price" consumers face under the contract, despite the highly non-linear nature of many observed contracts, and hence the difficulty in defining a single price induced by the non-linear budget set (Aron-Dine et al. 2013). In this respect, our attention to non-linear contract design for prescription drug insurance mirrors the recent flurry of interest in how non-drug medical spending responds to

[1] Notable exceptions are Tur-Prats et al. (2012), who examine the drug expenditure response to a discrete change in prescription drug cost-sharing only when individuals retire in Spain, and recent papers examining the impact of the Medicare Part D contract on drug use (e.g. Abaluck, Gruber, and Swanson 2013; Joyce et al. 2013).

non-linear medical insurance contracts (Vera-Hernandez 2003; Bajari et al. 2011; Kowalski 2011; Marsh 2011; Aron-Dine et al. 2012).

Finally, our analysis of "bunching" or "excess mass" in response to the price increase at the kink is related to a recent set of studies analyzing bunching of annual earnings in response to the non-linear budget set created by progressive income taxation, such as the Earned Income Tax Credit (Saez 2010; Chetty et al. forthcoming) and the Danish income tax schedule (Chetty et al. 2011). This literature has emphasized that since the amount of excess mass at a kink depends not only on the underlying behavioral elasticity but also on frictions, excess mass estimates alone cannot directly translate into an underlying behavioral elasticity (Chetty et al. 2011; Chetty 2012; Kleven and Waseen 2013). The frictions that are often pointed to in the labor supply context – such as supply side constraints on the number of hours that can be chosen to work and limited awareness of the budget set – are likely to be substantially less important in our setting. Individuals make an essentially continuous choice about drug spending (up to the lumpiness induced by the cost of a prescription) and get "real time" feedback on the current price they face for a drug at the point of purchase.[2] On the other hand, the translation of our bunching estimate into an underlying behavioral elasticity is not as direct as in the static framework developed by Saez (2010), since we must account for the fact that decisions are made sequentially throughout the year and information is obtained gradually as health shocks arrive. In this regard, our dynamic model is similar in spirit to the approach taken by Manoli and Weber (2011) in analyzing the response of retirement behavior to kinks in employer pension benefits as a function of job tenure.

2 A model of prescription drug use

Figure 1 showed that Medicare Part D plans provide highly non-linear coverage, with the out-of-pocket price changing sharply, and non-monotonically, during the year as the individual's prescription drug use accrues. In order to analyze the impact of different contract designs on prescription drug spending, we model the prescription drug use decisions of an individual with a specific contract. The model, which is similar to the one we developed previously in Aron-Dine et al. (2012), is designed to illustrate the key economic objects that determine the expenditure response to a contract. Our subsequent descriptive analysis will provide evidence on the presence and qualitative importance of these key economic objects, using minimal modeling assumptions. We will then parameterize and estimate the full model.

We consider a risk-neutral, forward looking individual who faces stochastic health shocks within the coverage period.[3] These health shocks can be treated by filling a prescription. The individual

[2] This real-time price salience may contribute to the difference between our finding of bunching and the absence of evidence of bunching by consumers at the convex kinks in the residential electricity pricing schedule, despite the ability to make an essentially continuous choice in that context as well (Ito 2012).

[3] Risk neutrality simplifies the inutition and estimation of the model. In the robustness section we describe and estimate a specification that uses a recursive utility model and allows for risk aversion and find that this has little

is covered by a non-linear prescription drug insurance contract j over a coverage period of T weeks. In our setting, as in virtually all health insurance contracts, the coverage period is annual (so that, typically $T = 52$).[4] Contract j is given by a function $c_j(\theta, x)$, which specifies the out-of-pocket amount c the individual would be charged for a prescription drug that costs θ dollars, given total (insurer plus out-of-pocket) spending of x dollars up until that point in the coverage period.

The individual's utility is linear and additive in health and residual income. Health events are given by a pair (θ, ω), where $\theta > 0$ denotes the dollar cost of the prescription and $\omega > 0$ denotes the (monetized) health consequences of not filling the prescription. We assume that individuals make a binary choice whether to fill the prescription, and a prescription that is not filled has a cumulative, additively separable effect on health. Thus, conditional on a health event (θ, ω), the individual's flow utility is given by

$$u(\theta, \omega; x) = \begin{cases} -c_j(\theta, x) & if \text{ prescription filled} \\ -\omega & if \text{ prescription not filled} \end{cases} . \tag{1}$$

Health events arrive with a weekly probability λ, and when they arrive they are drawn independently from a distribution $G(\theta, \omega)$. It is also convenient to define $G(\theta, \omega) \equiv G_2(\omega|\theta)G_1(\theta)$.

Given this setting, the only choice individuals make is whether to fill each prescription or not. Optimal behavior can be characterized by a simple finite horizon dynamic problem. The two state variables are the number of weeks left until the end of the coverage period, which we denote by t, and the total amount spent so far, denoted by x. The value function $v(x, t)$ represents the present discounted value of expected utility along the optimal path. Specifically, the value function is given by the solution to the following Bellman equation:

$$v(x,t) = (1-\lambda)\delta v(x, t-1) + \lambda \int max \left\{ \begin{array}{l} -c_j(\theta, x) + \delta v(x+\theta, t-1), \\ -\omega + \delta v(x, t-1) \end{array} \right\} dG(\theta, \omega), \tag{2}$$

with terminal conditions of $v(x, 0) = 0$ for all x. If a prescription arrives, the individual fills it if the value from doing so, $-c_j(\theta, x) + \delta v(x + \theta, t-1)$, exceeds the value obtained from not filling the prescription, $-\omega + \delta v(x, t-1)$.

The model thus boils down to a statistical description of the individual's health shocks, and two key economic objects which will be the focus of the rest of the paper. The individual's health shocks are captured by the arrival rate of prescriptions λ, and their associated (marginal) distribution of cost $G_1(\theta)$.

The first key economic object, summarized by $G_2(\omega|\theta)$, can be thought of as the "primitive" price elasticity that captures substitution between health and income. Specifically, $G_2(\omega|\theta)$ represents the distribution of the (monetized) utility loss ω from not filling a prescription of total cost θ. Of interest is the distribution of ω relative to θ, or simply the distribution of the ratio ω/θ. As ω/θ

affect on our main counterfactual estimates.

[4] Aggregating to the weekly level reduces the computational cost of estimating the model. This seems to be a not unreasonable approximation given that many prescriptions may arrive as a "bundle" that needs to be consumed together.

is higher (lower), the utility loss of not filling a prescription is greater (smaller) relative to the cost of filling the prescription, so (conditional on the cost) the prescription is more (less) likely to be filled. In particular, when $\omega \geq \theta$, the individual will fill the prescription even if she has to pay the full cost out of pocket, so the contract features will not affect the utilization decision. However, once $\omega < \theta$ the individual will fill the prescription only if some portion of the cost is (effectively) paid by the insurance. Thus, $G_2(\omega|\theta)$ can be thought of as capturing the price elasticity that would completely determine behavior in a constant price (linear) contract. In Section 4 we will provide descriptive evidence of a utilization response to the large jump in price as individuals enter the gap; this illustrates the presence of a non-zero price elasticity of demand for drugs, rejecting a model in which ω is always at least as large as θ.

The second key economic object in the model, summarized by the parameter $\delta \in [0, 1]$, captures the extent to which individuals understand and respond to the dynamic incentives associated with the non-linear contract. At one extreme, a "fully myopic" individual ($\delta = 0$) will not fill a prescription of cost θ if the negative health consequence of not filling the prescription, ω, is less than the immediate out-of-pocket expenditure required to fill the prescription, $-c_j(\theta, x)$. However, individuals with $\delta > 0$ take into account the dynamic incentives and will therefore make their decision based not only on the immediate out-of-pocket cost of filling the prescription, but also on the expected arrival of future health shocks and the associated sequence of prices associated with the non-linear contract. A greater value of δ increases the importance for the current utilization decision of subsequent out-of-pocket prices relative to the immediate out-of-pocket price.[5] Since price is non-monotone in total spending (for example, rising at the kink and then falling again at the catastrophic limit, as seen in Figure 1), whether an individual with $\delta > 0$ is more or less likely to fill a current prescription, relative to an individual with $\delta = 0$, will depend on their spending to date and their expectation regarding future health shocks. In Section 4 we will present descriptive evidence of a utilization response to the future budget set, thereby rejecting $\delta = 0$.

3 Setting and Data

Medicare provides medical insurance to the elderly and disabled. Medicare Parts A and B provide in-patient hospital and physician coverage respectively; Part D, which was introduced in 2006, provides prescription drug coverage. Our data are comprised of a 20% random sample of all Medicare Part D beneficiaries in 2007 through 2009. We observe the cost-sharing characteristics of each beneficiaries' plan as well as detailed, claim-level information on any prescription drugs purchased. We also observe basic demographic information (including age, gender, eligibility for various programs tailored to low income individuals). In addition, we have information on each beneficiary's Part

[5] For convenience we often refer to $\delta = 0$ as "myopia" and $\delta > 0$ as "forward looking". In practice, δ is affected not only by the "pure" discount rate, but also by the extent to which individuals understand and are aware of the budget set created by the non-linear contract, and by liquidity constraints. We thus think of δ as a parameter specific to our context.

A and B claims, which we feed into CMS-provided software to construct a summary proxy of the individual's predicted annual drug spending, which we refer to as the individual's "risk score."[6]

Unlike Medicare Parts A and B, which provide a uniform public insurance package for all enrollees (except those who select into the managed care option Medicare Advantage), in Medicare Part D enrollees can choose among different prescription drug plans offered by private insurers. The different plans have different plan features and premiums. All plans provide annual coverage by the calendar year, re-setting in January of each year, so that the individual is back on the first cost-sharing arm on January 1, regardless of how much was spent in the prior year.[7] Individuals newly eligible for Medicare can enroll in a Part D plan with coverage beginning the month that they turn 65 (CMS, 2011). Since their initial coverage period will be the remaining length of the calendar year, this generates variation in initial contract length among 65 year old beneficiaries depending on their birth month; we will exploit this variation in some of the descriptive analyses below, as well as in our estimation of the model.

Sample definitions and characteristics We make a number of sample restrictions to our initial sample of approximately 16 million beneficiary-year observations. We first limit our sample to those 65 and older who originally qualify for Medicare through the Old Age and Survivor's Insurance. This brings our sample down to about 11.6 million beneficiary-year observations. We further eliminate individuals who are dually eligible for Medicaid or other low-income subsidies, or are in special plans such as State Pharmaceutical Assistance Programs; such individuals face a very different budget set with zero, or extremely low consumer cost-sharing, for whom the contract design features that are the focus of the paper are essentially irrelevant. This further reduces our sample to about 7.4 million. Finally, we limit our attention to individuals in stand-alone prescription drug plans (PDPs), thereby excluding individuals in Medicare Advantage or other managed care plans which bundle healthcare coverage with prescription drug coverage. This brings our sample down to about 4.4 million. Several other more minor restrictions result in a baseline sample of about 3.9 million beneficiary-years, comprised of about 1.7 million unique beneficiaries.[8]

[6] Specifically we use CMS' 2012 RxHCC risk adjustment model which is designed to predict a beneficiary's prescription drug spending in year t as a function of their inpatient and outpatient diagnoses from year $t-1$ and demographic information (including age and sex and original route of eligibility onto Medicare). The risk scores are designed (by CMS) to be normalized to the average Part D beneficiary drug spending. They are used to adjust Medicare's per-beneficiary payments to insurance companies. More information on the risk scores can be found here: http://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk_adjustment.html

[7] During the open enrollment period in November and December, individuals can change their plan for the following calendar year. Otherwise, unless a specific qualifying event occurs, individuals cannot switch plans during the year.

[8] We exclude people who have missing plan details in any month of the year in which they are enrolled or who switch plans during the year. This excludes, among others, about 4% of the sample who die during the year. We also eliminate the small fraction of people in plans where the kink begins at a non-standard level; we use some of these individuals with non-standard kink levels for additional analyses in Appendix A.

We use the baseline sample for the first set of descriptive analyses in Section 4 and for the subsequent model estimation. For the second set of descriptive analyses, we restrict attention to the small subset of the baseline sample who is 65 years old and satisfy certain other requirements which we discuss later. Panel A of Table 1 presents some basic demographic characteristics of the original full sample, our baseline sample, and our 65 year old sub-sample. Our baseline sample has an average age of 76. It is about two-thirds female. The average risk score in our baseline sample is 0.88, implying that our baseline sample has, on average, 12% lower expected spending than the full Part D population.[9]

Prescription drug spending We have detailed, claim-level information on prescription drug spending. We use this to construct information on annual spending, as well as on the timing of drug purchases during the year. In some of the later analyses we will also use the National Drug Code (NDCs) to construct measures of the types of drugs consumed, in part relying on classifications provided by First Databank, a drug classification company.

Panel B of Table 1 presents summary statistics for annual total prescription drug spending. In our baseline sample, average annual drug spending is about $1,900 per beneficiary. As is typical, spending is right skewed; median spending is less than three-quarters of the mean. About 5% of the observations have zero spending, while the 90th percentile of annual spending is almost $4,000. Panel C reports the distribution of annual out-of-pocket spending, which ranges from zero to several thousand dollars annually**.**

Insurance contracts Since our analysis focuses on the impact of contract design, it is important to describe the contract features in some detail. Although there is substantial heterogeneity in plan features and premiums, the basic dimensions of the contract are determined by the government-defined "standard benefit." Insurance companies are required to offer a basic plan, which is either the "standard benefit" or a plan with "actuarially equivalent" value, defined as the same average share of total spending covered by the plan. Insurance companies may also offer more comprehensive plans, referred to as "enhanced plans."

Figure 1 shows the main features of the standard benefit plan in 2008. The total dollar amount of annual drug expenditures is summarized on the horizontal axis: this is the sum of both insurer payments and out-of-pocket payments by the beneficiary. The vertical axis indicates how this particular insurance contract translates total spending into out-of-pocket spending.

The figure illustrates the existence of several cost-sharing "arms" with different out-of-pocket prices. There is a $275 deductible, within which individuals pay for all drug expenditures out of pocket. That is, the individual faces a price of 1: she pays a full dollar out of pocket for every dollar spent at the pharmacy. After the individual has reached the deductible amount, the price

[9] We set the average risk score to missing for 65 year olds since risk scores for new Medicare Part D enrollees are, by necessity, a function of only a few demographics (primarily gender), so not fully comparable to risk scores of continuing enrollees.

drops sharply to 0.25. That is, for every additional dollar spent at the pharmacy, the individuals pays 25 cents out of pocket and the insurance company pays the remaining 75 cents. This 25% co-insurance applies until the individual's total expenditures (within the coverage period) reach the "initial coverage limit" (ICL), which we refer to as the kink. The kink location was $2,510 in the 2008 standard benefit plan. Once the kink is reached, the individual enters the famed "donut hole," or "gap," in which she once again pays all her drug expenditures out of pocket (price of 1) until her out-of-pocket spending reaches the "catastrophic coverage limit" (CCL). This limit, which is defined in terms of out-of-pocket spending (in contrast to the kink amount, which is defined in terms of total spending), was $4,050 in the 2008 standard benefit plan; this is equivalent to about $5,700 in total expenditure (see Figure 1). Only a small fraction of the beneficiaries (about 3% in our baseline sample) reach the catastrophic limit in a given year. Those who do face the larger of a price of 0.05 (i.e. a 5% co-insurance), or co-pays of $2.25 for a generic or preferred drug and $5.35 for other drugs. Empirically we estimate that this translates into a 7% co-insurance rate on average (in our baseline sample), which is the rate used in Figure 1.[10]

In analyzing the main cost-sharing features of the plans in our sample, we make two simplifying abstractions. First, we summarize cost-sharing in each plan-arm in terms of the percent of total claims that must be paid out of pocket by the beneficiary (co-insurance). Although this is how cost-sharing is defined in the standard benefit design, in practice, more than three-quarters of enrollees are in plans that specify a fixed dollar amount that must be paid by the beneficiary per claim (co-pays). To analyze the data in a single framework, we convert these co-pays to co-insurance rates for each plan-arm in the data by calculating the average ratio of out-of-pocket spending to total spending across all beneficiaries from our baseline sample in that plan-arm.[11] Second, we assume cost-sharing is uniform within a plan-arm, but actual plans often set cost-sharing within an arm differently by (up to six) drug "tiers"; drug tiers are defined by each plan's formulary and drugs are assigned to tiers based on whether the drug is branded or generic, among other factors.

Table 2, which summarizes our calculations of some of the main cost-sharing features of the plans in our sample, shows that our assumptions drive a (small) wedge between the stylized description of the plans and our empirical cost sharing calculations. For example, in our baseline sample we estimate average cost sharing in the gap for plans with "no gap coverage" of 0.98, and cost-sharing in the deductible for plans with a deductible of 0.88. In principle, both of these numbers "should be" 1, but in practice they are slightly less, reflecting some drug-specific exceptions.

There are several thousand different plans in our sample, although the differences among them

[10] The standard benefit has the same basic structure in all years, altough the level of the deductible, the kink, the catastrophic limit, and co-pays above it move around somewhat from year to year (see http://www.q1medicare.com/PartD-The-2009-Medicare-Part-D-Outlook.php).

[11] Since very few individuals reach the catastrophic limit, computing plan-specific cost sharing above this limit is difficult. We therefore caculate the average cost-sharing for all beneficiaries in our baseline sample in this arm across all plans. We note that almost all spending above the catastrophic limit is covered by the government directly, and therefore cost-sharing should be relatively uniform across plans.

are sometimes minimal. Table 2 attempts to summarize some of the main distinguishing features of the plans our sample is enrolled in. Many individuals in our baseline sample enroll in plans that offer different coverage than the standard plan. In particular, about three-quarters choose a plan with no deductible and almost one-fifth of those choose a plan with some gap coverage. Average cost-sharing below the kink is 0.34, often reflecting the fact that insurance companies find it attractive to offer an "actuarially equivalent" plan that, relative to the standard benefit design, has no deductible but charges higher co-insurance rate prior to hitting the kink Above the kink, the average cost sharing in the gap is close to 1, about 0.93. However, it varies substantially based on whether Medicare classifies the plan as one with no or "some" gap coverage.

4 Descriptive Analysis

4.1 Static price response: bunching at the kink

We focus first on behavior around the sharp price increase when individuals reach the kink. About 25% of beneficiaries in our baseline sample have spending at the kink or higher in a given year. Table 2 indicates that, at the kink, the price the individual faces increases on average by almost 60 cents for every dollar spent in the pharmacy. Standard economic theory suggests that, as long as preferences are convex and smoothly distributed in the population, we should observe individuals bunching at this convex kink point of their budget set. Saez (2010) provides a recent, formal discussion of this.

To see the intuition, consider a counterfactual linear budget set, i.e. the continuation of the co-insurance arm's cost sharing into the gap. In this case, individual spending would be distributed smoothly through the kink. For example, as illustrated in Appendix Figure A1, the solid and dashed indifference curves represent two individuals with different healthcare needs who would have different total drug spending under this linear contract. With the introduction of the kink, however, the spending of the sicker (dashed) individual will decrease and locate at the kink, as would all individuals whose spending under the linear contract was in between the solid and dashed individuals, thus generating "bunching." In a frictionless world, these individuals would pile up exactly at the kink. In practice, with real-world frictions such as the lumpiness of drug purchases and some uncertainty about future health shocks, individuals are instead expected to cluster in a narrow area around the kink.

An empirical illustration of this theoretical response to a non-linear budget set is evident in Figure 2, which reports a histogram of total annual prescription drug spending in 2008. The response to the kink is apparent: there appears to be a noticeable spike in the distribution of annual spending around the kink location.

The government changes the kink location each year. Figure 3 shows how the location of the bunching moves in virtual lock step as the location of the kink moves from $2,400 in 2007 to $2,510 in 2008, and to $2,700 in 2009. The fact that the location of the bunching moves with the location of the kink constitutes strong evidence that the bunching represents a behavioral response to the

sharp increase in out-of-pocket price as individuals enter the gap.[12]

Figure 4 pools the analyses across the three years and reports the frequency of spending relative to the (year-specific) kink location, which we normalize to zero. Focusing on the distribution of spending within $2,000 of the kink, Figure 4 presents our core, summary evidence of a behavioral response to the out-of-pocket price. It shows substantial "excess mass" of individuals around the convex kink in the budget set.

To quantify the amount of this excess mass, we follow the approach taken by Chetty et al. (2011) and approximate the counterfactual distribution of spending that would exist near the kink if there was no kink. Specifically, we fit a cubic approximation to the CDF, using only individuals whose spending is below the kink (between $2,000 and $200 from the kink), and subject to an integration constraint. The dashed line of Figure 4 presents this counterfactual distribution of spending. We estimate an excess mass of 29.1% (standard error = 0.003) in the -$200 to +$200 range of the kink, relative to the area under the counterfactual density in that range. That is, we estimate that the increase in price at the kink increases the number of individuals whose annual spending is within $200 of the kink by 29.1%; this increase is statistically significant. The presence of statistically significant excess mass around the kink indicates a behavioral response to price; with no behavioral response, there should have been no excess mass, a null that we reject.

One would expect that the magnitude of this excess mass would increase in how sharp the kink is. While, on average, beneficiaries experience about a 60 cents (per dollar spent) increase in price, this average masks considerable heterogeneity across plans, reflecting differences in cost-sharing both before and in the gap. Figure 5 therefore plots a plan-specific estimate of the plan-specific excess mass (constructed in the same way as before) against the size of the price change at the kink that is associated with each plan. We then fit a (weighted) regression line. It is reassuring to observe that the excess mass is increasing in the size of the price change, as economic theory would predict.

4.2 Dynamic price response: initial drug use by new enrollees

We now focus on the second object of the model of Section 2, by examining whether in fact individuals respond to dynamic incentives that arise from the non-linear pricing in their Part D contract (i.e. whether $\delta > 0$). Specifically, we consider whether, at the start of the contract period, individuals take into account the "future price" of drugs, or base their purchase decisions solely on the current "spot" price. The standard benefit design presented in Figure 1 shows how these

[12] In Appendix A we present additional corroborating evidence that the bunching at the kink represents a behavioral response to the gap. Specifically, Appendix Figure A2 shows that for the small subsample of individuals outside our baseline sample who are in contracts where the kink begins at a non-standard level of spending, there is no excess mass around the standard kink, but there is evidence of excess mass around the (non-standard) kink level. Interestingly, we also show in Appendix A that there is no evidence of missing mass at the concave kink created by the price decrease when individuals hit the deductible (see Appendix Figure A3); in the Appendix we speculate about a potential explanation.

spot and future prices may differ markedly. For example, an individual in a plan with the standard benefit design who has predictably high annual drug spending (for example, due to a chronic condition) faces an initial spot price of one (since he is in the deductible phase), but dynamic incentives could make her face an expected end-of-year price of 0.07 if she anticipates that, say, chronic medications will make her reach the catastrophic coverage threshold amount. Economic theory would predict that if the individual understands her dynamic incentives, this latter price is the more relevant price, which should enter her utilization decision.

The key empirical difficulty arises because, as Figure 1 also makes clear, the future price is a function of expected health spending, creating a mechanical endogeneity problem in an analysis of how the future price affects health spending. For individuals with the same expected spending, differences in their insurance contracts can create differences in future prices. However, different insurance contracts also create differences in spot prices. Testing for dynamic price response requires a setting in which individuals *with the same spot price* face different future prices for reasons unrelated to their health.

To address this challenge, we identify a setting in which we can compare initial drug use for individuals who face the same spot price but different future prices for a reason that is plausibly unrelated to prescription drug use. Specifically, individuals who newly enroll in a given Part D plan when they turn 65 face the *same initial spot price* for drugs. However, because the insurance contract resets at the end of each calendar year, different individuals in the same Part D plan face *different future prices* depending on which month of the year they turn 65 and enrolled in Part D. Furthermore, the sign and magnitude of the relationship between the month in which the individual joins the plan and the future price will vary depending on the type of plan.

This setting thus provides a natural set of contrasts. We examine how initial drug use varies across individuals within a plan by the individual's enrollment month, and how this within-plan pattern of initial-utilization-by-enrollment-month varies across plans with different relationships between enrollment month and future price. This empirical strategy is quite close in spirit to the approach we used previously to test for forward looking behavior in prime-age employees' medical utilization response to the non-linear pricing in their medical insurance (Aron-Dine et al. 2012). In that earlier work, variation in enrollment month came primarily from when within the year the employees joined the firm, while in the current setting it is primarily driven by the individual birth date.

Given the identification strategy, our analysis is limited to 65 year olds. With a few further restrictions for tractability, including limiting to individuals who enroll between February and October (this and other restrictions are described in Appendix B), we arrive at our "65 year old sub-sample" of about 137,000 beneficiary-years (see Table 1 and Table 2 for summary statistics).[13]

[13] Table 1 shows substantially lower annual spending for the 65 year old sub-sample than the baseline sample. This reflects two factors: first the 65 year olds are the youngest individuals in the baseline sample, therefore with the lowest expected spending. Second, we observe "annual spending" for a 65 year old for, on average, about 6 months, reflecting the fact that they do not enroll until their birth month.

Appendix B presents the analysis in detail. It includes formal regression analysis as well as a discussion of the key identifying assumption that any underlying seasonal patterns in initial drug use do not vary based on which plan the individual enrolled in except for dynamic incentives. For the sake of brevity, here we simply summarize the main empirical results graphically in Figure 6.

For this illustrative purpose, we show the pattern of expected end-of-year price and initial drug use by enrollment month separately for beneficiaries in two groups of plans: deductible and no-deductible plans. The expected end of year price depends on the cost-sharing features of the beneficiary's plan, the number of months of the contract, and the individual's expected spending. We proxy for expected end-of-year price with a "simulated future price" measure, in the spirit of Currie and Gruber's (1996) "simulated eligibility" instrument. Specifically, we calculate the simulated future price separately for each beneficiary based on his plan and his birth month, using the specific plan's cost-sharing features, the number of months in the contract if the individual enrolled in his birth month, and a common (across plan and birth month) distribution of monthly drug spending. On average, the simulated future price (shown on the right vertical axis) is increasing for deductible plans with enrollment month (since there is less time to spend past the deductible and into a lower priced arm) and decreasing for no-deductible plans (since there is less time to accumulate enough spending to enter the gap). We measure initial drug use as the number of days until the first claim, censored at 92 days. A longer number of days to first claim indicates less initial use.

The patterns of initial use by enrollment month present evidence against the null of no response to the dynamic incentives. In deductible plans, where the simulated future price is increasing with enrollment month, initial utilization is decreasing with enrollment month (i.e. time to first claim is increasing with enrollment month). By contrast, in the no-deductible plan, where the simulated future price is decreasing with enrollment month, days to first claim do not appear to vary systematically with the enrollment month. A "difference-in-difference" comparison of the pattern of initial drug use by enrollment month for people in plans in which the simulated future price increases with enrollment month *relative to* people in plans in which the simulated future price decreases with enrollment month thus suggests that initial use is decreasing in the expected end of year price.

5 Econometric model

Motivated by the foregoing evidence, we now turn to specify a more complete econometric model, based on the economic model of an individual's prescription drug use described in Section 2. As mentioned there, the model has three key primitives. The first is associated with how sick the individual is, which is summarized by the weekly arrival rate of prescriptions λ, and their associated (marginal) distribution of cost $G_1(\theta)$. The second is the "primitive" substitution between health and income, which can be summarized by $G_2(\omega|\theta)$. When $\omega \geq \theta$ the individual will consume the prescription even if she has to pay the full cost out of pocket. However, once $\omega < \theta$ the individual

will consume the prescription only if some portion of the cost is (effectively) paid by the insurance. Finally, the third object in the model is the extent to which individuals respond to the dynamic incentives associated with the non-linear contract, as summarized by the parameter δ.

Parameterization To estimate the model, we need to make two types of assumptions. One is about the parametric nature of the distributions that enter the individual's decisions, $G_1(\theta)$ and $G_2(\omega|\theta)$. We assume that $G_1(\theta)$ is a lognormal distribution with parameters μ and σ^2; that is,

$$\log\theta \sim N(\mu, \sigma^2). \tag{3}$$

We also assume that ω is equal to θ with probability $1-p$, and is drawn from a uniform distribution over $[0, \theta]$ with probability p. That is, $G_2(\omega|\theta)$ is given by

$$\omega|\theta \sim \begin{cases} U[0,\theta] & \text{with probability } p \\ \theta & \text{with probability } 1-p \end{cases}. \tag{4}$$

Recall that if $\omega \geq \theta$ it is always optimal for the individual to consume the prescription, so the assumption of a mass point at θ (rather than a smooth distribution with support over values greater than θ) is inconsequential. With probability $1-p$ the individual will consume the prescription regardless of the cost sharing features of the contract. A larger value of p implies that a larger fraction of shocks have $\omega < \theta$ and are therefore ones where drug purchasing may be responsive to the cost-sharing features of the contract.

With this parameterization, the extent of substitution between health and income is increasing in p, the probability that ω is lower than θ. To give a concrete interpretation, consider a person who faces a constant coinsurance rate of $c \in [0,1]$. That person will fill prescriptions whenever $\omega \geq c\theta$. This occurs with probability $1-pc$. A low value of p means that the person will fill most prescriptions regardless of the coinsurance rate. A high value of p indicates that the probability of filling prescriptions is more responsive to the coinsurance rate c.

The second type of assumption is about parameterization of heterogeneity across individuals in a given year. Since individuals' health is likely serially correlated, we introduce permanent unobserved heterogeneity in the form of discrete types, $m \in \{1, 2, ..., M\}$. An individual i is of type m with (logit) probability

$$\pi_m = \frac{\exp(z_i'\beta_m)}{\sum_{k=1}^{M}\exp(z_i'\beta_k)}, \tag{5}$$

where z_i is a vector of individual characteristics – our primary specification uses a constant, the risk score, and a 65 year-old indicator – and $\{\beta_m\}_{m=1}^{M}$ are type-specific vectors of coefficients (with one of the elements in each vector normalized to zero). All the parameters of the model – the weekly probability of a prescription λ_m, the parameters μ_m, σ_m, and p_m – are all allowed to vary across types, except the discount factor δ, which is more difficult to identify and is thus assumed to be the same for all types. As in Einav et al. (2013), our parameterization thus allows for heterogeneity in both individual health (λ, μ, σ), and in the responsiveness of individual spending to cost-sharing (p). Thus, the model has $4M$ parameters that define the M quadruplets $(\mu_m, \sigma_m, \lambda_m, p_m)$, the

single parameter δ, and $3(M-1)$ parameters that define the β_m's that shift the type probabilities. In our primary specification we use 5 types ($M = 5$), and thus have 33 parameters to estimate.

Our choice of parameterization imposes a number of limitations that deserve further discussion. First, the only source of correlation over time in health are the permanent unobserved types. This allows some people to be permanently sicker than others within a year, but does not allow people to become systematically healthier or sicker over time within the calendar year. Moreover, while we observe individuals for up to three years, we do not take explicit advantage of the panel structure of the data, and only use the persistence in risk scores from year to year to generate serial correlation in individuals' types over time.[14]

Second, our assumption that ω has a mass point at θ is completely innocuous. This is because our model implies that any individual (even one with no insurance) will fill every prescription when $\omega > \theta$. This means that we can never identify $G_2(\omega|\theta)$ above $\omega = \theta$, but also that the distribution of of ω above θ does not affect the model's predictions, and therefore has no effect on our counterfactual exercises.

Finally, our assumption that the ratio ω/θ is independent of θ implies that substitutability between health and income does not depend on the cost of a given prescription. Taken literally, this is not realistic: more expensive prescriptions are more likely associated with vital drugs with few close substitutes. We partially capture the idea that some drugs may be less substitutable by allowing both the distribution of θ and the distribution of ω/θ to depend on type. More generally, however, our goal is not to realistically model each individual's prescription choice but instead to capture the responsiveness of aggregate prescription spending to insurance coverage. We thus view our model as a reasonable approximation for this task.

Identification Loosely speaking, identification relies on three important features of our model and data. First, the non-linearity of Part D coverage generates variation in incentives that we use to recover the distribution of $\omega|\theta$, or the primitive substitution between health and income that would govern behavior in a linear contract. In particular, the bunching at the kink (shown in Section 4) allows us to identify the spending response to price where the spot and future price are the same (as in a linear contract). Second, the variation in initial spending by enrollment month (also shown in Section 4) helps in identifying δ, since variation in enrollment month generates variation in the future price among individuals who face the same spot price.[15] Finally, observing weekly claims made by the same individual over the entire year, along with our assumption that the (unobserved)

[14] Due to the dynamic nature of the model, adding unobserved persistence in individual types from year to year would require appropriate week-to-week type transitions. This would add at least one more state variable to the dynamic problem, and would make solving it and estimating the model considerably slower.

[15] To assess the importance of these moments for the actual identification, we follow the procedure recently proposed by Gentzkow and Shapiro (2013). We find that the estimation moments (described below) that are associated with the timing of Medicare enrollment account for approximately 40% of the contribution of all the moments to the estimation of δ. Analogously, we find that the estimation moments (again described below) that are associated with the bunching around the kink account for approximately 55% of the contribution of all the moments to the estimation

type is constant throughout the year, allows us to recover the distribution of health status for each type from the observed selected distribution of filled prescriptions.

More formally, we will consider identification conditional on plan characteristics and other covariates. To streamline the notation and discussion, we will leave the conditioning on covariates and plan characteristics implicit for the remainder of this section. We want to show that the observed distribution of prescription drug claims can uniquely identify the distribution of types, π_m, the distribution of health status given type, λ_m and $G_1(\theta|m)$, the substitutability between income and health, $G_2(\omega|\theta, m)$, and the parameter δ. The results of Kasahara and Shimotsu (2009) and Sasaki (2012) show the nonparametric identification of the distribution of types, π_m, the conditional (on type) distribution of θ, $G_1(\theta|m)$, and the conditional claim probabilities, $P(claim|m, \theta, x, t)$. Given the distribution of health status and conditional claim probabilities, the non-linearity of Medicare Part D coverage generates variation in incentives that trace out the distribution of ω.

To see this, note that an immediate consequence of equation (2) is that

$$\begin{aligned} P(claim|m,\theta,x,t) &= P(-c_j(\theta,x)+\delta v(x+\theta,t-1) \geq -\omega+\delta v(x,t-1)|m,\theta,x,t) = \\ &= P(\omega/\theta \geq \frac{1}{\theta}\left(c_j(\theta,x)+\delta v(x,t-1)-\delta v(x+\theta,t-1)\right)|m,\theta,x,t) = \\ &= 1-\overline{G}_2\left(\frac{1}{\theta}\left(c_j(\theta,x)+\delta v(x,t-1)-\delta v(x+\theta,t-1)\right)|m,\theta\right) \end{aligned} \tag{6}$$

where $\overline{G}_2(\cdot|m,\theta)$ is the conditional CDF of the ratio ω/θ. With linear insurance coverage, $c_j(\theta, x) = c\theta$, the value function does not depend on x, and equation (6) simplifies to

$$P(claim|m,\theta,x,t) = 1-\overline{G}_2(c|m,\theta). \tag{7}$$

In this case, without exogenous variation in insurance contracts, we would only be able to identify $\overline{G}_2(\cdot)$ at a single point. Fortunately, our data features nonlinear contracts, so we can identify $\overline{G}_2(\cdot|m,\theta)$ on a much larger range.

To eliminate the value function, consider the final week of the year. Then,

$$P(claim|m,\theta,x,1) = 1-\overline{G}_2\left(\frac{c_j(\theta,x)}{\theta}|m,\theta\right), \tag{8}$$

so we can identify $\overline{G}_2(\cdot|m,\theta)$ on the support of $c_j(\theta, x)/\theta$. The range of this support is an empirical question. Beyond the catastrophic limit, our contracts are linear with a coinsurance rate of around 7%. Below the deductible or in the coverage gap, the ratio $c_j(\theta, x)/\theta$ is as high as one. Thus, we can identify $\overline{G}_2(\cdot|m,\theta)$ on approximately $[0.07, 1]$. This is only approximate because there is variation in the coinsurance rates across plans, and we are showing identification conditional on plan.

Given $\overline{G}_2(\cdot|m,\theta)$, variation in claim probabilities with x and t allows us to identify δ from equation (6). If δ is near zero, then t will have little effect on the claim probabilities, given x. The larger is δ, the more important t will be. Although not strictly necessary, variation in enrollment month ensures that we observe a wide range of variation in t conditional on x from observing claim

of p, on average.

propensities just after enrollment for individuals that enrolled in different months, as well as wide range of variation in x conditional on t from observing claim propensities during the same month for individuals who have just enrolled and others who have been enrolled earlier and accumulated previous claims.

Estimation We estimate the model on our baseline sample using simulated minimum distance.[16] Let m_n denote a vector of sample statistics of the observed data. Let $m_s(\varphi)$ denote a vector of the same sample statistics of data simulated using our model with parameters φ. Our estimator is

$$\widehat{\varphi} \in \arg\min_{\varphi \in \Psi} (m_n - m_s(\varphi))' W_n (m_n - m_s(\varphi)). \tag{9}$$

The efficient choice of weighting matrix, W_n, is the inverse of the asymptotic variance of the sample statistics. We use this efficient weighting matrix, except that we overweight the sample statistics related to initial spending conditional on enrollment month. These statistics depend only on the portion of the sample that is 65 years old. As a result, their efficient weight is relatively low. However, these statistics are important for the identification of δ, so we increase their weight.[17] Appendix C describes in detail how we solve for the value function and simulate our model.

As moment conditions, we use the difference between observed and simulated moments that capture the key identifying variation in the data. In particular, we use moments that (i) summarize total "annual" spending; (ii) summarize the bunching around the kink; and (iii) summarize the variation in initial claims with enrollment month (for our 65 year old sub-sample). To summarize total annual spending, we use the probability of zero spending; the average of censored (at \$15,000) spending; the standard deviation of censored spending; the probability of annual spending being less than \$100, \$250, \$500, \$1000, \$1500, \$2000, \$3000, \$4000, and \$6000; and the covariance of annual spending with each of the covariates. To capture the persistence of individual spending over time, we use the covariance between spending in the first half and second half of the year. To capture the bunching around the kink, we use the histogram of total spending around the kink location, using twenty bins (each of width of \$50) within \$500 of the kink. That is, we divide the range of -\$500 to \$500 (relative to the kink location) into twenty equally sized bins and use frequency of each bin as a moment we try to match. Finally, to summarize the variation in initial claims with enrollment month, we use the average censored (at 12) weeks to first claim conditional

[16] In practice, to reduce computational cost, we make two inconsequential restrictions to our baseline sample when estimating the model. First, we limit the baseline sample to the 500 most common plans; this represents about 10 percent of plans but about 90% of beneficiary-years. Second, from this modified baseline sample we retain the entire 65 year-old sub-sample and a 10% random sample of older individuals. We weight the moments and the observations so that the results could be applied for the entire (modified) baseline sample.

[17] These moments are reweighted as though there are as many observations of the initial spending moments as the other moments. That is, if there are N total observations and n_j observations for the jth initial spending moments, then the (j,k) entry in the weight matrix is multiplied by $N/\sqrt{n_j n_k}$. These ratios range from 120-160 for the no-deductible plans and 700-750 for the deductible plans.

on the enrollment quarter and on having a deductible or not; this last set of moments is limited to the 65-year old sub-sample only and since we only include people who enrolled between February and October, we define the first quarter as February-April, the second as May-July, and the third as August-October.

It may be useful to highlight some computation challenges that we faced in our attempt to obtain estimates. Naive simulation of the model causes $m_s(\varphi)$ to be discontinuous due to the discrete claim decisions in our model. Due to the long sequence of discrete choices, conventional approaches for restoring continuity to $m_s(\varphi)$ fail. Each period an individual can fill a prescription or not, so there are 2^T possible sequences of claims. We cannot introduce logit errors to smooth over each period separately because the claims affect the state variable of total spending; calculating all 2^T possible sequences of claims and smoothing them is infeasible. While using importance sampling is possible in theory, in practice it is difficult to choose an initial sampling distribution that is close to the true distribution, resulting in inaccurate simulations. Instead, we use the naive simulation method to compute $m_s(\varphi)$ and utilize a minimization algorithm that is robust to discontinuity.

Specifically, we use the covariance matrix adaptation evolution strategy (CMA-ES) of Hansen and Kern (2004) and Hansen (2006). Like simulated annealing and various genetic algorithms, CMA-ES incorporates randomization, which makes it effective for global minimization. Like quasi-Newton methods, CMA-ES also builds a second order approximation to the objective function, which makes CMA-ES much more efficient than purely random or pattern based minimization algorithms. In comparisons of optimization algorithms, CMA-ES is among the most effective existing algorithms, especially for non-convex non-smooth objective functions (Hansen et al. 2010; Rios and Sahinidis 2012).

Our estimator has the typical asymptotic normal distribution for simulated GMM estimators. Although $m_s(\varphi)$ is not smooth for fixed n or number of simulations, it is smooth in the limit as $n \to \infty$ or $S \to \infty$. As a result,

$$\sqrt{n}\left(\hat{\varphi} - \varphi_0\right) \xrightarrow{d} N\left(0, (MWM)^{-1}MW(1+1/S)\Omega WM(MWM)^{-1}\right), \tag{10}$$

where $W = p\lim W_n$, $M = \nabla p\lim m_s(\varphi_0)$, Ω is the asymptotic variance of m_n, and S is the number of simulations per observation used to calculate m_s.

6 Results

6.1 Parameter estimates and model fit

Table 3 presents the parameter estimates. We find δ to be relatively close to one, at 0.93. Recall that our preferred interpretation of δ is not a (weekly) discount factor, which we would expect to be even closer to one, but simply a behavioral parameter that also reflects individuals' understanding of the insurance coverage contract, in particular the salience to them of the (future) non-linearities of the contract.

The rest of the parameters are allowed to vary by type, and our baseline specification allows for five discrete types. The types are ordered in terms of their expected annual spending (bottom rows of Table 3). The first, fourth, and fifth types are the most common and together account for about 85% of the individuals. As would be expected, increases in risk score are associated with increased probability of the highest spending type (type 5) and decreased probability of the lower spending types. Likewise, the 65 year olds are disproportionately lower spending types (i.e. type 3 relative to type 4).

A type's health is characterized by the rate of arrival of prescription drug events (λ) and the distribution of their size (θ). The first type is fairly healthy, with relatively low event probability (λ) and small claim amounts when there is a (potential) claim (i.e. low $E(\theta)$). The fourth type has a similarly low event probability as the first type, but more than double the (potential) claim amount ($E(\theta)$), while the fifth type's (potential) claim amounts are similar to the first type, but the fifth type experiences drug events on average every other week, almost six times more frequently than the first type.

Annual spending depends not only on health but also on the propensity to purchase (i.e. fill a prescription in response to a drug event), which depends on the parameter p. The parameter p likewise determines how responsive drug purchasing may potentially be to the cost-sharing features of the contract. The fourth and first types are relatively responsive in their drug purchase decisions to cost sharing features compared to the fifth type. For the fifth type, most of the prescription drug events will be purchased regardless of the insurance coverage (i.e. $p = 0.44$), whereas virtually all drug events of the first and fourth types will be sensitive to the cost sharing features of the insurance (p is very close to 1).

Overall, the model fits the data well. To assess the goodness of fit, we generated the model predictions by simulating optimal spending as a function of the estimated parameters and the observable characteristics for each beneficiary-year of our baseline sample. The top panel of Table 4 presents the observed and predicted summary statistics for the annual spending distribution; the model seems to fit well. The fit is also very good for two key patterns in the data that are important in identifying the model. Figure 7 presents the distribution of spending, for both the observed and the predicted data; we show both the fit of the overall spending distribution and the "zoomed in" fit without $1,000 of the kink. The two bottom panels of Table 4 present the observed and predicted number of weeks until the first claim, by deductible and no deductible plans, for individuals in the 65-year-old-subsample who enrolled in Part D at different times of the year. In both cases, the model tracks the observed patterns very well. Although it may not be surprising that we fit well these moments, given that these are some of the moments we try to match in estimation, it is still encouraging to observe that the model is flexible enough to be able to fit all of these rich patterns very well.

6.2 Spending response to counterfactual contract designs

Our primary objective of the paper is to explore how counterfactual contract designs affect prescription drug spending. We are interested in both mean spending effects (which are arguably the most policy-relevant) and also heterogeneity in the spending effects. In particular, we wish to examine how changes in non-linear contracts affect individuals at different points in the expected spending distribution.

The model and its estimated parameters allow us to accomplish precisely this. We will discuss some of the results in Tables 5 through 7 and in Figures 8 and 9. These are generated in the same way we assessed goodness of fit, except that we now simulate optimal spending under counterfactual (in addition to observed) contracts, again as a function of the estimated parameters and the observable variables in our sample. When we do this, we use the same set of simulation draws to generate individual-specific predictions, so simulation noise is essentially differenced out.

Initial illustration: filling the gap in the 2008 standard benefit design We begin by examining the spending implications of counterfactual changes to the 2008 standard contract shown in Figure 1. This focus on a single contract is useful for illustrating particular aspects of the spending response to alternative contract designs. In the next section we will consider counterfactuals involving a wider array of initial contract designs.

For illustrative purposes, we focus initially on perhaps our most policy-relevant exercise of "filling" the gap. As part of the Affordable Care Act (ACA), by 2020 the standard contract will no longer have a gap: the pre-gap coinsurance rate (of 25%) will instead be maintained from the deductible amount until catastrophic coverage (of about 7% coinsurance rate) kicks in at the current out-of-pocket catastrophic limit. We refer to this policy colloquially by the short-hand of "filling the gap."

Row 1 of Table 5 shows spending under the 2008 standard contract, and row 2 shows the results of filling the gap. On average, total spending increases by \$245, or about 14%, from \$1,710 to \$1,955. This increase in total spending reflects the combined effect of about \$140 decline in average out-of-pocket spending and about \$385 increase in insurer spending (right most columns). By way of comparison, we estimate that if utilization behavior were held constant, filling the gap would decrease out of pocket spending on average by about \$200 (and naturally increase average insurer spending by the same amount).

The spending effects of filling the gap are quite heterogeneous. For example, comparing rows 1 and 2 of Table 5, we see that the median increase in total spending is only about \$30, while the 90th percentile change is about \$1,000. Figure 8 provides a look at which individuals are affected by the change. The figure plots the distribution of the change in spending from filling the gap as a function of the individual's predicted spending under the 2008 standard contract. It shows, not surprisingly, that most of the change in spending from filling the gap is driven by changes in spending by individuals whose predicted spending under the standard contract would be in the gap.

However, the figure also highlights two somewhat subtle implications of non-linear contracts. First, recall that due to dynamic considerations, there is the possibility of an "anticipatory" positive spending effect from filling the gap for people who do not eventually hit the gap. Figure 8 shows that, indeed, there is an increase in spending from filling the gap for people whose predicted spending under the standard contract is quite far below the gap. This highlights the potential importance of considering the entire non-linear budget set in analyzing the response of health care use to health insurance contract. Quantitatively, we estimate that the increase in spending among people more than $200 below the kink location under the standard plan accounts for almost 20% of the average $245 per person increase in annual drug spending.

A second, somewhat counter-intuitive result is that the ACA policy of filling the gap causes some individuals to actually *decrease* their spending. Because the catastrophic limit is held constant with respect to out-of-pocket rather than total spending when the gap is "filled," it takes a greater amount of total spending to hit the catastrophic limit. Thus, holding behavior constant, some high-spending individuals who under the old standard contract had out-of-pocket spending that put them in the catastrophic coverage range where the marginal price is only 7 cents on the dollar would, under the "filled gap" contract, have out-of-pocket spending that leaves them still within the ("filled") gap, where the marginal price would be 25 cents on the dollar. We see this in Figure 8 where, among individuals whose total spending put them above the catastrophic limit under the 2008 standard contract, some reduce their spending in response to filling of the gap. This illustrates a more general point that, with non-linear contracts, a given change in contract design can provide more coverage (less cost sharing) on the margin to some individuals but less coverage to others.

To gain more insight into the response to "filling the gap," we provide a comparison analysis of the effects of "filling the deductible." We "fill" the $275 deductible from the 2008 standard benefit in an analogous manner to our filling of the gap. That is, we make the deductible zero, and individuals pay 25 cents of the dollar until they hit the kink. Holding behavior constant, this change is of roughly the same order of magnitude as filling the gap; the deductible is $275 and affects all individuals, while the gap covers a range of spending that is about ten times greater, but only about a quarter of the individuals reach the gap and only few reach its end (the catastrophic limit).

Our findings indicate that filling the deductible increases spending by less than filling the gap. Row 3 in Table 5 shows that filling the deductible raises total spending, on average, by about $170 (about 10%) compared to the baseline in row 1. As would be expected, the set of people affected by filling the deductible is very different than those affected by filling the gap. This can be seen in Figure 9 which, analogous to Figure 8, plots the distribution of the change in spending from filling the deductible as a function of predicted spending under the 2008 standard coverage contract. Here, we see that the increase in spending comes predominantly from lower spending individuals.

Spending effects of other counterfactual contracts Thus far we have considered the spending effect of changes to only the 2008 standard contract. However, in practice, as seen in Table 2, many people have coverage that exceeds the standard contract, including some gap coverage. To

more accurately forecast the expected spending effects from a given policy change, we therefore examine the implications of changes in contract design for the existing distribution of contracts, rather than only for the standard contract. In all these counterfactuals, we assume that firms do not respond by making other changes to contracts, and that the distribution of beneficiaries across contracts remains fixed; specifying and estimating the demand and supply of contracts is beyond the scope of the paper, although in the robustness analysis below we do explore sensitivity to one relatively crude way of accounting for beneficiary selection of contracts.

In rows 4-6 of Table 5 we examine the impact of filling the gap or filling the deductible on the observed distribution of contracts in our data. The existence of more comprehensive coverage than the standard plan is reflected in the higher mean baseline spending (compare row 4 to row 1), and the smaller spending increase from filling the gap (row 5) or filling the deductible (row 6). Given the observed distribution of plans in the data, we estimate that filling the gap will raise total annual drug spending by about $180 per beneficiary, or about 10%, from $1,744 to $1,925. Almost half of this increase in total average annual spending ($81 out of $181) comes from increases in spending by individuals who are predicted to spend more than $200 below the kink location under their original plan, suggesting a quantitatively important role for "anticipatory" behavior. The $180 average increase in total annual spending reflects the combined effect of about a $95 decline in average out of pocket spending and about a $275 increase in average Medicare spending (right most columns).[18] By contrast, ignoring the behavioral response to the contract, we calculate that "filling the gap" would increase average Medicare spending by only $150, just over half of our estimated increase in average Medicare spending.

Finally, the last three rows of Table 5 report the results from counterfactual contracts that are further out of sample, so should be interpreted with more caution. We provide them here as a way of quantifying the overall spending effect of insurance and the overall "money at stake." In rows 7 and 8 we analyze spending under the extremes of full prescription drug insurance (i.e. 0% consumer co-insurance everywhere) and no prescription drug insurance (i.e. 100% consumer co-insurance everywhere). These estimates imply that going from full to no insurance would decrease spending by about $2,000, or about 80%. A less extreme exercise is reported in row 9, where we eliminate insurance, but retain the catastrophic coverage. This exercise does not expose individuals to extreme risk, yet spending still declines, relative to full insurance, by 62%, an enormous effect.

Comparison to other estimates As one way to compare our results to prior estimates, we note that our estimates imply that relative to no insurance (Table 5, row 8), spending under the observed set of contracts (Table 5, row 4) increases spending on average by about $1,145, or about 300%. At a broad level, this exercise is similar to various "reduced form" estimates of the impact of the introduction of Part D on drug spending. Using various difference-in-differences empirical

[18] We estimate the increase in insurer spending and assume this higher spending is completely passed through to Medicare in the form of higher Medicare reimbursement of insurers. See Duggan et al. (2008) for more information on how Medicare reimburses insurers.

strategies and data sources, estimates of the impact of the introduction of Part D on increased drug use range from about 6% to about 70% (Duggan and Scott Morton, 2010, Yin et al. 2008, Ketcham and Simon, 2008). To compare these estimates to our estimate of the increase in spending due to typical Part D coverage, one must take into account that, on average, the introduction of Part D increased the fraction of the elderly with prescription drug coverage by only 10 percentage points (Englehardt and Gruber 2010); the implied increase in drug use coming from Part D *coverage* is therefore 10 times higher than these "reduced form" estimates of impact of the introduction of Part D, which makes our (highly out-of-sample) prediction fit within this (wide) range of the existing estimates.

To try to move beyond policy-specific counterfactuals to a more general economic object that applying to other budget sets or comparing to other estimates, we also calculate the implied elasticity of drug spending for a given percent reduction in cost-sharing on every arm of the 2008 standard benefit budget set.[19] Table 6 shows the results. Perhaps not surprisingly, the elasticity decreases (in absolute value) as the price change is greater; at some point the probability of claiming in response to a shock becomes sufficiently high that further price reductions have a smaller effect. The implied elasticity of drug spending with respect to the price ranges from about -0.75 (for a 1.5% reduction in cost-sharing throughout the budget set) to -0.5 (for a 75% reduction in cost-sharing throughout the budget set). Outside of the Part D context, "reduced form" estimates of a price elasticity of demand for drug spending range from around -0.1 to -0.4 (see Chandra, Gruber, and McKnight 2007 for a summary of the literature).

We can also compare the implied elasticity estimates from our dynamic model to what we would obtain by adapting the Saez (2010) approach of translating excess mass (or "bunching") estimates into an elasticity to our setting. In a static, frictionless environment, Saez (2010) presents a stylized model that allows conversion of the excess mass of income tax filers at convex kinks created by the progressive income tax schedule into a (local) estimate of the compensated elasticity of income with respect to the net of tax rate. In Appendix D we adapt Saez's approach to our setting, replacing his model of a constant elasticity of income with respect to the net of tax rate with a model of constant elasticity of medical spending with respect to a function of the co-insurance rate. We use this to translate our plan-specific excess mass estimate (see Figure 5) into (local) elasticities of drug spending with respect to the coinsurance rate, evaluated at each plan's pre-kink cost sharing rate. Averaging across plans (weighted by enrollment), we estimate an elasticity of -0.024, or about an order of magnitude lower than the implied elasticities from our dynamic model.

Despite the appeal of taking the static Saez (2010) framework "off the shelf" and translating it to our setting, this framework makes many assumptions that are poorly suited to our problem. Most importantly, annual spending in our setting is the result of individuals making many sequential prescription drug purchase decisions throughout the year as health shocks arrive (and information

[19] In Table 6 we report elasticity estimates by computing the ratio of the percent change in spending to the percent change in price. When price changes are large, this calculation is not the same, of course, as a "pure" elasticity which is defined locally (i.e. for a marginal change in price).

is revealed) and the price of treating each shock changes as individuals move along their non-linear budget set. This is in sharp contrast to the assumption of the static framework in which all the uncertainty is realized prior to a (single) annual spending decision. Relatedly, if (as our descriptive analysis and model estimates indicate) individuals respond to the dynamic incentives provided by the non-linear contract, then not only does information arrives gradually, but also early purchase decisions reflect individuals' expectations about future health shocks and their associated out-of-pocket price, adding yet another important dynamic effect. For example, the static analysis by construction limits the behavioral response to the kink to those near the kink. Yet the set of people "near" the kink and therefore "at risk" of bunching may in fact by endogenously affected by the presence of the kink if forward looking individuals, anticipating the increase in price if a series of negative health shocks puts their spending near the kink, make purchase decisions when they are far away from the kink that decrease their chance of ending up near it.[20] Indeed, our estimates from the dynamic model pointed to a non-trivial role for such "anticipatory" behavioral responses by people who expect to end up far below the kink, which would not be captured by the static model. This is one reason why the static model produces estimates that are so much lower than the dynamic one.[21]

6.3 Robustness

In our model and parameterization, we have made many assumptions. In this section we briefly assess the sensitivity of our main findings to some of these assumptions. Table 7 summarizes the results, by reporting, for each robustness check, the implied effect on "filling the gap" on total and insurer spending. Overall, the results appear quite stable across specifications. Specifically, across the specifications (discussed below), the estimated increase in total annual drug expenditures from filling the gap ranges from 7.5% to 12%, which is qualitatively similar to our baseline estimate (of 10.4%). Similarly, the estimated increase in insurer expenditure ranges between 25% and 30%, while our baseline estimate was 28.4%. Results from other counterfactual exercises discussed above also appear quite stable (not reported in the table).

The first row of Table 7 reports the estimates from the baseline specification. Rows 2 and 3 assess the sensitivity of the results to changing the number of discrete types. In our baseline specification we assumed, somewhat arbitrarily, that heterogeneity is captured by a mixture of five discrete types ($M = 5$). In row 2 we estimate the model using three types ($M = 3$), and in row 3 we use six types ($M = 6$). Since the share of one of the five types in our baseline specification was

[20] Manoli and Weber (2011) note a related set of dynamic forces at play when estimating the response of retirement behavior to kinks in employer pension benefits as a function of job tenure.

[21] Another potential contributor to the smaller static estimates may be that the "bunching" estimator estimates a behavioral response that is local to people around the kink. In practice, we estimate that these people are disproportionately of the type that exhibits a lower behavioral response to price (i..e, the fifth type in Table 3 is found disproportionately around the kink and has a relatively low sensitivity of drug purchasing to cost sharing features, as captured by the paramter p).

only about 2% (see Table 3), it may not be surprising that adding a sixth type does not affect the results much. Indeed, the six-type specification gives rise to results that are quite similar to the baseline ones, and the share of the sixth type is close to zero, suggesting that adding additional types (beyond six) – an exercise that we have not done and is computationally intensive due to the increase in the number of parameters – is also unlikely to affect much the results.

Row 4 and 5 of Table 7 assess the sensitivity of the results to the choice of covariates z_i. In our baseline specification we use the beneficiary's risk score and an indicator for whether he is 65 years old as the two covariates (see Table 3). In row 4 we use only a constant and no other covariates, while in row 5 we add an indicator that is equal to one if the beneficiary selected a plan that provides no gap coverage (in addition to the included covariates of risk score and a 65 year old indicator). The latter specification is a rough, "reduced form" attempt to capture potential plan selection on unobservables, for example, that healthier beneficiaries may be more likely to select plans with no gap coverage.

While modeling plan selection is outside the scope of our current exercise, one potential concern with using our baseline model to assess the counterfactual effects of changes in contract design is that it does not allow for any effect that such contract changes may have on inducing some beneficiaries to select different plans. A related concern is the possibility that the effect of a contract change like "filling the gap" is heterogeneous across individuals and that the selection of plans is correlated with this heterogeneity (i.e. "selection on moral hazard" as in Einav et al., 2013), so that the size of the treatment varies across individuals with different treatment effects (e.g. individuals with a larger "treatment effect" due to higher p select plans that offer gap coverage and therefore experience less of a "treatment" from the counterfactual of "filling the gap"). However, the fact that our estimates do not change much once we include a "no gap" indicator as a covariate suggest that plan selection is unlikely to have a first-order effect on our primary estimates of interest.

Finally, in rows 6 and 7 of Table 7 we examine the sensitivity of our results to our modeling assumption of individuals as risk neutral. While the assumption of risk neutrality appears odd in the context of insurance, risk neutrality may not be a bad approximation for a week-to-week decision making, even when the utility function over annual quantities (of income and/or health) is concave. To assess this conjecture, we extend the model of Section 2 and specify a utility model that allows for a concave utility function. Specifically, we introduce risk aversion while maintaining perfect intertemporal substitution by specifying recursive preferences as in Kreps and Porteus (1978) or Epstein and Zin (1989). As in our baseline model, an individual's flow utility is linear and additive in health and residual income. Since we do not observe residual income, we assume constant absolute risk aversion so that residual income does not affect claiming decisions. Thus, individual preferences over a stochastic sequence of flow utilities, $\{u_t\}$, are defined recursively as

$$V_t = u_t + \delta \left(\frac{-1}{\alpha} \right) \log E_t[e^{-\alpha V_{t+1}}] \tag{11}$$

where α is the coefficient of absolute risk aversion.[22] The limit, as α approaches zero, is equivalent

[22] These preferences are equivalent to $V_0 = E_0 \left[-e^{-\alpha \sum_{t=0}^{T} \delta^t u_t} \right]$.

to our baseline specification. For the results reported in Table 7 we set the values of α to span the range of (absolute) risk aversion estimates that are obtained in a similar health-related context by Handel (2013). The main results remain qualitatively similar.

7 Nature of spending response

Thus far we have analyzed the overall drug expenditure response to contract design. In this final section we try to shed some (qualitative) light on the source and nature of the response. We do so by returning to the graphical analysis of bunching around the kink as in Section 4.

7.1 Timing of purchases

The finding of bunching in response to the kink (e.g., Figure 4) presumably reflects individuals foregoing (or postponing, as we discuss next) prescriptions that they would otherwise have filled.[23] An attractive feature of our setting is that, unlike in the classic labor supply setting for studying bunching (Saez 2010), we observe within-year behavior rather than just annual behavior. We use the information on the date of purchase (i.e. date of claim) to explore the nature of the utilization response to price as it shows up in the end-of-year purchasing behavior.

Figure 10 shows the propensity to purchase at least one drug during the month of December as a function of the total annual spending.[24] As in the earlier graphical analysis, the horizontal axis reflects the annual total spending of each individual, normalized relative to the year-specific kink location. The vertical axis now presents, for each $20 bin of total spending, the share of individuals with at least one December prescription drug purchase. Absent a price response, it seems plausible to assume that the seasonal pattern of purchases would be similar across individuals with different levels of spending, and thus the share of individuals with a December purchase would monotonically increase with the level of spending (that is, with the overall frequency of purchases) and would approach one for sufficiently sick individuals who visit the pharmacy every month.

Indeed, this is the pattern that is shown in Figure 10 for individuals whose spending is far enough from the kink. Yet, the figure shows a sharp slowdown in the probability of end-of-year purchases as individuals get close to the kink. Once they enter the gap the pattern reverts to

[23] It is highly unlikely that the bunching pattern reflects purchases that are being made but simply not being claimed (due to the reduced incentives to claim in the gap). First, most prescription drug purchases are automatically registered with Medicare directly via the pharmacy (there is no need for the individual to separately file a claim). Moreover, given that most contracts have some gap coverage (see Table 2) and all have catastrophic coverage if individuals spend sufficiently far past the kink, individuals have an incentive to report ("claim") any drug spending in the gap.

[24] The choice of the last calendar month is somewhat arbitrary. Analyses of claiming behavior during shorter or longer end-of-year periods (e.g. last two weeks of December or November and December) yield similar qualitative patterns.

the original monotone pattern, albeit at a lower frequency of December purchases, presumably reflecting the higher cost-sharing in the gap. The fitted line in the graph illustrates the difference between actual purchase probabilities and what would be predicted in the absence of the kink.[25] Overall, the figure illustrates that an important way by which individuals respond to the kink is by purchasing less in the end of the year.

7.2 Heterogeneity in spending response

Another dimension along which one can better understand the overall drug expenditure response to cost sharing is to see how the response varies across different types of drugs or people (as in e.g. Goldman et al. 1994, and Chandra et al. 2010). Table 8 examines how the decline in end-of-year purchases around the kink varies across types of drugs. The first row reports our estimates for the entire sample, which mirrors the graphical analysis presented in Figure 10. On average, individuals who reach the gap appear to reduce their propensity to purchase in December by just over 8%. This reduction appears to be about two percentage point larger for chronic or "maintenance" drugs relative to acute and "non maintenance" drugs, respectively; this may reflect greater flexibility regarding the timing of drug purchases for chronic conditions.[26] The purchase of branded drugs slows down much more sharply around the kink than generic drugs – a 20% decline in the probability of purchasing a branded drug in December, compared to 8.5% decline for generics. This presumably occurs because branded drugs tend to be much more expensive (on average in our sample about $130 per drug compared to about $20 for generics), so the per-prescription (rather than per-dollar) price effect of entering the gap is significantly greater for branded drugs. [27] Our finding of a larger reduction in branded drug purchases than generics in response to the donut hole is similar to previous findings for diabetics that their use of branded drugs decreases much more than use of generics when they enter the donut hole (Joyce et al. 2013). Finally, the bottom row of Table 8 shows a greater (12%) decline in purchasing of "inappropriate" drugs compared to the average 8%

[25] To fit the line, we run a simple regression of the logarithm of the share of individuals with no December claim in each $20 spending bin on the mid-point of the spending amount of the bin, weighting each bin by the number of beneficiaries in that bin. We fit this regression using all bins between -$2,000 and -$500. This specification is designed to make the share of December claims (purchases) monotone in the spending bin and asymptote to one as the bin amount approaches infinity. As can be seen in Figure 10, the fit appears quite well (prior to getting close to the donut hole, where the price effects kick in).

[26] Following the spirit of Alpert (2012), we classify a drug as chronic if, empirically, conditional on consuming the drug, the median beneficiary consumes the drug more than two times within the year. We classify a drug as "maintenance" vs. "non maintenance" using the classification from First Databank, a drug classification company. This classification is roughly analagous to being a drug for a chronic condition or not.

[27] The size of the kink is roughly the same for branded and generic drugs. In the 2008 standard benefit plan, the price goes from 0.25 to 1 at the kink for both branded and generic drugs. Looking across the observed contracts in our baseline sample, on average the consumer price rises at the kink by about 60 cents for branded drugs and about 55 cents for generic drugs.

decline in drug purchasing, providing some evidence that higher prices may lead to a somewhat more careful selection of drugs.[28]

We also explored heterogeneity in the response to cost-sharing across different types of individuals by returning to the excess mass analysis (as shown in Figure 4) and computing it separately for different types of individuals. Table 9 shows the results. We find statistically significant excess mass in all sub-groups. The size of the excess mass increases with year of the Part D program, from 9% in the first year (2006) to 30% in the last year we observe in the data (2009). This may reflect a "learning" effect (by individuals or pharmacists) about the presence of the gap.[29] The behavioral response to the contract, as measured by the excess mass, is slightly higher for men than for women, and tends to be larger for healthier individuals, as measured by age, the number of hierarchical conditions the individual has, or by the individual's risk score.[30]

7.3 Inter-temporal substitution

The evidence in Figure 10 that individuals stop purchasing drugs late in the year once they are near the kink raises the question of whether they never purchase these prescriptions or simply shift the purchase to the beginning of the next calendar year, when the coverage schedule "resets" and the spot price and the expected end-of-year price are lower. To examine this, we explore whether there is a relationship between "excess" January spending in year $t+1$ and total annual expenditures relative to the kink in year t.

We define an individual's "excess" spending in January $t+1$ as the ratio of her January spending in year $t+1$ to her average monthly spending in February to December of $t+1$. The top panel of Figure 11 graphs this measure of "excess" January spending in year $t+1$ against total annual expenditures in year t. If the slowdown in purchasing propensity toward the end of the calendar year as individuals approach the gap that we saw in Figure 10 simply reflects a decline in drug purchases, there should be no systematic relationship between excess January spending in the subsequent year and prior year's spending. However, if some of the slowdown in purchasing reflects intertemporal substitution toward filling the same prescriptions in the subsequent year, we should expect to see excess January spending in year $t+1$ for individuals who approach (or enter) the gap in year t.

The results in the top panel of Figure 11 strongly suggest that such intertemporal substitution

[28] Following Zhang et al. (2010), we proxy for innappropriate drug use using an indicator from the Healthcare Effectiveness Data and Information Set (HEDIS) on whether the drug is considered high-risk for the elderly (HEDIS 2010).

[29] For the analysis by year we add in the 2006 data on the first year of the program, which we have otherwise excluded from the sample; we limit the "by year" analysis to the approximately two-fifths of individuals who joined in January of 2006 and who remained in the data through 2009.

[30] For the analysis by age we exclude 65 year olds since they join throughout the year and therefore the set of 65 year olds near the kink likely differ than at other ages. The Hierarchical Conditions are inputs into the CMS risk score; they are meant to capture conditions that are predictive of higher drug spending in the next year, such as diabetes and hypertension.

occurs. For individuals whose spending is far below the gap, spending in the subsequent January appears representative of any other month later that year. Yet, as individuals come close to the gap (or end up in it), their subsequent January spending jumps up to be 30-40% higher than a "regular" month, presumably due to accumulated prescription drugs whose purchase could be deferred from the previous year, when the out-of-pocket price was higher.

This raises the question of whether inter-temporal substitution can largely or entirely explain the annual spending reduction that we estimated. Empirical analyses of the spending effects of health insurance contracts have typically focused on annual spending effects, and our paper follows that tradition. However, if the annual spending response is largely or entirely undone over a longer period, one might want to assess the budgetary implications of cost-sharing in insurance contracts beyond a one-year horizon as well as consider the wisdom of designing incentives around annual contracts.

Assessing the quantitative magnitude of intertemporal substitution is not easy. It would require a complete, infinite-horizon model in which an individual decides about the timing of purchases as a function of her expectation about the expected price of next year's purchases, which in turn would have to take into account the price expectation in the subsequent year, and so on. (By contrast, our crude exercise shown in the top panel of Figure 11 above only looks at the effects of the kink for spending in the first month of the new year). This type of exercise is beyond the scope of the current paper.

However, the bottom panel of Figure 11 provides qualitative evidence suggesting that the annual spending response to the kink we documented in the paper is unlikely to be fully explained by intertemporal substitution to January of the subsequent year. We follow a strategy used by Chetty et al. (2011) and present a density of the sum of year t's spending and the average dollar difference, for each spending bin in year t, between the average January spending in year $t+1$ and the average monthly spending in February to December of year $t+1$. As the figure shows, the bunching around the kink (seen previously in Figure 4 when the spending density was plotted as a function of year t spending relative to the kink) remains when the density is plotted as a function of this "adjusted" year t spending: it would have been eliminated if the entire response would have been driven by shifting claims to January. We should note that these results get weaker as we shrink the size of the bin used in the adjustment. This is to be expected; in the extreme, when adjusting using spending at the individual-level (rather than at the bin-average), we see no evidence of bunching relative to "adjusted" annual spending, presumably reflecting the addition of a large amount of individual-specific realization noise.[31] Yet, under the null that the entire response is driven by intertemporal substitution, the (adjusted) bunching would be eliminated for any bin size. Therefore, the fact that it appears large and significant for the bin size plotted in the bottom panel of Figure 11 ($50 bin) is sufficient to reject the null that the entire response is driven by intertemporal substitution.

[31] Consistent with the addition of individual-specific realization noise in this exercise, an alternative, "placebo" exercise, which adjusts for the difference between the individual's year $t+1$ *July* spending and the average monthly spending in August - December of year $t+1$, also make most of the bunching disappear.

8 Conclusions

This paper has explored the spending response to changes in non-linear health insurance contracts. Non-linear contracts are the norm in health insurance, yet most of the prior, voluminous literature on the spending effects of health insurance contracts has tried to summarize the spending response with respect to a single price. In this paper, we instead specify and estimate a dynamic model of drug use decisions made by an optimizing individual facing a specific non-linear budget set.

We do so in the particular context of Medicare Part D prescription drug contracts, which are highly non-linear contracts whose spending impacts are arguably of considerable interest in their own right. We present a simple dynamic model of prescription drug use and descriptive evidence of the two key economic objects of the model. First, we document the presence of a static price response by finding significant bunching of annual spending around the convex kink in the budget set created by the famous Part D "donut hole." Second, we document a dynamic price response by showing that initial drug use is lower for individuals in the same contract who face the same initial "spot" price of drugs but higher expected end-of-year prices.

Estimation of the model allows us to analyze the response to the entire non-linear budget set. As one example, we analyze the impact on spending of "filling of the donut hole" in Medicare Part D, which the 2010 Affordable Care Act legislates will go fully into effect by 2020. We estimate that this policy will increase total drug spending by $180 per beneficiary (or about 10%), and Medicare drug spending by much more ($275 per beneficiary, or about 30%). Beyond the average spending effect, our analysis also allows us to examine some of the subtleties in the behavioral response to a non-linear contract. For example, we find that even individuals whose predicted spending does not reach the gap would still increase their drug use in response to filling the gap, consistent with a dynamic price response. This illustrates that the set of beneficiaries affected by this policy is not limited to those near or in the gap. It also illustrates the importance of estimating a dynamic utilization model, since a static analysis of the utilization response would not capture this effect, which we estimate accounts for almost half of the increase in annual drug spending from filling the gap. We note that while our analyses focuses on what is likely to be the most direct effect of a change in contract design, secondary margins could adjust as well; one potentially important direction for further work would be to extend the analysis of the impact of contract design to incorporate potential supply side responses by insurance firms and potential plan-selection responses by beneficiaries.

The analysis in this paper has been entirely positive. Another important and interesting set of issues for further study concern the normative implications of our findings. In the last section of the paper, we presented additional descriptive results on the nature of the spending response to the kink, examining heterogeneity in the response across groups of individuals and types of drugs, as well as analyzing the response of drug purchase timing and re-timing. Some of the findings there may be useful in beginning to informally assess the normative implications of the drug utilization response. For example, we found evidence that the kink induces a larger reduction in chronic relative to acute drugs, a larger reduction in drug use by healthier individuals, and that some (but not all) of the

reduction in drug use at the kink represents purchases postponed to the following year rather than foregone entirely. Additional evidence on whether there are spillover effects from prescription drug cost-sharing onto non-drug healthcare spending (such as doctor visits and hospitalization) and to health might also be informative on the normative dimension. More formal welfare analysis would also need to take into account the optimality of drug consumption in the absence of insurance. For example, since the policy of granting monopolies through the patent system produces drug prices above social marginal cost, an insurance-induced increase in drug expenditures need not be socially inefficient (Lakdawalla and Sood 2009). Likewise, if one is concerned that incomplete information or potential failures of rationality may lead individuals to under-consume drugs in the absence of insurance, insurance-induced increases in drug consumption may be efficiency enhancing (Baicker, Mullainathan, and Schwartzstein 2012).

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicare Part D program." *American Economic Review* 101(4), 1180-1210.

Abaluck, Jason, Jonathan Gruber, and Ashley Swanson. 2013. "Prescription Drug Utilization under Medicare Part D: A Dynamic Perspective." Presentation slides, MIT.

Alpert, Abby. 2012. "The Anticipatory Effects of Medicare Part D on Drug Utilization." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2161669

Aron-Dine, Aviva, Liran Einav and Amy Finkelstein. 2012. "Moral hazard in health insurance: how important is forward looking behavior?" NBER Working Paper No. 17802.

Aron-Dine, Aviva, Liran Einav and Amy Finkelstein. 2013. "The RAND Health Insurance Experiment, Three Decades Later." *Journal of Economic Perspectives* 27(1), 197-222.

Baicker, Katherine, Sendhil Mullainathan and Joshua Schwartzstein. 2012. "Behavioral Hazard in Health Insurance." NBER Working Paper 18468.

Bajari, Pat, Han Hong, Minjung Park, and Robert Town. 2011. "Regression Discontinuity Designs with an Endogenous Forcing Variable and an Application to Contracting in Health Care." Mimeo, UC Berkeley.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight. 2010.. "Patient Cost-Sharing, Hospitalization Offsets, and the Design of Optimal Health Insurance for the Elderly." *American Economic Review* 100(1), 193-213.

Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics* 126(2), 749-804.

Chetty, Raj, John Friedman and Emmanuel Saez, forthcoming. "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Review.*

Chetty, Raj. 2012. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro

and Macro Evidence on Labor Supply." *Econometrica* 80(3), 969-1018.

Currie, Janet, and Jonathan Gruber. 1996. "Health Insurance Eligibility, Utilization of Medical Care, and Child Health." *Quarterly Journal of Economics* 111(2), 431-466.

CMS, 2011. "Understanding Medicare Enrollment Periods." CMS Product No. 11219. http://www.medicare.gov/Pubs/pdf/11219.pdf.

Decarolis, Francesco. 2012. "What Does Medicare D Share with LIBOR and Procurement Auctions? The Distortionary Effects of the Low Income Subsidy." Mimeo, Boston University.

Duggan, Mark, Patrick Healy, and Fiona Scott Morton. 2008. "Providing Prescription Drug Coverage to the Elderly: America's Experiment with Medicare Part D." *Journal of Economic Perspectives* 22(4): 69-92.

Duggan, Mark and Fiona Scott Morton. 2010. "The Impact of Medicare Part D on Pharmaceutical Prices and Utilization." *American Economic Review* 100(1): 590-607

Einav, Liran, Amy Finkelstein, Stephen Ryan, Paul Schrimpf and Mark Cullen. 2013. "Selection on Moral Hazard in Health Insurance." *American Economic Review* 103(1), 178-219.

Engelhardt, Gary and Jonathan Gruber. 2011. "Medicare Part D and the Financial Protection of the Elderly," *American Economic Journal: Economic Policy*, 3(4): 77-102.

Epstein, Larry G., and Stanley E. Zin. 1989. "Substitution, Risk Aversion, and The Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57(4), 937-969.

Gentzkow, Matthew, and Jesse Shapiro. 2013. "Measuring Sources of Identification in Nonlinear Econometric Models." Mimeo, University of Chicago.

Goldman, Dana, Geoffrey Joyce, Jose Escarce, Jennifer Pace, Matthew Solomon, Marianne Laouri, Pamela Landsman and Steven Teutsch, 1994. "Pharmacy Benefits and the Use of Drugs by the Chronically Ill." *Journal of the American Medical Association* , May 19, 29(19): 2344-2350.

Handel, Ben. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review*, forthcoming.

Hansen, Nikolaus. 2006. "The CMA evolution strategy: a comparing review." In J.A.Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, eds., *Towards a new evolutionary computation. Advances on estimation of distribution algorithms* 192, 75-102.

Hansen, Nikolaus, and S. Kern. 2004. "Evaluating the CMA Evolution Strategy on Multimodal Test Functions." In X. Yao et al. eds., *Parallel Problem Solving from Nature PPSN VIII, LNCS* 3242, 282-291.

Hansen, Nikolaus, Anne Auger, Raymond Ros, Steffen Finck, and Petr Posik. 2010. "Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009." In Proceedings of the 12th annual conference companion on Genetic and evolutionary computation ACM, 1689-1696.

Healthcare Effectiveness Data and Information Set (HEDIS). Washington, DC: National Committee for Quality Assurance, 2010. (http://www.ncqa.org/tabid/59/default.aspx.)

Heiss, Florian, Daniel McFadden and Joachim Winter. 2010. "Mind the Gap! Consumer Perceptions and Choices in Medicare Part D Prescription Drug Plans." *Research Findings in the*

Economics of Aging, David Wise (ed), University of Chicago Press.

Heiss, Florian, Adam Leive, Daniel McFadden and Joachim Winter. 2012. "Plan Selection in Medicare Part D: Evidence from Administrative Data" NBER Working Paper 18166.

Ito, Koichiro (2012). "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." NBER Working Paper 18533.

Joyce, Geoffrey, Julie Zissimopoulos and Dana Goldman. 2013. "Digesting the donut hole." *Journal of Health Economics.* Available on-line at: http://www.sciencedirect.com/science/article/pii/S0167629613

Kaiser Family Foundation 2010, http://kaiserfamilyfoundation.files.wordpress.com/2013/01/8059.pdf

Kaiser Family Foundation 2012a, http://www.kff.org/medicare/upload/7044-13.pdf

Kaiser Family Foundation 2012b,http://www.kff.org/medicare/upload/1066-15.pdf

Kasahara, Hiroyuki, and Katsumi Shimotsu. 2009. "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices." *Econometrica* 77 (1), 135-175.

Ketcham, Jonathan. and Kosali Simon. 2008. "Medicare Part D's Effects on Elderly Drug Costs and Utilization" *American Journal of Managed Care*. November. p14-22

Ketcham, Jonathan, Claudio Lucarelli, Eugenio Miravete and M. Christopher Roebuck. 2012. "Sinking Swimming, or Learning to Swim in Medicare Part D." *American Economic Review* 102(6).

Kleven , Henrik and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics* (128): 669-723.

Kling, Jeffrey, Sendhil Mullainathan, Eldar Shafir, Lee Vermeulen, and Marian Wrobel. 2012 "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127(1).

Kowalski, Amanda (2011). "Estimating the Tradeoff Between Risk Protection and Moral Hazard with a Nonlinear Budget Set Model of Health Insurance." Mimeo, Yale University.

Kreps, David M., and Evan L. Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory." *Econometrica* 46(1), 185-200.

Lakdawalla, Darius and Neeraj Sood. 2009. "Innovation and the Welfare Effects of Public Drug Insurance." *Journal of Public Economics* 93(3-4): 541-548.

Manoli, Day and Andrea Weber. 2011. "Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions." NBER Working Paper No. 17320.

Marsh, Christina (2011). "Estimating Health Expenditure Elasticities using Nonlinear Reimbursement." Mimeo, University of Georgia.

Rios, Luis Miguel, and Nikolaos V. Sahinidis. 2012. "Derivative-free optimization: a review of algorithms and comparison of software implementations." *Journal of Global Optimization*, 1-47.

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2(3): 180-212.

Sasaki, Yuya. 2012. "Heterogeneity and selection in dynamic panel data." Working paper. Available at http://www.sfu.ca/content/dam/sfu/economics/Documents/SeminarPapers/Sasaki_Nov_8_2012_seminar_paper.pdf.

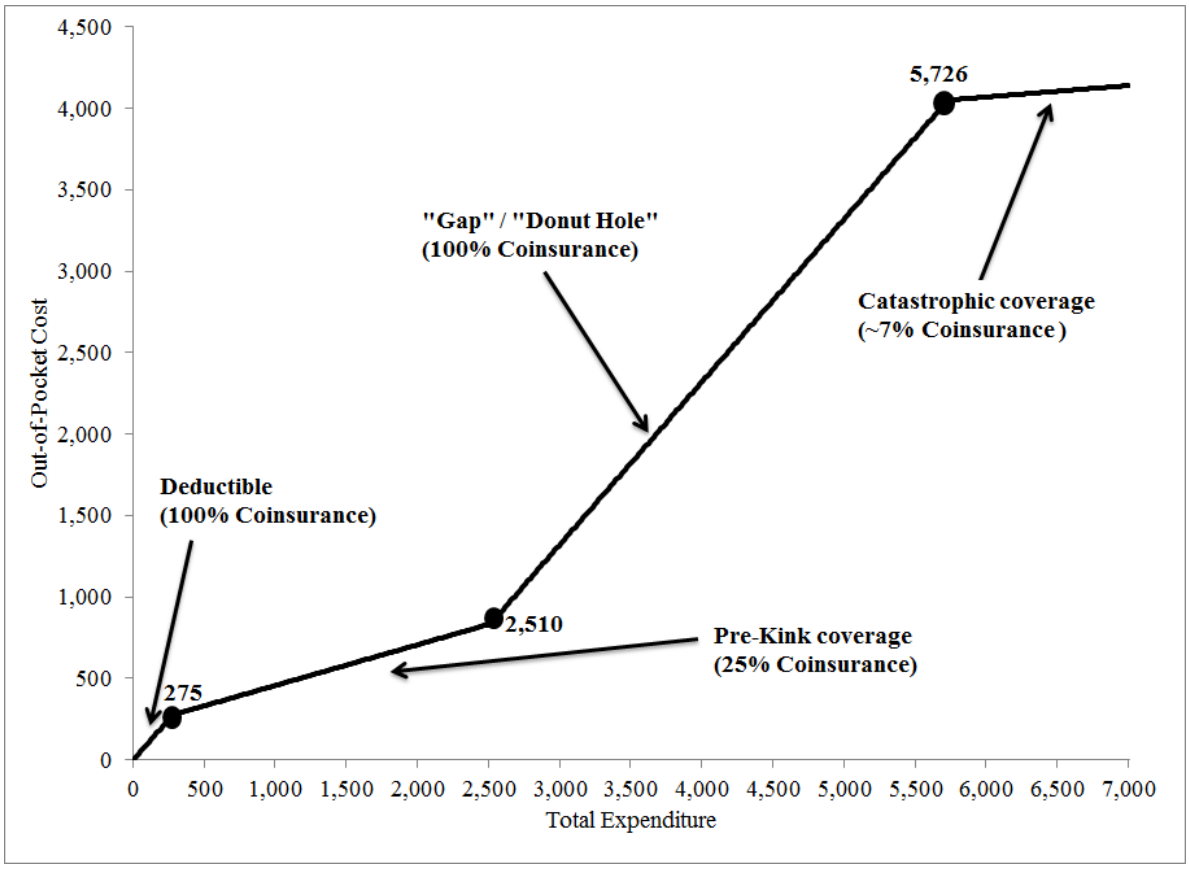
Tur-Prats, Ana, Marcos Vera-Hernanzed, Jaume Puig Junoy. 2012. "Estimates of Price Elasticities of Pharmaceutical Consumption for the Elderly." http://www.homepages.ucl.ac.uk/~uctpamv/papers/TurPrats_etal2012.pdf

Vera-Hernandez, Marcos. 2003. "Structural estimation of a principal-agent model: moral hazard in medical insurance." *RAND Journal of Economics* 34(4): 670-893.

Yin, Wesley, Anirban Basu, James Zhang, Antonu Rabbani, David Meltzer, and Caleb Alexander. 2008. "The Effect of the Medicare Part D Prescription Benefit on Drug Utilization and Expenditures." Annals of Internal Medicine 148: 169-177.

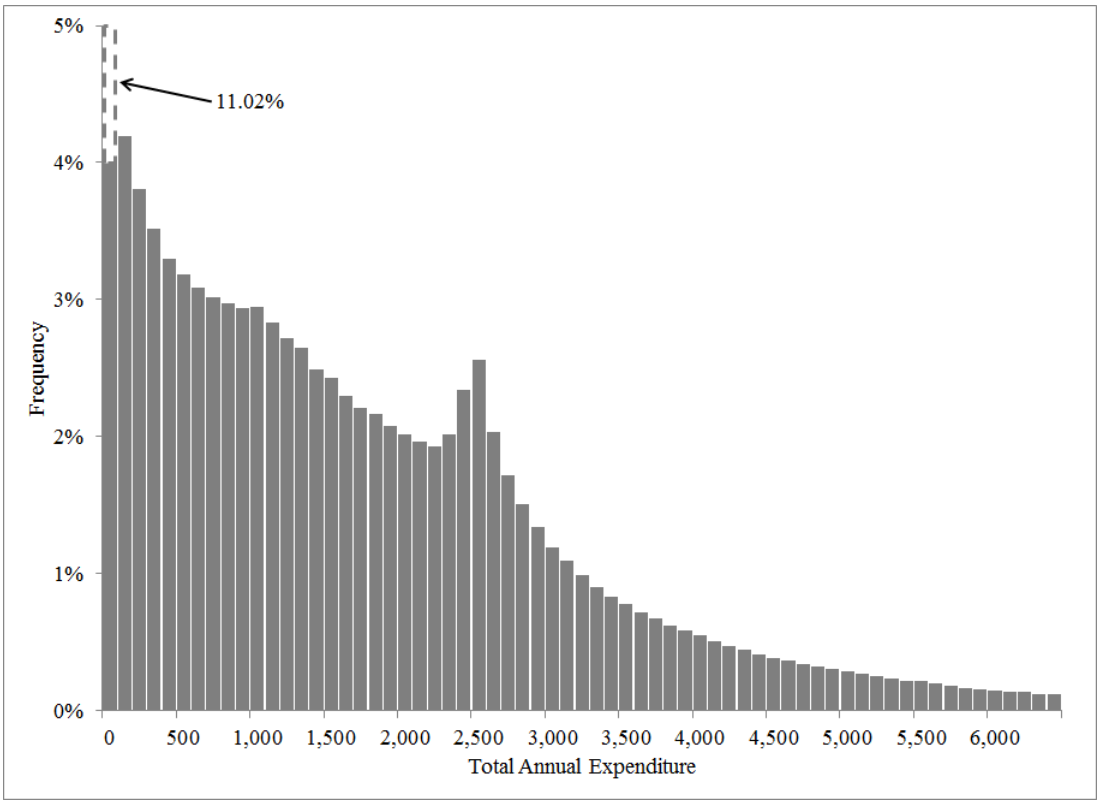
Zhang, Yuting, Katherine Baicker and Joseph P. Newhouse. 2010. "Geographic Variation in the Quality of Prescribing." *New England Journal of Medicine,* 363:1985-1988; Perspective, November 18, 2010.

Figure 1: Standard benefit design (in 2008)



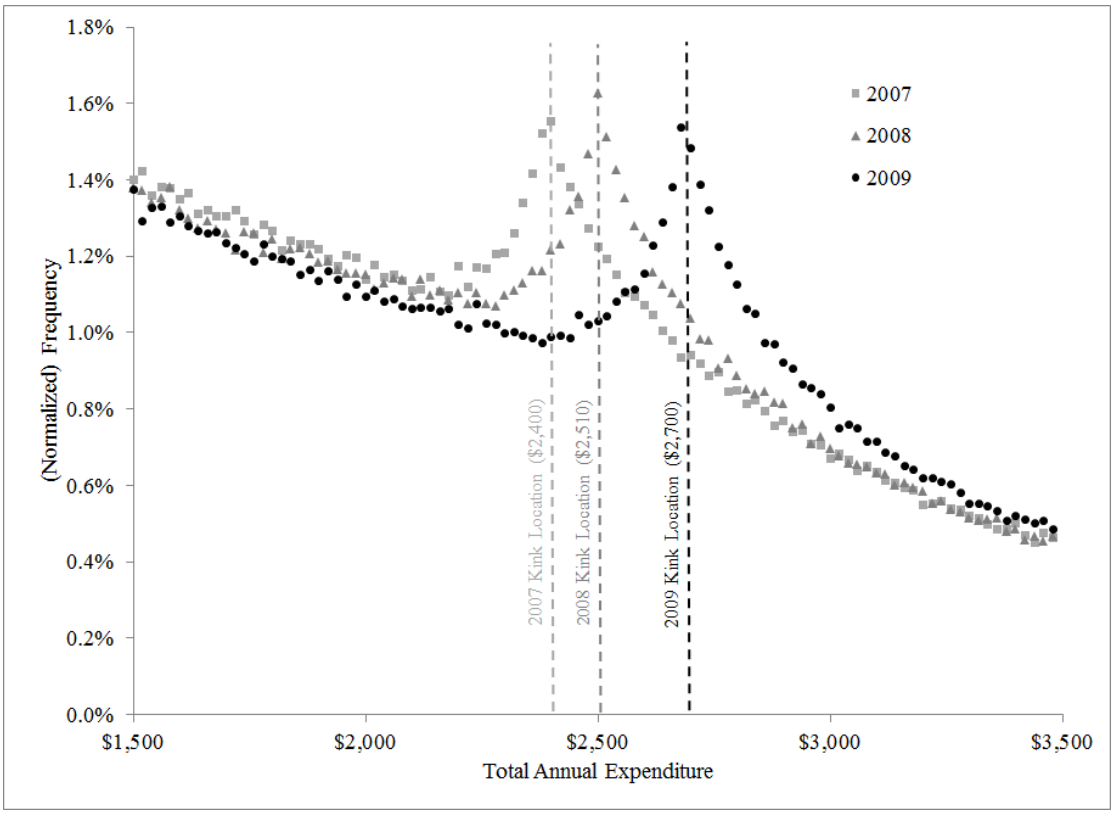
The figure shows the standard benefit design in 2008. "Pre-Kink coverage" refers to coverage prior to the Initial Coverage Limit (ICL) which is where there is a kink in the budget set and the gap, or donut hole, begins. As described in the text, the actual level at which the catastrophic coverage kicks in is defined in terms of out-of-pocket spending (of $4,050), which we convert to the total expenditure amount provided in the figure. Once catastrophic coverage kicks in, the actual standard coverage specifies a set of co-pays (dollar amounts) for particular types of drugs, while in the figure we use instead a 7% co-insurance rate, which is the empirical average of these co-pays in our data.

Figure 2: Annual spending distribution (in 2008)



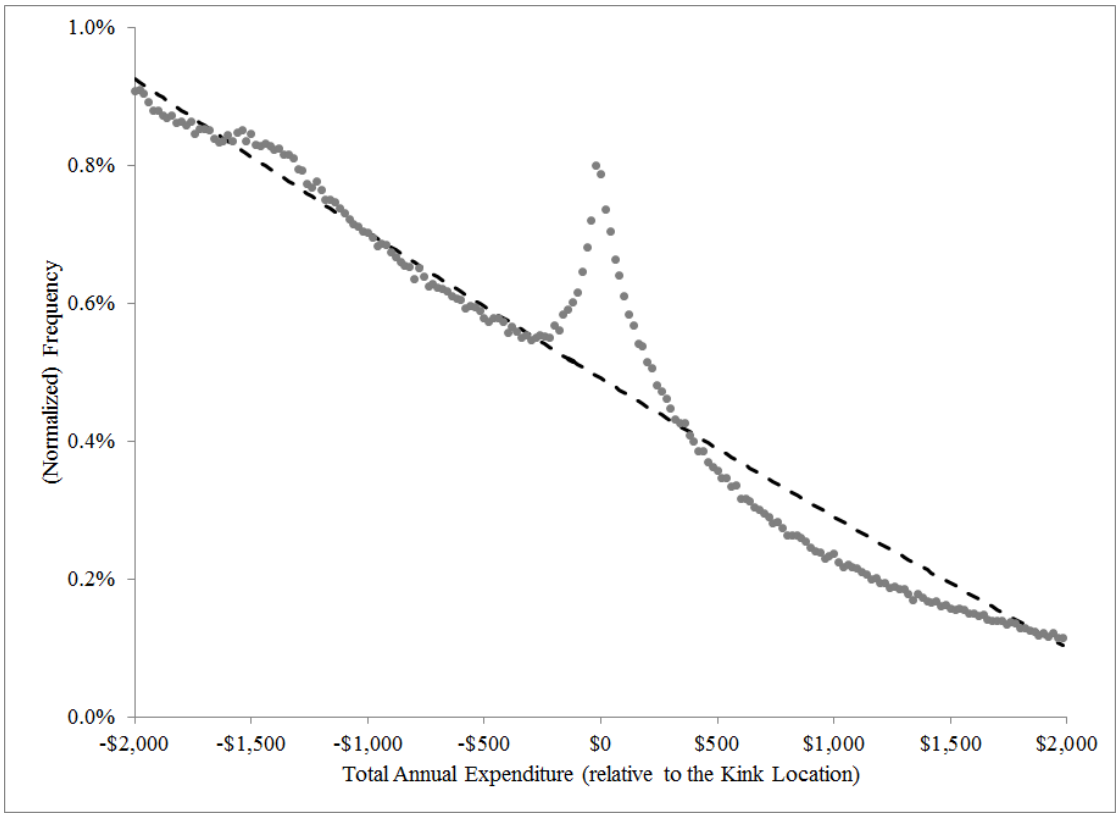
The figure displays the distribution of total annual prescription drug spending in 2008 for our baseline sample. Each bar represents the set of people that spent up to $100 above the value that is on the x-axis, so that the first bar represents individuals who spent less than $100 during the year, the second bar represents $100-200 spending, and so on. For visual clarity, we omit from the graph the 3% of the sample whose spending exceeds $6,500. The kink location (in 2008) is at $2,510. N =1,251,969.

Figure 3: Distribution of spending around the kink, by year



The figure displays the distribution of total annual prescription drug spending, separately by year, for individuals in our baseline sample whose annual spending in a given year was between $1,500 and $3,500 (N=1,332,733 overall; by year it is 447,006 (2007), 442,317 (2008), and 442,410 (2009)). Each point in the graph represents the set of people that spent up to $20 above the value that is on the x-axis, so that the first point represents individuals who spent between $1,500 and $1,520, the second bar represents $1,520-1,540 spending, and so on. We normalize the frequencies so that they add up to one for each series (year) shown.

Figure 4: Magnitude of Excess Mass



Total annual prescription drug spending on the x-axis is reported relative to the (year-specific) location of the kink, which is normalized to zero. Sample uses beneficiary-years in our 2007-2009 baseline sample whose annual spending is within $2,000 of the (year-specific) kink location. The points in the figure display the distribution of annual spending; each point represents the set of people that spent up to $20 above the value that is on the x-axis, so that the first point represents individuals who spent between -$2,000 and -$1,980 from the kink, the second point represents individuals between -$1,980 and -$1,960, and so on. We normalize the frequencies so that they add up to one for the range of annual spending shown. The dashed line presents the counterfactual distribution of spending in the absence of a kink. This is calculated by fitting a cubic CDF function – that is, for each $20 bin of spending (x, y) we fit $F(y) - F(x)$, where $F(z) = a + bz + cz^2 + dz^3$ – using *only* individuals with annual spending (relative to the kink location) between -$2,000 and -$200, and subject to the integration constraints that $F(-2000) = 0$ and $F(+2000) = 1$. N = 2,589,420.

Figure 5: Excess mass by plan

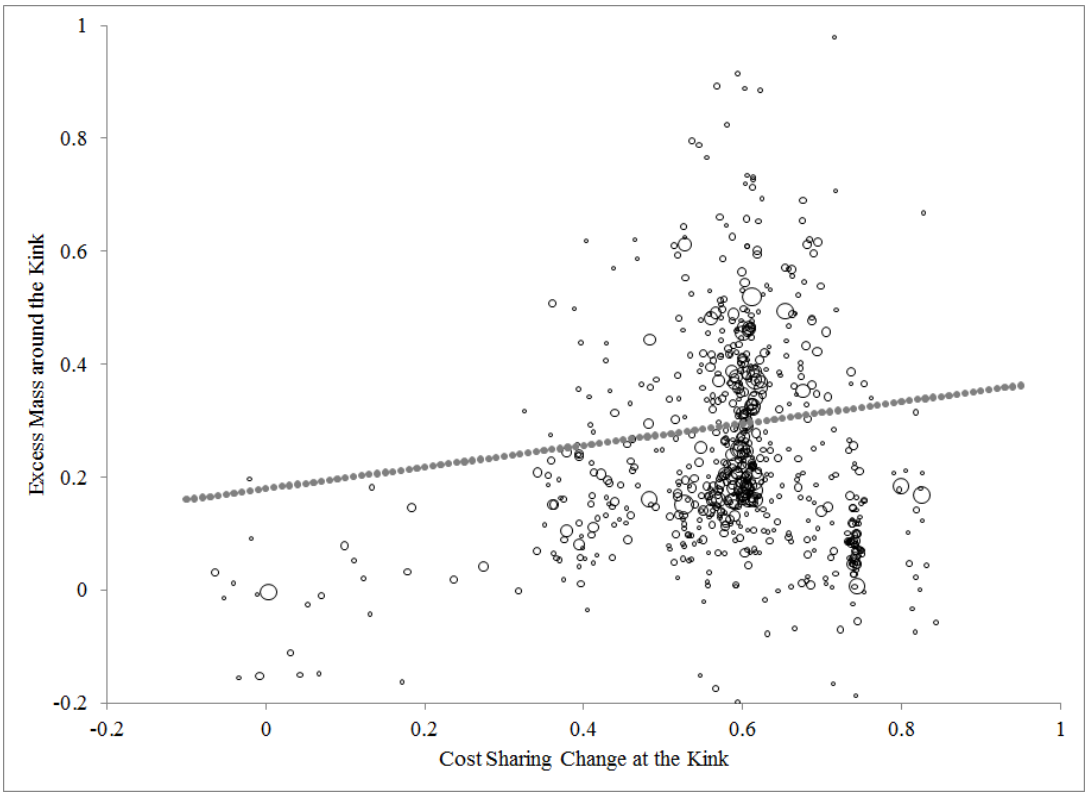


Figure graphs the excess mass in different plans against the size of the kink (i.e. the size of the price increase faced by the consumer as she moves into the gap). The size of the circles is proportional to the number of beneficiaries in the plan. Analysis is limited to the approximately 80% of our baseline sample who are in plans with at least 1,000 beneficiaries within $2,000 of the kink. Excess mass is calculated separately for each plan using the exact same procedure described above for Figure 4. The dashed line in the figure represents the enrollee-weighted regression line of the relationship between excess mass and kink size. N =1,985,676.

Figure 6: Days to first claim and expected end-of-year-price by enrollment month

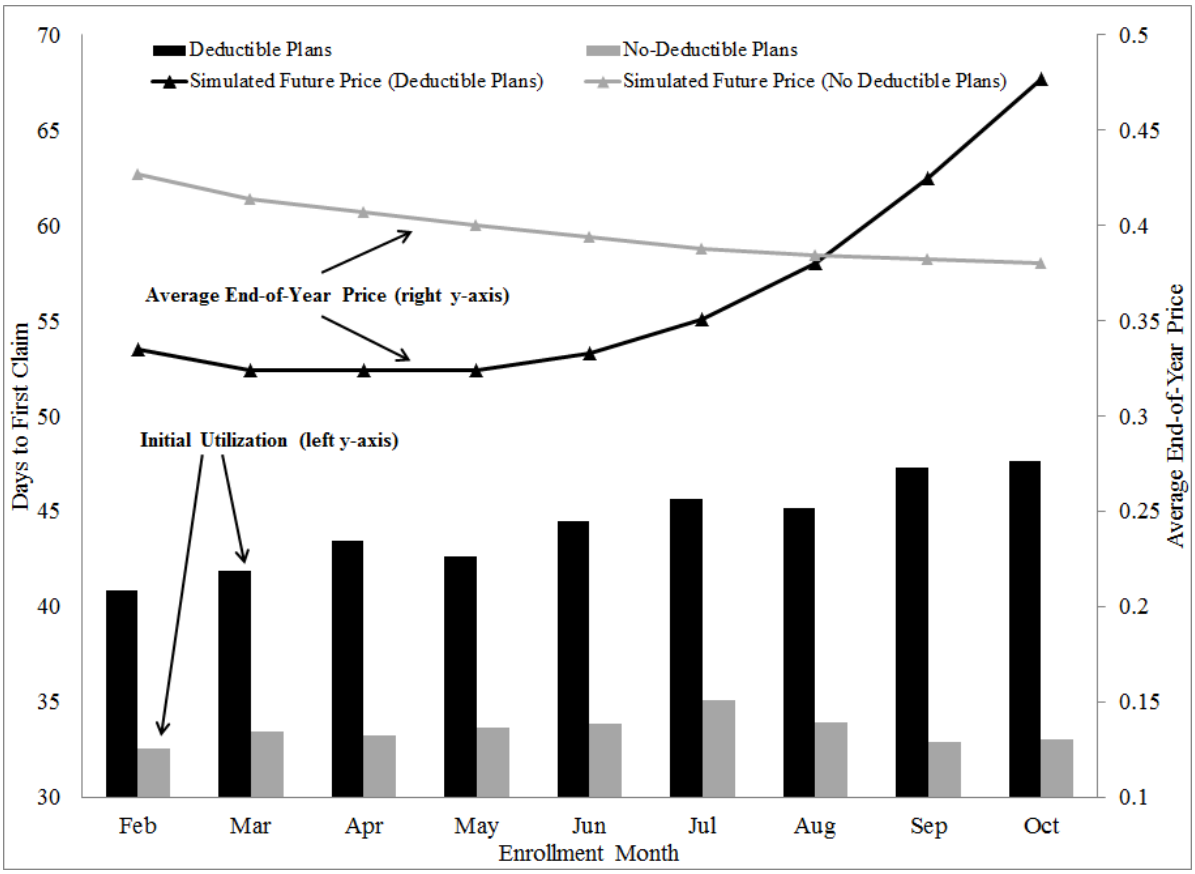


Figure graphs the pattern of expected end-of-year price and of initial drug use by enrollment month for individuals in our 65 year old sub-sample. We graph results separately for individuals in deductible plans and no deductible plans. We calculate the expected end-of-year price separately for each individual based on his plan and birth month, and the same (common) distribution of monthly drug use of all individuals in the 65 year old sub-sample; we refer to this in the text as the "simulated future price." Days to first claim is censored for all beneficiaries at 92 days (about 19% of the sample is censored). See Appendix B for more details on the construction of variables used in this figure. We note that more days to first claim implies lower initial drug use. N =137,536 (N=108,577 for no deductible plans, and N=28,959 for deductible plans).

Figure 7: Model fit

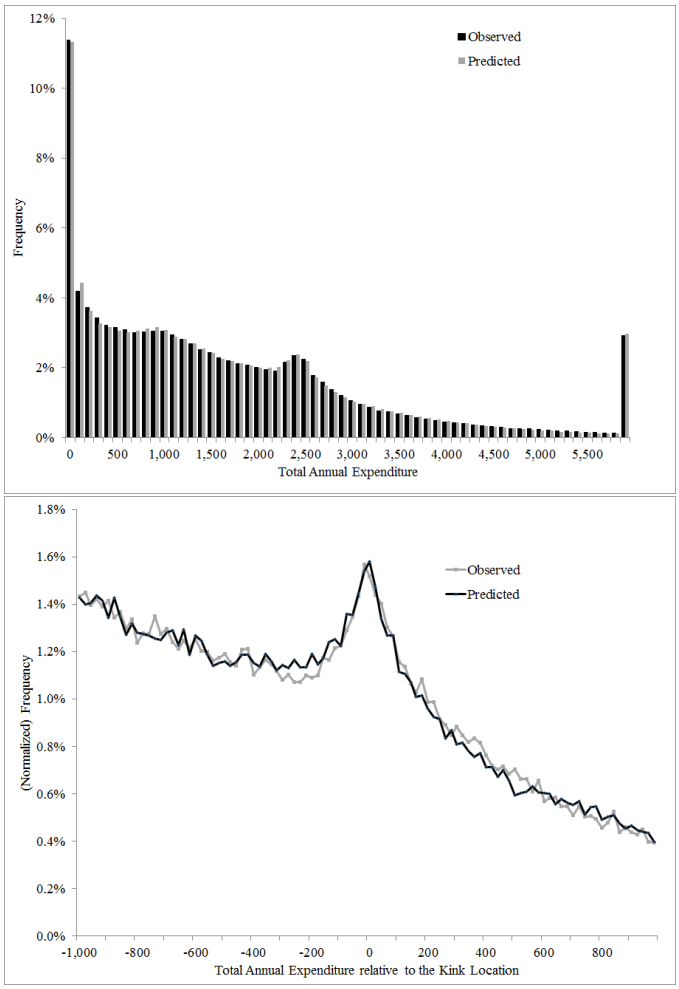


Figure shows the distribution of observed and predicted total annual drug spending. The top panel shows the results for the whole distribution, where each bar represents a $100 spending bin above the value on the x-axis (except for the last bar, which includes all spending above $5,900). The bottom panel "zooms in" on spending within $1,000 of the (year-specific) kink (which is normalized to 0) and shows observed and predicted spending in $20 bins, where each point represents individuals who spend within $20 above the value on the x-axis. Frequencies in the bottom panel are normalized to sum to 1 across the displayed range.

Figure 8: Change in spending from "filling the gap," by pre-change spending level

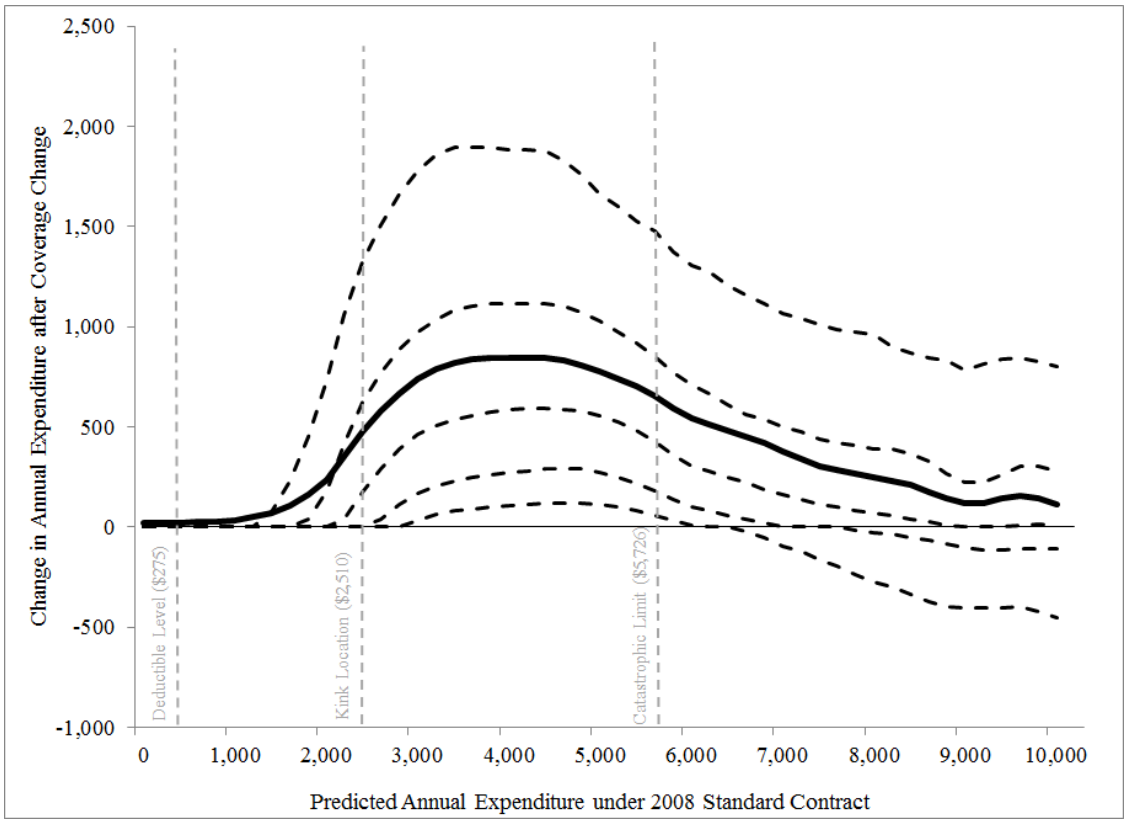


Figure shows the change in spending from "filling the gap" (i.e. providing 25% cost-sharing in the gap) for the 2008 standard benefit (which provides no coverage in the gap) The x-axis shows predicted spending under the 2008 standard benefit. The solid black line shows the mean change in spending for individuals whose predicted spending under the 2008 standard contract is on the x-axis. The dashed lines show the 10th, 25th, 50th, 75th, and 90th percentile changes in spending.

Figure 9: Change in spending from "filling the deductible," by pre-change spending level

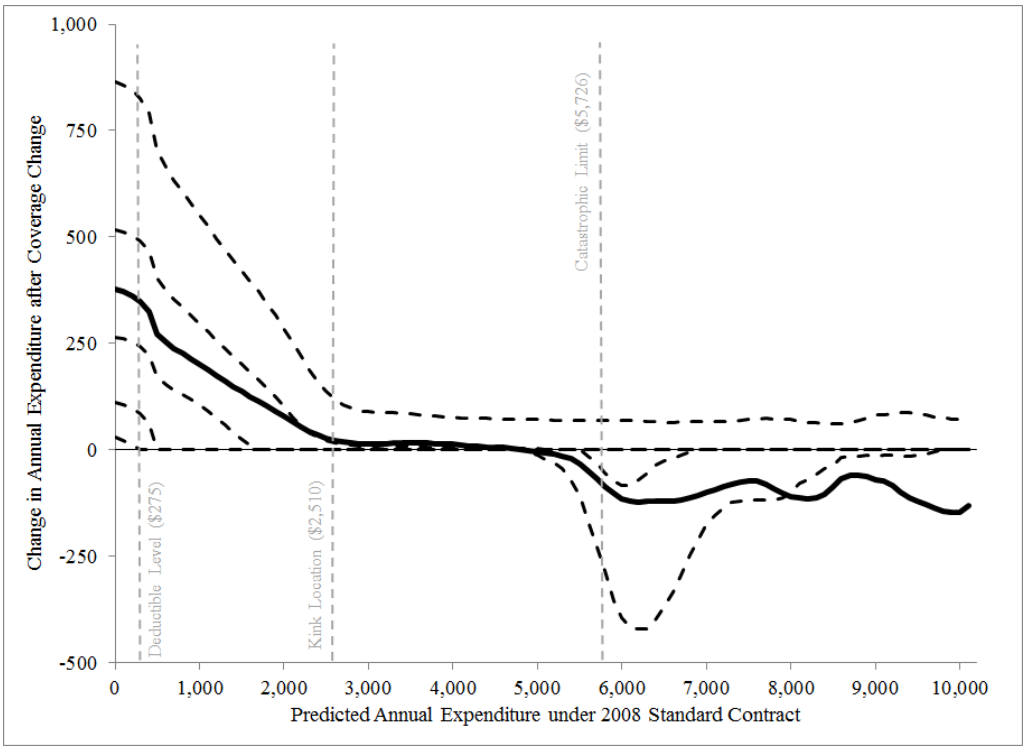
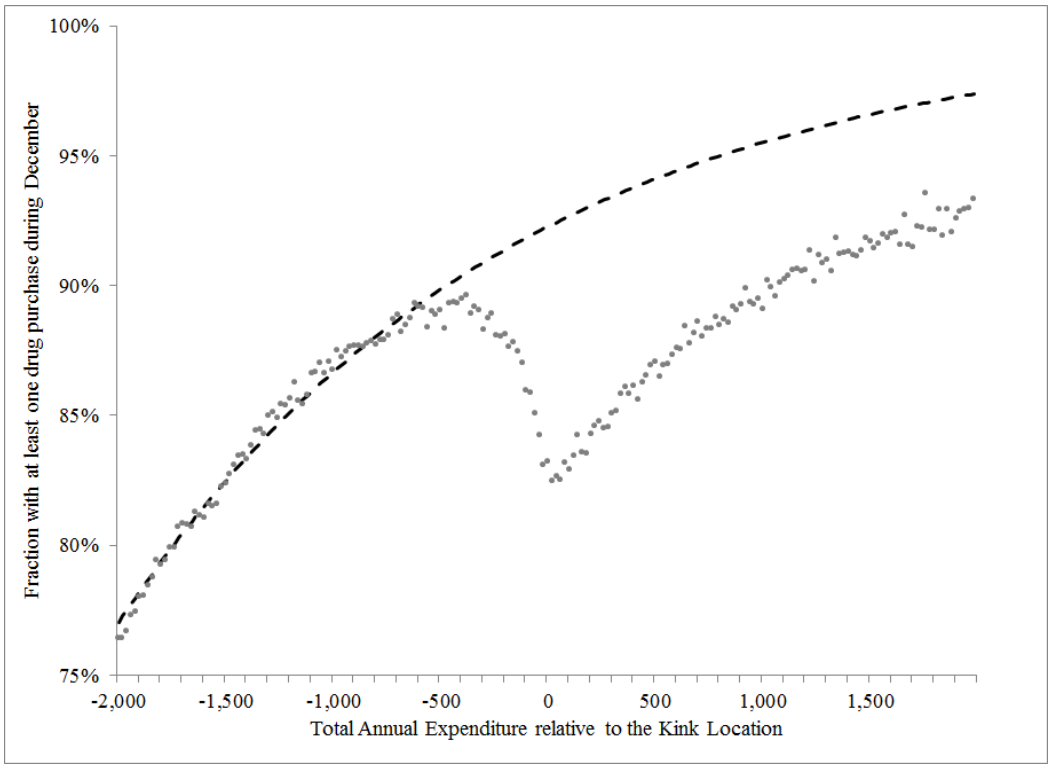


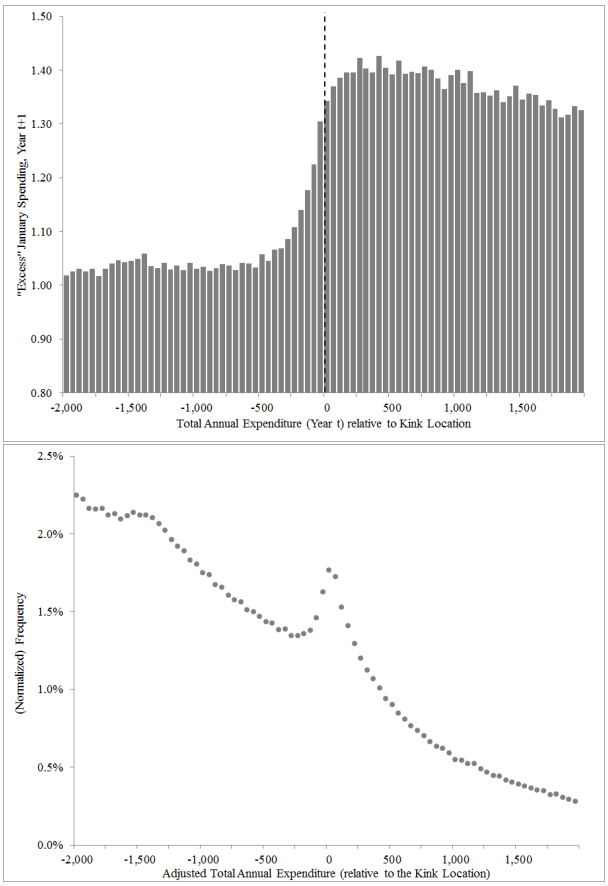
Figure shows the change in spending from "filling the deductible" (i.e. providing 25% cost-sharing in the deductible) for the 2008 standard benefit (which provides no coverage in the deductible) The x-axis shows predicted spending under the 2008 standard benefit. The solid black line shows the mean change in spending for individuals whose predicted spending under the 2008 standard contract is on the x-axis. The dashed lines show the 10th, 25th, 50th, 75th, and 90th percentile changes in spending.

Figure 10: Timing of drug purchases



The figure shows the fraction of individuals who have at least one drug purchase in December as a function of their total annual spending. The x-axis reports total annual spending relative to the (year-specific) kink location, which is normalized to zero. Each point in the graph represents individuals who spend within $20 above the value on the x-axis. N =2,589,420. The dashed line is generated by regressing the logarithm of the share of individuals with no December purchase in each $20 spending bin, using *only* individuals with annual spending (relative to the kink location) between -$2,000 and -$500, on the mid-point of the spending amount in the bin, weighting each bin by the number of beneficiaries in that bin. Data are our baseline sample in 2007-2009 whose annual spending is within $2,000 of the kink location (N=2,589,420).

Figure 11: Inter-temporal substitution



The top panel shows the individual's "excess" January spending in year $t+1$ as a function of her total annual spending (relative to the kink location, which is normalized to 0) in the prior year (year t). "Excess" January spending in year $t+1$ is defined as the ratio of January spending in year $t+1$ to average monthly spending in all other months (of year $t+1$). Each bar on the graph represents individuals within \$50 above the value on the x-axis. The y-axis reports the average, for each year t spending bin, of the "excess January spending" measure. This analysis is limited to individuals in our baseline sample in 2007 and 2008 whom we observe in the subsequent year; we exclude individuals in our baseline sample in 2009 since we do not observe their year $t+1$ spending. N=1,525,678.

The bottom panel shows the bunching analysis (from Figure 4), but now as a function of "adjusted" annual spending. "Adjusted" annual spending is computed by taking total annual spending (in year t) and, for each \$50 bin of annual spending in year t, adding the the average dollar difference between January spending in year $t+1$ and the average monthly spending in February to December of year $t+1$. "Adjusted" annual spending is reported relative to the kink location in year t (which is normalized to 0). The sample shown is once again limited to our baseline sample in 2007 and 2008 whom we observe in the subsequent year (N=1,511,353).

Table 1: Summary statistics

Sample	Full Sample	Baseline Sample	65 y.o. Sub-Sample
Panel A: Demographics			
Obs. (beneficiary years)	16,036,236	3,898,247	137,538
Unique beneficiaries	6,208,076	1,689,308	137,538
Age	70.9 (13.3)	75.6 (7.7)	65 (0)
Female	0.60	0.65	0.60
Risk score[a]	n/a	0.88 (0.34)	n/a
Panel B: Annual Total Spending			
Mean	2,433	1,888	933
Std. Deviation	4,065	2,675	1,618
Pct with no spending	7.35	5.65	12.18
25th pctile	378	487	114
Median	1,360	1,373	513
75th pctile	2,942	2,566	1,240
90th pctile	5,571	3,901	2,355
Panel C: Annual Out of Pocket Spending			
Mean	418	778	325
Std. Deviation	744	968	544
Pct with no spending	14.64	7.11	14.19
25th pctile	29	183	41
Median	144	464	175
75th pctile	476	900	395
90th pctile	1,040	1,971	703

Table shows summary statistics for the full 20% random sample of Medicare Part D beneficiaries (column 1), our baseline sample (column 2), and our 65 year-old sub-sample of the baseline sample (column 3). We show standard deviations (in parentheses). The major restrictions from the full sample to the baseline sample are the exclusion of individuals under 65, dually eligible for Medicaid or other low-income subsidies, or not in stand-alone prescription drug plans. The 65 year-old sub-sample restricts our baseline sample to individuals who are 65 and who joined Medicare Part D between February and October. See text for more details on sample restrictions.

[a] Risk scores are predictions of Part D annual spending using CMS' 2012 RxHCC risk adjustment model (see text for details). They are designed to be 1 on average for Part D beneficiaries. Risk scores in our baseline sample are reported exclusive of 65 year olds, since risk scores for newly enrolling 65 year olds are only a function of a few crude demographics like gender.

Table 2: Cost-sharing features

Sample	Baseline Sample		65 y.o. Sub-Sample	
	Deductible plans	No Ded. plans	Deductible plans	No Ded. plans
Obs. (beneficiary years)	1,036,824	2,861,423	28,960	108,578
Avg. Deductible Amount	265.9	0	257.1	0
Avg. Deductible Coins. Rate	0.88	--	0.85	--
Avg kink location[a]	2,523.0	2,541.7	2,516.4	2,534.6
Avg. pre-kink Coins. Rate	0.26	0.37	0.27	0.37
Pct w/ Some Gap Coverage	0.01	0.17	0.00	0.12
Avg. Gap Coins. Rate (no gap Coverage)	0.88	0.98	0.88	0.98
Avg Gap Coins. Rate (some gap coverage)	0.71	0.77	--	0.76
Avg catastrophic limit (out of pocket)[a]	4,060.0	4,091.8	4,048.3	4,080.4
Catastrophic Coins. Rate	0.07	0.07	0.07	0.07

[a]The kink location is defined based on total expenditures; the catastrophic coverage limit is defined based on out-of-pocket expenditures.

Table 3: Parameter estimates

	j=1	*j=2*	*j=3*	*j=4*	*j=5*
Parameter estimates:					
Beta_0	2.98	-0.51	0.00	2.47	-3.45
	(0.060)	(0.244)	--	(0.082)	(0.154)
Beta_Risk	-2.61	-1.81	0.00	-1.35	4.98
	(0.117)	(0.807)	--	(0.163)	(0.106)
Beta_65	-3.29	-2.21	0.00	-3.99	3.34
	(0.424)	(1.055)	--	(0.437)	(0.384)
δ	---------- 0.929 (0.030) ----------				
μ	3.87	3.93	2.90	5.62	3.97
	(0.021)	(0.020)	(0.009)	(0.011)	(0.020)
σ	1.45	1.41	1.32	0.70	1.49
	(0.094)	(0.110)	(0.046)	(0.020)	(0.017)
p	0.99	0.38	0.89	0.99	0.44
	(0.021)	(0.268)	(0.062)	(0.055)	(0.006)
λ	0.10	0.18	0.85	0.12	0.58
	(0.002)	(0.023)	(0.025)	(0.006)	(0.012)
Implied shares:					
Overall	0.25	0.02	0.15	0.28	0.31
For age=65	0.25	0.02	0.34	0.07	0.31
For age>65	0.24	0.01	0.08	0.35	0.31
Other implied quantities:					
d(Share)/d(Risk)	-0.56	-0.03	-0.07	-0.29	0.94
E(θ)	137	137	44	353	160
Implied annual expected spending:					
Full insurance	714	1,270	1,933	2,144	4,801
0.25 coins. Rate	538	1,151	1,491	1,612	4,272

Top panel reports parameter estimates, with standard errors in parentheses. Standard errors are calculated using the asymptotic variance of the estimates (see equation (10)), with M estimated by the numeric derivative of the objective function. Bottom panels report implied quantities based on these parameters. Note that "health" depends on both the arrival rate of drug events (λ) and the distribution of event size (θ). Spending depends on these parameters as well as on the decision to claim, which is affected by the features of the contract and the parameter p.

Table 4: Model fit

	Observed	Predicted
Annual Total Spending (All individual-year observations)		
Mean	1,757	1,723
Median	1,300	1,299
Std. Dev	2,335	1,887
Percent Spend 0	5.9	5.9
Average weeks to first claim, No deductible		
Feb-April joiners, No deductible plans	4.74	4.73
May-July joiners, No deductible plans	4.84	4.81
August-Oct. joiners, No deductible plans	4.65	4.72
Average weeks to first claim, Deductible		
Feb-April joiners, Deductible plans	5.97	6.01
May-July joiners, Deductible plans	6.10	6.08
August-Oct. joiners, Deductible plans	6.49	6.49

Table reports observed and predicted summary statistics based on model estimation. Model is estimated on the modified baseline sample (described in footnote 16), which limits the baseline sample to the 90 percent of individuals in the 500 most common plans and then retains the entire 65-year-old sub-sample plus a 10 percent random draw of the remaining baseline sample; each observation in the 10 percent random draw is weighted by 10. Top panel shows results for this modified baseline sample (N=482,448). In the bottom two panels, sample is limited to the 65-year-old sub-sample of the modified baseline sample (N=125,463).

Table 5: Spending effect of alternative insurance contract design

		Mean	Std. Dev.	25th pctile	Median	75th pctile	90th pctile	Mean OOP	Mean Insurer
1	Baseline (Standard 2008 contract)	1,710	1,928	243	1,356	2,503	3,607	785	925
2	"Filling" the gap[a]	1,955	2,190	270	1,388	2,890	4,617	646	1,309
3	"Filling" the Deductible[b]	1,882	1,836	695	1,549	2,531	3,625	678	1,204
4	Baseline (Actual contracts)	1,744	1,895	495	1,320	2,435	3,632	777	967
5	"Filling" the gap[a]	1,925	2,079	501	1,345	2,729	4,406	682	1,242
6	"Filling" the Deductible[b]	1,783	1,874	560	1,366	2,444	3,636	754	1,029
7	Full insurance[c]	2,571	2,251	1,018	2,077	3,520	5,246	0	2,571
8	No Insurance[d]	599	1,291	0	0	812	1,896	599	0
9	Only catastrophic coverage[e]	965	1,975	0	0	1,407	3,047	805	159

Table reports the predicted annual drug spending under various observed and counterfactual contracts. All columns report total annual drug spending except the right-most two which separately report out of pocket and insurer spending. Rows 1-3 report predicted spending under the standard contract in 2008, which was illustrated in Figure 1, and counter-factual changes to it. Rows 4-6 report predicted spending for the observed contracts in our sample, and counterfactual changes to them. Rows 7-9 report predicted spending under additional counterfactual contracts. For all of the simulations, we assume individuals are in the contract for a full 12 months. (Predicted mean spending for observed contracts - row 4 - is slightly higher than the estimate reported in Table 4 because of the assumption here that everyone is in the contract for 12 months).

[a] "Filling the gap" means that, above the deductible, the plan now has a constant co-insurance rate, without a kink, until out-of-pocket expenditure hits the catastrophic limit. For each plan, we use the observed (pre-kink) co-insurance rate (which is 25% in the 2008 standard benefit plan.). For the less than 1% of plans where our calculated pre-kink co-insurance rate is higher than our calculated co-insurance rate in the gap, we do not adjust cost-sharing in the counterfactual.

[b] "Filling the deductible" means reducing the deductible amount to zero, so that the observed (pre-kink) co-insurance rate (which is 25% in the 2008 standard benefit plan) kicks in with the first dollar of spending.

[c] "Full Insurance" means 0% consumer co-insurance everywhere.

[d] "No Insurance" means 100% consumer co-insurance everywhere.

[e] "Only catastrophic coverage" means 100% consumer co-insurance up to the catastrophic coverage limit.

Table 6: Implied price elasticities

(Uniform) Price Reduction[a]	Average Annual Spending	Implied "Elasticity"[b]
0% (Baseline)	1,710	—
1.0%	1,722	-0.73
2.5%	1,736	-0.60
3.0%	1,740	-0.59
3.5%	1,745	-0.58
5.0%	1,757	-0.56
7.5%	1,779	-0.54
10.0%	1,800	-0.53
15.0%	1,841	-0.51
25.0%	1,921	-0.49
50.0%	2,123	-0.48
75.0%	2,338	-0.49

Table shows the model's estimate of the impact of various changes to the 2008 standard benefit budget set (shown in Figure 1). The first row shows predicted average annual spending under the existing budget set. Other rows show predicted average annual spending (and the implied "elasticity") of various *uniform* price reductions to this budget set.

[a] "Uniform price reduction" is achieved by reducing the price (i.e. consumer coinsurance) in every arm of the 2008 standard benefit by the percent shown in the table.

[b] The implied "elasticity" is calculated by computing the ratio of the percent change in spending (relative to the baseline) to the percent change in price (relative to the baseline).

Table 7: Robustness

	Total Spending			Insurer Spending		
	Basline	"Filled" Gap	Change	Basline	"Filled" Gap	Change
1 Baseline model	1,744	1,925	10.4%	967	1,242	28.4%
Number of types:						
2 Three types	1,743	1,872	7.4%	966	1,208	25.1%
3 Six types	1,738	1,902	9.4%	961	1,226	27.6%
Different sets of covariates:						
4 Remove all covariates	1,762	1,901	7.9%	979	1,230	25.6%
5 Add "no gap" covariate	1,737	1,942	11.8%	965	1,254	29.9%
Concave (risk averse) utility function:						
6 CARA = exp(-6.5)	1,718	1,893	10.2%	941	1,214	29.0%
7 CARA = exp(-9.5)	1,723	1,872	8.6%	948	1,202	26.8%

Table shows how the predictions of the model are affected by different specifications of the economic or econometric model. The first row presents the results from the baseline model (as reported in rows 4 and 5 of Table 5). Row 2 and 3 report results in which we explore the sensitivity of the results to the number of discrete types (M): the baseline model assumes $M = 5$, while row 2 assumes $M = 3$, and row 3 assumes $M = 6$. Row 4 reports results in which we do not use any covariate z_i to estimate the propensity of each individual to be of each type, while row 5 uses the baseline covariates (a risk score, and an indicator for a 65 year old) and also adds an additional covariate, an indicator for a beneficiary who selected a plan with no gap coverage. Rows 6 and 7 report results that are based on estimating a recursive utility model that allows for risk aversion as described in the main text (relative to the risk neutrality assumption of the baseline model). The two values of the (absolute) risk aversion parameter are imposed, such that they span the range of risk aversion estimates reported in Handel (2013).

Table 8: Heterogeneity in response across drug types

Drug Type	Percent of purchases	Percent of spending ($)	Actual P(Dec. Purchase)	Predicted P(Dec. Purchase)	Percent decrease in purchase probability
All	100.0	100.0	0.846	0.922	0.083 (0.001)
Chronic	69.6	77.1	0.762	0.867	0.121 (0.001)
Acute	30.4	22.9	0.532	0.590	0.099 (0.002)
Maintenance	85.1	90.1	0.807	0.899	0.103 (0.001)
Non-Maintenance	14.9	9.9	0.303	0.330	0.080 (0.004)
Brand	33.4	74.8	0.624	0.780	0.199 (0.001)
Generic	66.6	25.2	0.731	0.798	0.085 (0.001)
"Inappropriate" [a]	2.7	1.3	0.075	0.085	0.119 (0.008)

Table shows overall, and then separately by drug type, the change in the probability of a purchase in December around the kink location. The analysis is done on the same sample analyzed in Figure 10 (N=2,589,420). The "actual" probability of a December purchase is the probability someone whose annual spending is within $200 of the kink location fills at least one prescription in December. The "predicted" probability of a December purchase within $200 of the kink location is designed to reflect what the probability of a December purchase would have been in the absence of the kink. It is estimated (as in Figure 10) by regressing the logarithm of the share of individuals with no December purchase in each $20 spending bin (between -$2,000 and -$500) on the mid-point of the spending amount in the bin, weighting each bin by the number of beneficiaries in that bin. When looking separately by drug type, the "actual" and "predicted" probabilities are calculated based on the probability of a purchase for a specific drug type. The right-most column shows the percent decrease in purchase probability, i.e. the predicted minus actual December purchase probability as a percent of the predicted. Bootstrapped standard errors (in parentheses) are calculated by estimating the actual and predicted probability of a December purchase of a given type based on 500 bootstrap replications.

[a] Following Zhang et al. (2010), we proxy for inappropriate drug use using an indicator from the Healthcare Effectiveness Data and Information Set (HEDIS) on whether the drug is considered high-risk for the elderly (HEDIS, 2010).

Table 9: Heterogeneity in response across individuals

Population	Share of Sample	Share of Spending	Excess Mass
All	100.0	100.0	0.291 (0.003)
Year[a]			
2006	n/a	n/a	0.088 (0.008)
2007	n/a	n/a	0.150 (0.008)
2008	n/a	n/a	0.213 (0.009)
2009	n/a	n/a	0.293 (0.010)
Gender			
Male	35.0	34.3	0.348 (0.005)
Female	65.0	65.7	0.262 (0.004)
Age group			
66	5.4	4.6	0.519 (0.016)
67	5.7	4.9	0.426 (0.015)
68-69	11.0	9.8	0.383 (0.009)
70-74	24.0	23.0	0.334 (0.006)
75-79	19.2	20.2	0.255 (0.006)
80-84	15.4	17.6	0.194 (0.006)
85+	14.8	17.8	0.136 (0.007)
Number of Hierarchical Condition Categories[b]			
0	15.5	6.6	0.837 (0.020)
1	8.9	5.0	0.494 (0.018)
2	14.9	10.7	0.191 (0.008)
3	17.6	16.2	0.197 (0.006)
4	15.0	16.9	0.236 (0.007)
5+	28.1	44.6	0.316 (0.005)
Risk Score Quartile			
1 (healthiest)	25.1	11.9	0.448 (0.011)
2	25.1	19.3	0.155 (0.005)
3	24.9	26.7	0.250 (0.005)
4 (least healthy)	25.0	42.2	0.346 (0.005)

Table investigates the excess mass at the kink separately for different populations. Excess mass is calculated, separately for each group. For a given group of individuals, we compute the number of people within $200 of the kink and estimate (counterfactually, using the approached described in Figure 4) how many people (in that group) would have been in that range in the absence of the kink. Our "excess" mass estimate is the percentage increase in the people observed at the kink relative to the number we estimate would be there in the absence of the kink. Bootstrapped standard errors (in parentheses) are calculated based on 500 replications of the bootstrap. Row 1 shows the results for the baseline sample (from Figure 4; N =2,589,420). Subsequent rows show results for the indicated sub-samples.

[a] For the analysis by year we add in the first year of the Part D program (2006); the results "by year" are shown for the sub-sample of approximately two-fifths of individuals who joined in January 2006 and remain in the sample for complete years through 2009.

[b] The hierarchical condition categories are inputs in the CMS risk score; they are meant to capture conditions that are predictive of higher drug spending in the next year.

NOT FOR PUBLICATION APPENDICES

Appendix A: Additional bunching analysis

Plans without a standard kink location Our baseline sample consists of individuals with a standard kink location. A small sample of individuals excluded from the baseline sample have a kink at an amount that is different from the standard level. The modal non-standard kink amount is $2,100; most of these plans are in 2007 or 2008. Figure A2 shows, as expected, that for individuals in plans with the $2,100 kink location, there is evidence of excess mass around $2,100 but not at the standard kink locations. Naturally, the figure is somewhat noisier than the baseline analyses that use the considerably larger baseline sample.

Distribution of spending around the deductible The same standard economic theory that generates bunching at the (convex) kink as individuals enter the gap, should also generate "missing mass" at the concave kink created by the sharp price decreases when individuals hit the deductible amount or hit the catastrophic coverage limit (see Figure 1). It is difficult to analyze the distribution of spending around the catastrophic limit.[32] Figure A3, however, shows no evidence of such missing mass around the deductible level for individuals in plans with the standard deductibles.[33]

This finding of excess mass (bunching), but not missing mass, is mirrored in the labor supply context where previous research has similarly found excess mass in annual earnings in convex kinks but not missing mass at concave kinks (Saez 2010). One potential rationale for the bunching at the gap but the lack of "missing mass" at the deductible amount might be that it is easier to stop (or delay) utilization in response to an increase in price at the gap than it is to increase (or speed up) utilization because of an anticipated decrease in price if one were to hit the deductible level. It would be interesting to see if this lack of missing mass at non-convex kinks is a broader phenomenon, and if so to understand why. In the context of health insurance, typical contracts specify a price that is decrease in total spending, so that most of the generated kinks are non-convex. Some health insurance contracts, however, have convex kinks, such as high-deductible Health Reimbursement Accounts, where the price the consumer faces increases discontinuously when the employer contribution to help cover the deductible is exhausted (Lo Sasso et al. 2010).

Appendix B: Testing for dynamic response

As described in Section 4, we utilize variation in the birth month of beneficiaries, which creates variation in coverage duration during the first year of eligibility, to examine whether individuals

[32] Analysis of the spending distribution around the catastrophic limit is made noisy for two reasons. First, few people spend enough to put them in the range of the catastrophic limit, so sample sizes are small. Second, the catastrophic limit is a function of out-of-pocket spending, not total spending. However, the distribution of out-of-pocket spending changes mechanically when cost-sharing changes. We therefore would need to analyze the distribution of total spending around the catastrophic limit, but the mapping (from out-of-pocket spending to its associated total spending) introduces additional noise. Therefore, although we find no evidence of missing mass at the catastrophic limit, given these data issues we do not consider the result particularly informative.

[33] We exclude from the analysis the roughly 10% of people in plans with a (non-zero) deductible that is not the standard deductible level. As with the location of the kink. the level of the deductible is set differently each year in the standard benefit. It is $265 in 2007, $275 in 2008, and $295 in 2009.

respond to the non-linearity of the contract. Specifically, we test for a dynamic response by comparing the pattern of initial prescription drug use within a plan across newly-enrolled 65 year old beneficiaries who turn 65 at different points in the year. This creates variation in the expected end-of-year price (or future price) across individuals who face the same initial, spot price for drugs. This allows us to empirically test the null that individuals do not respond to the future price. This appendix presents the analysis and its results in more detail.

Analytical framework

The 65 year old sub-sample Given the identification strategy, we limit the baseline sample to 65 year olds, thus generating variation in join month based on birth month. (Almost all existing enrollees over 65 can change plans only during the fall open enrollment period for coverage starting January 1). For analytical tractability, we remove the small number (<0.25%) of individuals who choose a deductible that is less than $100 or who choose a deductible plan with gap coverage. As in Aron-Dine et al. (2012), we further limit our analysis to individuals who join between February and October.[34]

These restrictions produce the "65 year old sub-sample" described in Tables 1 and 2. It consists of about 137,000 beneficiary-years, a substantial restriction relative to our baseline sample of 3.9 million beneficiary-years. Not surprisingly, Table 1 shows average annual prescription drug spending substantially lower in the 65 year old sub-sample. This reflects both that the 65 year olds are the youngest individuals in the baseline sample (and therefore on average the healthiest) and also that, because we restrict to February-October enrollees, we observe "annual" spending in this sample for only about six months on average. Our analysis below will condition on the plan, which we will index by j. We observe 3,575 different plans covering our 65 year old sub-sample.

An assumption of our empirical strategy is that individuals face different contract durations depending on which month of the year they were born. Table B1 corroborates this, showing the relationship between birth month and enrollment month for our 65 year old sub-sample. The vast majority (over 70%) of our sample enrolls in their birth month. Virtually no one (less than 2%) enrolls prior to their birth month (this 2% presumably reflects measurement error in our data or some idiosyncratic circumstances). About one-quarter enrolls after their birth month (usually shortly thereafter), presumably reflecting some delay in signing up. In the empirical work below we will often instrument for enrollment month with birth month.

Measuring future price For purposes of our descriptive work, we define the future price to be the expected end-of-the year price.[35] The expected end-of-year price depends on three elements:

[34] We exclude November and December enrollees because we want to observe our initial utilization measure over a reasonable time horizon. We exclude January enrollees because empirically they turn out to conflate both individuals whose birth month is in January with a reasonable number of people who join in January for other idiosyncratic reasons.

[35] If individuals are risk neutral, this is the only moment of the future price that should matter for their utilization decisions. In practice, individuals may not be risk neutral and other moments of the end-of-year price may affect initial utilization. Limiting our analysis to the impact of the expected end-of-year price can therefore bias us against

the cost-sharing features of the beneficiary's plan, the duration (number of months) of the contract, and the expected spending of individuals.

For illustrative purposes, Table B2 shows how the fraction of individuals ending up in different cost-sharing arms varies by enrollment month. We show this pattern separately for deductible and no-deductible plans. We see, for example, in the deductible plan that the fraction still in the deductible (high cost sharing) arm at the end of the calendar year is increasing in enrollment month; this is what drives the pattern of increasing future price with enrollment month in the deductible plans. In the no-deductible plans, the fraction in the (high cost sharing) gap at the end of the calendar year is decreasing in enrollment month, which is what drives the pattern of decreasing future price with enrollment month in the no-deductible plan.

In practice, we calculate the future price $fp_{j,m}$ separately for each plan j and enrollment month m in the sample. Let $\Pr(j, m, a)$ denote the probability an individual who enrolls in plan j in month m ends up in the cost sharing arm a at the end of the year, and let $c_{j,a}$ denote the consumer cost-sharing rate for plan j in arm a. We calculate the empirical analog of $\Pr(j, m, a)$ using the data on the fraction of individuals who enrolled in plan j in month m and ended up at the end of the calendar year on each arm a. We calculate $c_{j,a}$ using the data for the baseline sample on the ratio of out-of-pocket spending to total spending for each plan-cost sharing arm (as described in the main text). Thus, we have:

$$fp_{j,m} = \sum_{a \in A} \Pr(j, m, a) \cdot c_{j,a}, \tag{12}$$

where $A = \{ded, pre-kink, gap, catastrophic\}$. Thus, the future price is the mean of the realized end-of-year cost sharing for each plan j and enrollment month m.We describe below a related variable ("simulated future price") that we use to instrument for the future price in some of our analyses.

Measuring initial drug use The key to our empirical strategy is that the spot price is *initially* constant within a plan across enrollment months while the future price varies. Therefore, we need a measure of initial drug use for which the spot price is the same. The most natural measure is therefore a measure of time to first claim, since the spot price should be the same for everyone within a plan at the time of their first claim.[36]

We analyze the distribution of log days to first claim through the first 91 days (the maximum time we observe everyone regardless of enrollment month). Over 80% of our sample has a claim within the first 92 days (not surprisingly, this fraction is lower for those in deductible plans (71%) than those in no deductible plans (84%)). Conditional on having a claim, the average days to first claim is about 23 days. We assign individuals without a claim in the first 91 days a time to a first claim of 92 days.

rejecting the null of no response to the future price, which is the exercise here.

[36] The only exception to this would arise due to the lumpiness of initial claims (e.g. if a plan has a $100 deductible and the drug costs $150).

Estimating equations Our main estimating equations analyze the relationship between a measure of initial drug use and enrollment month, or initial drug use and the future price (with variation in the future price being driven by the plan and enrollment month).

The simplest way by which we can implement our strategy is to look within a given plan type (such as deductible plans) and regress log days to first claim (y_i) on join month (m), which runs from 2 (February) to 10 (October), and plan fixed effects (α_j).[37] We estimate:

$$y_i = \beta m_i + a_j + u_i. \tag{13}$$

Equation (13) analyzes how initial drug use varies with enrollment month, conditional on plan fixed effects.[38] The plan fixed effects control for any fixed difference in initial drug use across plans; plans differ in, among other things, their cost sharing in the pre-kink arm and in the gap, and standard selection effects (or effects of the spot price for no deductible plans) could therefore generate differences in initial drug use across plans. Our analysis focuses on whether, within plans, initial drug use varies by enrollment month.

Recall that future price is increasing in enrollment month for the deductible plan (see e.g. Table B2). Therefore, absent any confounding influences of join month on y_i, we would expect an estimate of $\beta = 0$ for deductible plans if individuals are fully myopic and $\beta > 0$ if individuals are not. If individuals exhibit a dynamic response, then initial drug use should be decreasing in enrollment month and therefore time to first claim should be increasing.

If individuals' health varies by birth month or if seasonality in drug use is an important factor, it could confound the interpretation of β in the above equation. We address this concern by contrasting the pattern of initial drug use by enrollment month in deductible plans with that in no-deductible plans, where the future price is decreasing (rather than increasing) with enrollment month (again, see e.g. Table B2). Formalizing this in a difference-in-difference analysis we examine:

$$y_i = \beta' m_i D_j + \alpha_j + \tau_m + \upsilon_i, \tag{14}$$

where τ_m are enrollment-month fixed effects, and D_j is a dummy variable that is equal to one when plan j is a deductible plan. The enrollment month fixed effects control flexibly for any pattern of initial drug use by enrollment month that is common across plans, and the approximately 3,500 plan fixed effects control for fixed differences in initial drug use across plans, regardless of enrollment month. Again, our coefficient of interest is β', where $\beta' = 0$ would be consistent with the lack of response to dynamic incentives (i.e., full myopia) while $\beta' > 0$ implies that the evidence is consistent with dynamic response.

One limitation to the analysis in equation (14) is that it uses variation by enrollment month, which is in principle a choice. We can purge the choice element of this variation by estimating an

[37] Van der Berg (2001) discusses the trade-offs involved in analyzing a duration model using a linear model with the logarithm of the duration as the dependent variable, as we do here, relative to a proportional hazard model. As he explains, neither model strictly dominates the other.

[38] Because the features of a plan can change slightly from year to year, we define a "plan" at the yearly level. Thus, we have separate fixed effects for each plan in each year.

instrumental variable version of equation (14) in which we instrument for enrollment month with birth month. Specifically, we use as instrumental variables a series of birth month dummies and an interaction of (linear) birth month with a deductible dummy. Not surprisingly, given the patterns seen in Table B1, the relationship between birth month and enrollment month is quite strong.[39]

Another limitation to the analysis in equation (14) is that it constrains the relationship between enrollment month and initial drug use to be linear in enrollment month and to be constant across different deductible plans or across different no-deductible plans. In practice, however, the relationship between enrollment month and future price will vary within these broad plan types depending on the specific plan details, nor is the relationship necessarily linear (or in some cases even monotonic). To account for these features, as well as to provide an estimate with more of an economic interpretation, we will estimate regressions with the future price on the right hand side rather than enrollment month. Specifically, we will estimate

$$y_i = \widetilde{\beta} fp_{j,m} + \alpha_j + \tau_m + \widetilde{u}_i, \tag{15}$$

where (as before) y_i is log days to first claim, a_j are plan fixed effects, and τ_m are enrollment-month fixed effects. Equation (15) thus compares initial drug use across individuals within the same plan, controlling for a flexible relationship between initial utilization and enrollment month that is common across all plans. Variation in the key right-hand-side variable, the future price, comes from variation across individuals in the plans they enrolled in, the month in which they enrolled, and the spending of the group of people who enrolled in that plan during that month.

Simulated future price There are two sets of potential concerns with the analysis in equation (15). One class of concerns is that, as previously discussed, individuals choose when to enroll in a plan. We would prefer to use variation in the future price that comes from birth month rather than enrollment month.

A second class of concerns is that we calculate the future price as a function of the spending of individuals in a given plan and enrollment month, and our dependent variable is a function of that spending for an individual in that plan and enrollment month. This second class of concerns in turn raises three issues. The first issue is the potential endogeneity of the future price (our key right-hand-side variable) to initial drug use (the dependent variable). The future price is calculated based on the drug spending of individuals who enroll in a given plan in a given month. If this spending responds to the future price, this will bias our estimate of the impact of the future price on drug spending away from zero.[40] A second issue is reflection bias. The future price is calculated based on the total spending of the set of people who enroll in a given plan in a given month and

[39] For example, a regression of (linear) enrollment month on (linear) birth month (controlling for plan fixed effects) has a coefficient of 0.858 (standard error = 0.002). In the regression results below we report the F-statistics for the excluded instruments in each IV specification.

[40] Note, however, that in our context this is not a concern. This potential endogeneity is not a problem if the only goal is to test the null of complete myopia (i.e., testing whether the coefficient on the future price is zero) because under the null of complete myopia drug spending is not a function of the future price. Our focus in this appendix is only on testing that null, not on quantifying the response to the future price.

the dependent variable is the initial spending of a given person who enrolled in the plan in that month. This problem is more acute the smaller is the sample size of people enrolling in a given plan in a given month (and thus the larger the contribution of the individual to the plan-month mean total spending). A final issue is the potential for common shocks. If there is a shock to health or spending that is specific to individuals enrolling in a specific plan in a specific month (e.g. the flu hits particularly virulently those who enroll in a particular plan in a different month), this introduces an omitted variable that is driving both the future price and initial drug use.

We address both classes of concerns with an instrumental variable strategy in which we instrument for the future price with a simulated future price. Like the future price, the simulated future price is computed based on the characteristics of the plan chosen. However, unlike the future price, it uses data on monthly spending for a *common sample* of individuals for all calculations, thus "purging" any variation in monthly spending that is correlated with plan or enrollment month, while at the same time addressing reflection bias and common shocks concerns. It addition, for the simulated future price we calculate contract duration (i.e. number of months of spending to draw) based on birth month, not join month; this is designed to address the concern that enrollment month may be endogenous.[41] Our simulated future price variable is very much in the spirit of Currie and Gruber's (1996) simulated Medicaid eligibility instrument.

Identifying assumption Our key identifying assumption is that conditional on any fixed spending differences by plan and any (flexible) spending pattern by enrollment month, the within year pattern of initial drug use by enrollment month does not vary based on which plan the individual enrolled in, except for the dynamic incentives. This strategy allows initial drug use levels to vary across people in different plans due to selection differences (not surprisingly, we see in Figure 6 higher initial drug use – i.e. shorter days to first claim – for individuals in no-deductible plans, as would be expected from plan selection). It also allows for seasonal patterns in initial drug use either because of demographic differences in the population by birth month or because of seasonal differences in drug use based on which three-month window is being used to define "initial utilization."

One reason the identifying assumption could be violated is if the same dynamic response that may lead to differential initial drug use among people in the same plan with different contract length also leads to differential selection into plans on the basis of enrollment month. A priori, it is not clear if individuals would engage in differential selection into a deductible vs. no-deductible plan based on the month they are enrolling in the plan. In practice, we find that the probability of enrolling in the no-deductible plan is increasing in enrollment month in a statistically significant but

[41] Specifically, for every individual in our sample regardless of plan and enrollment month, we compute their monthly spending for all months that we observe them during the year that they enroll in the plan, creating a common monthly spending pool. We then simulate the future price faced by an indivdiual who enrolls in a particular plan in his birth month by drawing (with replacement) 10,000 draws of monthly spending from this common pool, for every month we need a monthly spending measure. For the first month we draw from the pool of first month spending (since people may join the plan in the middle of the month, the first month's spending has a different distribution from other months) whereas for all other months in the plan that year we draw from the pool (across plans and months) of non first month spending. For each simulation we then compute the expected end-of-year price based on the draws.

economically trivial manner (one extra month is associated with a 0.4 percentage point increase in the probability of choosing a deductible plan, relative to a mean probability of choosing the no-deductible of about 75 percent).

Results

Graphical results: deductible vs. no-deductible plans Figure 6, which is presented in the main text, illustrates our main finding graphically. For the deductible plans, the simulated future price is increasing with enrollment month and initial drug use is decreasing with enrollment month (i.e. average time to first claim is increasing with enrollment month). For the no-deductible plan the simulated future price is decreasing with enrollment month, and initial drug use does not appear to vary systematically with enrollment month.

Regression results Although for purposes of graphical presentation we grouped plans into deductible and no-deductible plans, in the regression analysis we exploit the finer variation across plans in the patterns of future price and initial drug use by enrollment month. For example, some no-deductible plans have gap coverage and others do not, which creates different patterns of future price by enrollment month. Table B3 reports the analysis of patterns of initial drug use by enrollment month. Throughout, the dependent variable is log days to first claim. Columns (1) and (2) report the results from estimating equation (13) for deductible plans and no-deductible plans, respectively. Column (1) shows that initial drug use declines (i.e. time to first claim increases) in enrollment month for the deductible plans. The effect is statistically significant, and the point estimate indicates that a one-month increase in enrollment month is associated with a 2.3 percent increase in days to first claim (corresponding to a decrease in initial drug use). Column (2) shows no economically or statistically significant pattern in the relationship between initial drug use and enrollment month in the no-deductible plans. As a result, the difference-in-differences analysis in column (3) shows an effect of enrollment month for deductible plans that is virtually identical to the deductible plan analysis in column (1). When we instrument for enrollment month with birth month in column (4), the effect remains statistically significant although the magnitude attenuates. The point estimates indicates that a one month increase in enrollment month is associated with a 1.4 percent increase in time to first claim (i.e. decrease in initial drug use) for individuals in the deductible plan, relative to the no-deductible plan.

The rest of Table B3 reports the relationship between initial drug use and future price. Column (5) shows the OLS estimates of equation (15). It indicates that a 10 cent increase in the future price is associated with a 4.6 percent increase in time to first claim (i.e. decrease in initial drug use). The IV analysis in column (6), which uses the simulated future price and birth month fixed effects as instruments for the future price and enrollment month fixed effects, indicates that a 10 cent increase in the future price is associated with a 6.7 percent increase in time to first claim (i.e. a decrease in initial drug use). Both estimates are statistically significant. We reject the null of no response of initial drug use to the future price, and thus reject the null of completely myopic behavior. These results are robust to estimating the regressions in levels instead of logs (not shown).

Appendix C: Estimation details

Simulation We estimate our model using simulated minimum distance. As described in Section 5,

$$\widehat{\varphi} \in \arg\min_{\varphi \in \Psi} (m_n - m_s(\varphi))' W_n (m_n - m_s(\varphi)). \tag{16}$$

To calculate $m_s(\varphi)$ we simulate data given a vector of parameters. To do so, we first calculate the value function for each latent type and plan combination as described below. For each observation we then simulate S claim histories. Given a person's chosen plan, age, and other characteristics we simulate a sequence of claims. We first draw the person's type m_{is} from a multinomial distribution with probabilities $e^{z_i\beta_m} / \left(\sum_{l=1}^{M} e^{z_i\beta_l}\right)$. Then, starting from the first week of the year ($t = 51$) and going until the final week of the year ($t = 0$), we simulate a claim history.[42] Cumulative spending begins with $x_{is0} = 0$. Each week there is an event with probability $\lambda_{m_{is}}$. When there is an event, the log potential claim is $\log \theta_{ist} \sim N(\mu_{m_{is}}, \sigma^2_{m_{is}})$. The utility cost of not filling the claim is ω_{ist}, which is equal to θ_{ist} with probability $1 - p_{m_{is}}$ and uniform on $(0, \theta_{ist})$ with probability $p_{m_{is}}$. The claim is filled if

$$-c_j(\theta_{ist}, x_{ist}) + \delta v_{jm}(x_{ist} + \theta_{ist}, t-1) \geq -\omega_{ist} + \delta v_{jm}(x_{ist}, t-1). \tag{17}$$

In this case, $x_{ist-1} = x_{ist} + \theta_{ist}$. Otherwise, $x_{ist-1} = x_{ist}$. We repeat this simulation until $t = 0$. We then use the simulated data to calculate the statistics described in Section 5. Since the number of observations is large, we use one simulation per observation ($S = 1$).

Minimization Throughout the minimization of the objective function, the underlying random draws are kept constant and only shifted and/or rescaled as the parameters change. Nonetheless, the simulated objective is not continuous with respect to φ due to discrete changes in whether some simulated potential claims are filled or not. The large number of potential sequences of claims makes smoothing the objective function difficult. Instead, we use a minimization algorithm that is robust to poorly behaved objectives, the covariance matrix adaptation evolution strategy (CMA-ES) of Hansen. Like simulated annealing and various genetic algorithms, CMA-ES incorporates randomization, which makes it effective for global minimization. Like quasi-Newton methods, CMA-ES also builds a second order approximation to the objective function, which makes CMA-ES much more efficient than purely random or pattern based minimization algorithms. In comparisons of optimization algorithms, CMA-ES is among the most effective existing algorithms, especially for non-convex non-smooth objective functions (Hansen et al. 2010; Rios and Sahinidis 2012). Andreason (2010) shows that CMA-ES performs well for maximum likelihood estimation of DSGE models. As discussed by Hansen and Kern (2004), an important parameter for the global convergence of CMA-ES is the population size. We initially set the population size to the default value of 15 (which is proportional to the logarithm of the dimension of the parameters), and then increased it

[42] For 65 year old we start from the week they enrolled in Medicare Part D. Since our data only contains the month, but not week, of enrollment, we draw the enrollment week from a uniform distribution within the enrollment month.

to 100. The computation is primarily CPU bound. The estimation takes roughly four days to run on a server with two Intel Xeon E5-2670 eight-core processors.

Calculation of value function

Each individual's value function depends on her chosen plan, j,and her unobserved type, m. As in equation (2), the Bellman equation is

$$v_{jm}(x,t) = (1-\lambda_m)\delta v_{jm}(x,t-1) + \lambda_m E_m \left[\max \left\{ \begin{array}{c} -c_j(\theta,x) + \delta v_{jm}(x+\theta,t-1), \\ -\omega + \delta v_{jm}(x,t-1) \end{array} \right\} \right], \tag{18}$$

where the subscripts denote plan j and type m. The expectation is subscripted by m to emphasize that it depends on the type-specific distribution of θ and ω. Given that $v_{jm}(x,0) = 0$, we can compute an approximation to v_{jm} sequentially. First, we approximate $v_{jm}(x,1)$. Then, we use that approximation to compute $v_{jm}(x,2)$, and so on. To be more specific, let $\{x_{k,j}\}_{k=1}^{K}$ be a large set of values of x that cover the support of x. Then, given some approximation to $v_{jm}(x,t-1)$, say $\tilde{v}_{jm}(x,t-1)$, we compute

$$v_{k,jm} = (1-\lambda_m)\delta \tilde{v}_{jm}(x_{k,j},t-1) + \lambda_m E_m \left[\max \left\{ \begin{array}{c} -c_j(\theta,x_{k,j}) + \delta \tilde{v}_{jm}(x_{k,j}+\theta,t-1), \\ -\omega + \delta \tilde{v}_{jm}(x_{k,j},t-1) \end{array} \right\} \right]. \tag{19}$$

We then calculate $\tilde{v}_{jm}(x,t)$ using linear interpolation between the $\{(x_{k,j}, v_{k,jm})\}$ values.[43] We allow $x_{k,j}$ to differ for each plan. For each plan, $x_{k,j}$ is set to 20 evenly spaced points between 0 and the deductible amount, 20 evenly spaced points between the deductible amount and the kink location, 20 evenly spaced points in the gap, and only 2 points above the catastrophic limit. Thus, plans with a deductible use $K = 62$ interpolation points and plans without a deductible use $K = 42$ interpolation points. Above the catastrophic limit, $c(\theta,x) = C\theta$ for some constant C, so the value function is constant and two interpolation points suffice.

To calculate $v_{k,jm}$, we must integrate over θ and ω to compute

$$E_m \left[\max \left\{ -c_j(\theta,x_k) + \delta \tilde{v}(x_k+\theta,t-1) \ , \ -\omega + \delta \tilde{v}_{jm}(x_k,t-1) \right\} \right]. \tag{20}$$

We approximate the expectation over θ using Gauss-Hermite quadrature with 30 integration points. Given the assume distribution of ω/θ, the remaining conditional expectation over ω given θ has a closed form. In particular,

$$\begin{aligned} &E_m \left[\max \left\{ \begin{array}{c} -c_j(\theta,x_{k,j}) + \delta \tilde{v}_{jm}(x_{k,j}+\theta,t-1), \\ -\omega + \delta \tilde{v}_{jm}(x_{k,j},t-1) \end{array} \right\} \right] = \\ &= E_m \left[\begin{array}{c} P\left(\frac{c_j(\theta,x_{k,j}) - \delta \tilde{v}_{jm}(x_{k,j}+\theta,t-1) + \delta \tilde{v}_{jm}(x_{k,j},t-1)}{\theta} \leq \frac{\omega}{\theta} \middle| \theta \right) \left(-c_j(\theta,x_{k,j}) + \delta \tilde{v}_{jm}(x_{k,j}+\theta,t-1) \right) + \\ + \left(\begin{array}{c} P\left(\frac{c_j(\theta,x_{k,j}) - \delta \tilde{v}_{jm}(x_{k,j}+\theta,t-1) + \delta \tilde{v}_{jm}(x_{k,j},t-1)}{\theta} > \frac{\omega}{\theta} \middle| \theta \right) \cdot \\ \cdot \left(E\left[-\omega \mid \frac{c_j(\theta,x_{k,j}) - \delta \tilde{v}_{jm}(x_{k,j}+\theta,t-1) + \delta \tilde{v}_{jm}(x_{k,j},t-1)}{\theta} > \frac{\omega}{\theta} \right] - \delta \tilde{v}_{jm}(x_{k,j},t-1) \right) \end{array} \right) \end{array} \right], \end{aligned} \tag{21}$$

[43] We also experimented with shape preserving cubic interpolation. The resulting value function approximation is very similar. We use linear interpolation in the estimation because it is less computationally intensive.

where

$$P\left(C \leq \frac{\omega}{\theta}\middle|\, \theta\right) = \begin{cases} 0 & \text{if } C <= 0 \\ p_m C & \text{if } C \in (0,1) \\ 1 & \text{if } C \geq 1 \end{cases} \tag{22}$$

and

$$E\left[\omega|\frac{\omega}{\theta} < C\right] = \begin{cases} \frac{Cp_m}{2} & \text{if } C \in [0,1) \\ 1 - p_m + \frac{p_m}{2} & \text{if } C \geq 1 \end{cases}. \tag{23}$$

Code

The estimation code is written in C++. It is available at *https://bitbucket.org/paulschrimpf/medicared/overview*. It uses the covariance matrix adaptation evolution strategy (CMA-ES) of Hansen and Kern (2004) and Hansen (2006) to minimize the objective function. ALGLIB (*www.alglib.net*) is used for random number generation, interpolation, and integration.

Appendix D: Using Saez (2010) to map the excess mass to a price elasticity

In this appendix we explain how we can adapt Saez's (2010) model of the response of the income distribution to a progressive income tax schedule to our context in order to translate our bunching estimate into a price elasticity. Unlike the full dynamic model we develop and estimate in the paper, the model in this appendix (as in Saez 2010) is static and assumes complete information (i.e. no uncertainty about health shocks at the time of decision making).

Individual i obtains utility

$$u_i(m, y) = g_i(m) + y \tag{24}$$

from (total) drug spending m and residual income y, as in Einav et al. (2013). As in Einav et al. (2013) and Saez (2010), we assume that utility is quasi-linear. We make further parametric assumptions, so that

$$u_i(m, y) = \underbrace{\left[2m - \frac{\zeta_i}{1 + \frac{1}{\alpha}}\left(\frac{m}{\zeta_i}\right)^{1+\frac{1}{\alpha}}\right]}_{g_i(m)} + \underbrace{[I_i - c_j(m)]}_{y}. \tag{25}$$

That is, residual income y is given by the individual's income I_i minus his (annual) out-of-pocket cost $c_j(m)$, where $c_j(\cdot)$ defines the function that, given the individual's insurance coverage j, maps total spending m to the fraction of it that is paid out of pocket as illustrated, for example, in Figure 1.

The choice of $g_i(m)$ in equation (25) is less standard, and is motivated by our attempt to obtain a tractable, constant elasticity form of the spending function that would be similar to Saez (2010) despite the different context. As will be clear soon, in our specification of $g_i(m)$ above one can

think of ζ_i as representing an individual's health needs, which vary across individuals, and α as a parameter, common across individuals, that affects individuals' elasticity of drug spending with respect to the out-of-pocket price.

To see the motivation for this particular parameterization, consider its implication in the context of a linear coverage. Suppose coverage is linear and is given by $c_j(m) = c_j \cdot m$ with $c_j \in [0,1]$, so that $c_j = 0$ represents full coverage and $c_j = 1$ represents no coverage. In such a case, an individual solves

$$\max_m \left[2m - \frac{\zeta_i}{1+\frac{1}{\alpha}} \left(\frac{m}{\zeta_i} \right)^{1+\frac{1}{\alpha}} + I_i - c_j \cdot m \right], \tag{26}$$

and the optimal choice of healthcare utilization is given by

$$m = \zeta_i (2 - c_j)^\alpha. \tag{27}$$

That is, with no insurance ($c_j = 1$) the individual spends $m = \zeta_i$, while with full insurance he spends $m = 2^\alpha \zeta_i$. Thus our specification implies a constant elasticity α of spending with respect to $(2 - c_j)$.

This constant elasticity form of the spending function is now very similar to Saez's choice of labor supply function, and for the rest of this appendix we can follow closely his strategy. We assume that ζ_i is distributed in the population with cdf $F(\zeta)$ and pdf $f(\zeta)$, analogously to individual's ability n in Saez. m is analogous to z in Saez, and $(2 - c_j)^\alpha$ is analogous to $(1 - t)^e$ in Saez (where t is the marginal tax rate on income). Applying these analogies, we can start with, say, equation (2) in Saez (page 186), which is identical (after applying the analogies) to equation (27) above.

Consider now $H_0(m)$ to be the cdf of spending when the marginal price (before the gap) is c_{j0}. Denote by $h_0(m) = H_0'(m)$ the corresponding pdf. Because $m = \zeta_i(2 - c_{j0})^\alpha$ we have $H_0(m) = \Pr(\zeta_i(2 - c_{j0})^\alpha \leq m) = F(m/(2 - c_{j0})^\alpha)$. So $h_0(m) = f(m/(2 - c_{j0})^\alpha)/(2 - c_{j0})^\alpha$. Consider now the gap, where there is a kink and the marginal price $c_{j1} >> c_{01}$ becomes much higher, so above the kink we have $m = \zeta_i(2 - c_{j1})^\alpha$. H is then the distribution of spending under the kink scenario. If the kink is at m^*, then distribution of spending up to m^* is given by $H_0(m)$. That is, spending is such that:

$$m(\zeta_i) = \begin{cases} \zeta_i(2 - c_{j0})^\alpha & if\ \zeta_i < m^*/(2 - c_{j0})^\alpha \\ m^* & if\ \zeta_i \in [m^*/(2 - c_{j0})^\alpha, m^*/(2 - c_{j1})^\alpha] \\ \zeta_i(2 - c_{j1})^\alpha & if\ \zeta_i > m^*/(2 - c_{j1})^\alpha \end{cases}. \tag{28}$$

Thus, for spending above the kink ($m > m^*$) we have $H(m) = F(m/(2 - c_{j1})^\alpha)$.

The rest continues as in Saez and the above analogies. For example, Saez's equation (3) becomes:

$$\frac{\Delta m^*}{m^*} = (2 - c_0)^\alpha - 1 \tag{29}$$

and his equation (5) becomes

$$B = m^* \left[\left(\frac{2 - c_{j0}}{2 - c_{j1}} \right)^\alpha - 1 \right] \frac{h(m^*)_- + h(m^*)_+ / \left(\frac{2-c_{j0}}{2-c_{j1}} \right)^\alpha}{2}. \tag{30}$$

Equation (30) can then be used to express α as a function of estimable objects, allowing us to convert our bunching estimate of B to an elasticity estimate α. We should note that, unlike in Saez (2010), our specification implies a constant elasticity α of spending with respect to $(2 - c_j)$ rather than $(1 - t)$. The elasticity of spending with respect to c_j is therefore not constant, and would depend on the level of c_j. In our calculations below we compute a plan-specific α_j, and then report the elasticity estimate evaluated at the pre-kink cost sharing rate c_{j0}, and then taking (weighted) average across all plans.

Table D1 shows our estimate of the elasticity α under alternative specifications for estimating B. We estimate B by estimating the counterfactual distribution of spending that would exist around the kink in the absence of the kink; B is the number of people who are empirically in the area around the kink over and above the number of people whom we (counterfactually) estimate would be in this area if the kink did not exist (B is thus the numerator for the "excess mass" estimate we report in Section 4.1).

Because calculation of α requires estimates of B separately by plans (so that B can be translated into an elasticity separately for plans with different cost-sharing features), we limit this analysis (as in Figure 5) to the approximately 80% of our baseline sample that are in plans with at least 1,000 beneficiaries with spending within $2,000 of the kink. We report the beneficiary-weighted average elasticity across plans.

The first row of Table D1 shows our baseline specification, which approximates the counterfactual distribution of spending that would exist near the kink if there was no kink by fitting a cubic approximation to the CDF, using only individuals whose spending is below the kink (between $2,000 and $200 from the kink), and subject to an integration constraint (see Figure 4 for more details). We then use a $200 window around the kink to produce our bunching estimate B, which we then translate to an elasticity α. The next two rows show the sensitivity of our elasticity estimate to fitting the cubic approximation using individuals whose spending is below the kink between $2,000 and $100 from the kink (and using a $100 window around the kink to produce our bunching estimate) or to fitting the cubic approximation using individuals whose spending is below the kink between $2,000 and $300 from the kink (and using a $300 window around the kink for the bunching estimate). In the bottom two rows we return to the $200 exclusion range, but, unlike the baseline specification, we approximate the counterfactual distribution with a locally uniform distribution (see table notes for details), or with a quadratic approximation fit using individuals whose spending is below the kink between $2,000 and $200 from the kink. These exercises produce alternative estimates of the elasticity of spending with respect to the pre-kink cost-sharing rate, which range from -0.015 to -0.029.

Additional Appendix References

Andreasen, Martin Møller. (2010). "How to Maximize the Likelihood Function for a DSGE Model." *Computational Economics*. 35(2), 127-154.

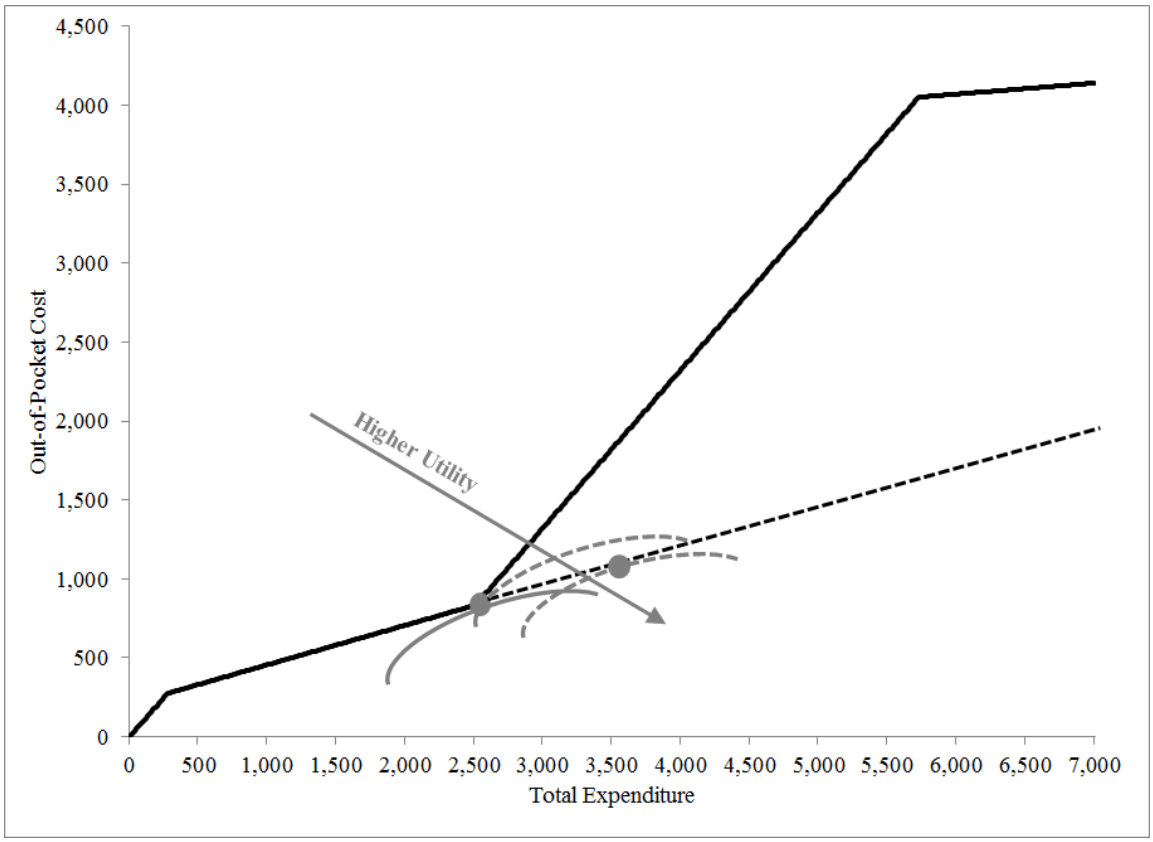
Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiri-*

cist's Companion. Princeton, New Jersey: Princeton University Press.

Lo Sasso, Anthony, Lorens Helmchen and Robert Kaestner. 2010. "The Effect of Consumer Directed Health Plans on Health Care Spending." *Journal of Risk and Insurance* 77(1): 85-103.

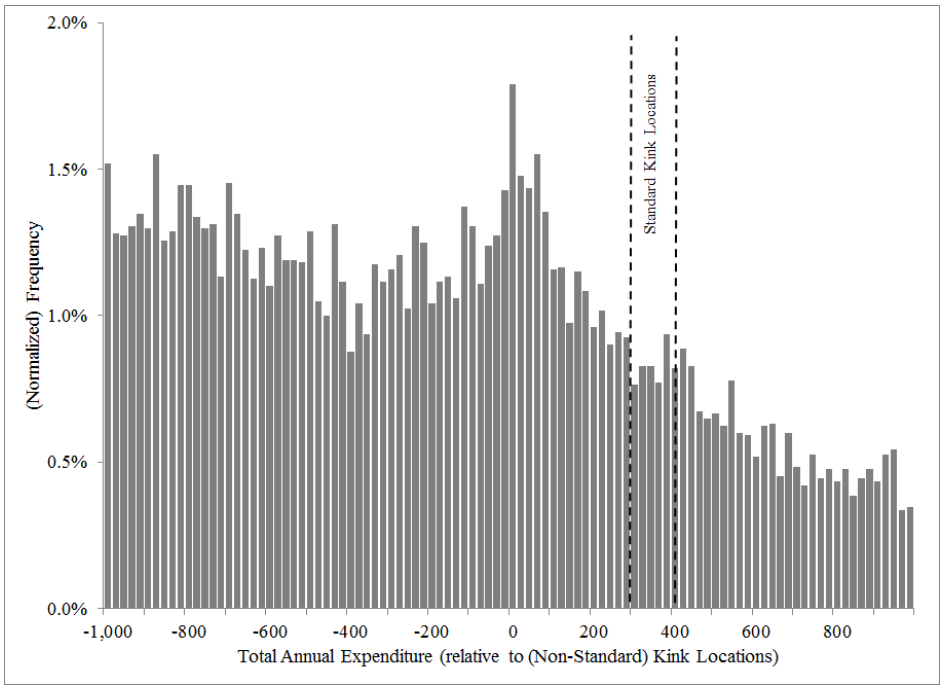
Van den Berg, Gerard J. (2001). "Duration Models: Specification, Identification and Multiple Durations." In J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics* (First Ed.), Vol. 5, Amsterdam: Elsevier, Chapter 55, 3381–3460.

Figure A1: Rationale for bunching



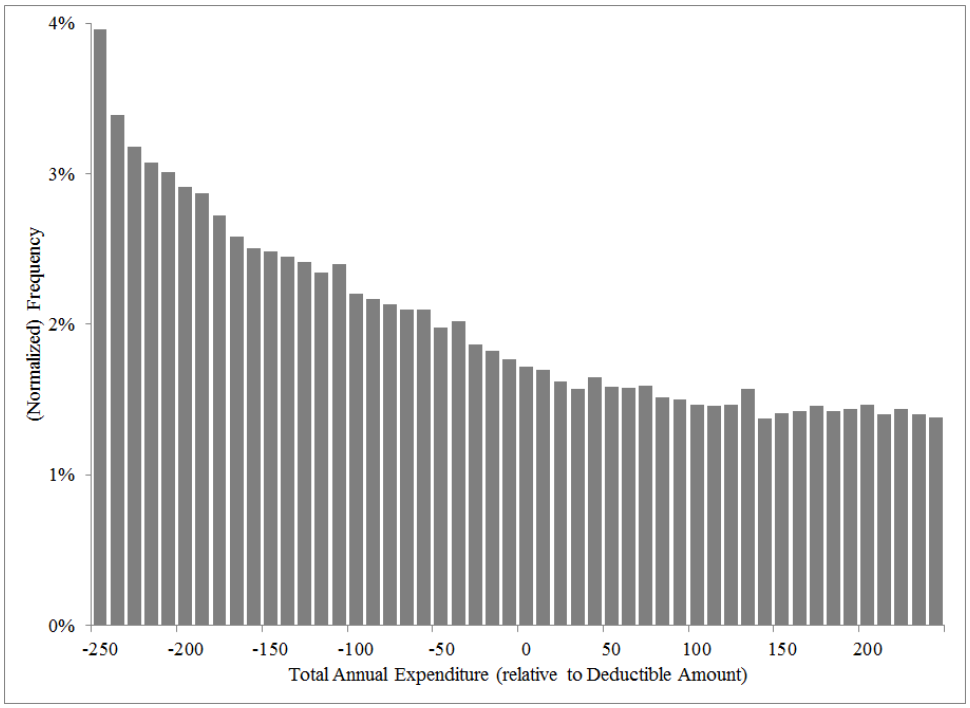
The solid line illustrates the budget set of the same standard benefit design as in Figure 1. The dashed line considers an alternative budget set with a linear budget (above the deductible) at the co-insurance arm's cost sharing rate. By contrast, the standard budget set has a kink (price increase) at $2,510 in total spending. The individual denoted by the solid indifference curve is not affected by the introduction of this kink; his indifference curve remains tangent to the lower part of the budget set. The individual with the dashed indifference curves consumed above the kink under the linear budget set; with the introduction of the kink her indifference curve is now exactly tangent to the upper part of the budget set at the kink. With the introduction of the kink, this latter individual would therefore decrease total spending to the level of the kink location. By extension, any individual whose indifference curve was tangent to the linear budget set at a spending level between that of the two individuals shown would likewise decrease total spending to the level of the kink location, thereby creating "bunching" at the kink.

Figure A2: Distribution of spending for those individuals with non-standard kink location



The figure displays the histogram of total annual prescription drug spending (in $20 bins) for individuals with the modal ($2,100) non-standard kink location in 2007 or 2008. Such individuals are not in our baseline sample, which is limited to those with the standard kink location. The x-axis reports total spending relative to the $2,100 kink location. The dashed vertical lines indicate the level of the standard kink locations in 2007 ($2,400) and 2008 ($2,510). Frequencies are normalized to sum to 1 across the displayed range. N =12,188.

Figure A3: Distribution of spending around the deductible amount



The figure displays the histogram of total annual prescription drug spending (in $10 bins) for individuals in our baseline sample in plans with the (year-specific) standard deductible amount (which was $265 in 2007, $275 in 2008 and $295 in 2009). The x-axis reports total spending relative to the (year-specific) deductible amount. Frequencies are normalized to sum to 1 across the displayed range. N =186,535.

Table B1: Relationship between birth month and enrollment month

Birth Month	Join Month									Total N
	2	3	4	5	6	7	8	9	10	
1	37.7	17.9	19.2	9.6	3.5	4.3	2.6	2.6	2.8	4,947
2	68.5	11.9	7.0	5.1	3.4	1.4	1.0	1.0	0.8	14,861
3	1.8	67.1	13.7	6.0	4.8	3.7	1.1	1.0	0.8	15,878
4	0.0	1.9	69.8	11.9	6.1	5.0	3.2	1.2	1.0	14,640
5	0.0	0.0	1.8	70.2	12.3	6.7	4.6	3.4	1.0	14,674
6	0.0	0.0	0.0	2.1	70.4	12.4	6.5	5.0	3.6	14,754
7	0.0	0.0	0.0	0.0	1.9	72.3	13.2	7.2	5.5	16,247
8	0.0	0.0	0.0	0.0	0.0	2.2	76.6	14.4	6.8	15,359
9	0.0	0.0	0.0	0.0	0.0	0.0	2.2	83.5	14.3	14,058
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	97.3	11,823
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	297
Total (%)	8.9	9.9	10.7	10.6	10.8	12.1	12.2	12.6	12.3	137,538

Table shows the relationship between birth month and enrollment month for our 65-year old sub-sample (N=137,536). Specifically, it indicates the percent of each birth month who enrolled in each month. The last column shows the sample size for each birth month.

Table B2: Relationship between enrollment month and final cost sharing phase

Enrollment Month	Deductible	Pre-kink	Gap	Catastrophic	N
			Deductible Plans		
2	34.0	50.0	14.1	1.9	2,840
3	35.3	51.5	11.7	1.5	3,035
4	39.3	48.7	10.5	1.5	3,235
5	42.2	48.6	8.3	1.0	3,147
6	45.8	46.7	6.8	0.7	3,185
7	49.2	45.9	4.4	0.6	3,518
8	54.7	42.6	2.6	0.2	3,314
9	60.9	37.6	1.4	0.2	3,352
10	70.3	28.6	1.1	0.0	3,334
Total	48.4	44.2	6.5	0.8	28,960
			No-Deductible Plans		
2	0.0	81.2	17.1	1.8	9,496
3	0.0	83.8	14.9	1.4	10,543
4	0.0	86.7	12.2	1.2	11,411
5	0.0	88.9	10.4	0.7	11,387
6	0.0	91.9	7.6	0.5	11,646
7	0.0	94.5	5.0	0.6	13,106
8	0.0	96.0	3.8	0.2	13,449
9	0.0	97.9	1.9	0.2	13,931
10	0.0	98.9	1.0	0.1	13,609
Total	0.0	91.8	7.6	0.7	108,578

Table shows the relationship between enrollment month and the final (end of year) cost sharing phase the employee ends up in. Specifically, it shows the percent of beneficiaries, for each enrollment month, who end up in each cost-sharing arm. We show results separately for deductible and no-deductible plans for our 65-year-old sub sample (N=137,536).

Table B3: Relationship between initial drug use and enrollment month and future price

	Dependent Variable: Log (days to first claim)					
Sample	Deductible plans	No-deduct. plans	All	All	All	All
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS (DD)	IV (DD)	OLS	IV
Enrollment month	0.023***	0.002				
	(0.003)	(0.002)				
Deductible*Enrollment month			0.021***	0.014***		
			(0.003)	(0.004)		
Future price					0.463***	0.671***
					(0.055)	(0.127)
N	28,626	108,234	136,860	136,860	136,860	136,860

Columns (1)-(4) show the relationship between initial drug use and enrollment month. Throughout our measure of initial drug use (the dependent variable) is log days to first claim. Column (1) shows the coefficient on join month from estimating the relationship between log days to first claim and enrollment month, controlling for plan fixed effects (equation (13)) for deductible plans (for which the future price on average increases with enrollment month). Column (2) shows the coefficient on enrollment month from estimating the same equation (13) for no-deductible plans (for which the future price on average decreases with enrollment month). Column (3) shows the coefficient on the interaction of enrollment month and a deductible dummy from estimating the "difference in difference" equation (14), which controls for plan fixed effects and enrollment month fixed effects, on the combined sample of individuals in all plans. Column (4) shows the instrumental variables estimation of the difference in difference equation (14) shown in column (3), where we instrument for enrollment month fixed effects and the enrollment month variable interacted with a deductible dummy using birth month fixed effects and a birth month variable interacted with a deductible dummy; Angrist-Pischke (2009) F-statistics for the first stage models are all above 2,000. Standard errors are clustered at the plan level in all specifications. Columns (5) and (6) show the relationship between initial drug use and future price. Column (5) shows the coefficient on future price from estimating the relationship between initial drug use and future price, controlling for plan fixed effects and enrollment month fixed effects (equation (15)). Column (6) shows the instrumental variables estimation of equation (15), where we instrument for the future price and enrollment month fixed effects with the simulated future price and birth month fixed effects; Angrist-Pischke (2009) F-statistics for the first stage models are all above 200. Standard errors are clustered at the plan level in all specifications.

Table D1: Excess mass and Saez (2010) elasticity calculations

Counterfacutal distribution	"Exclusion" window[a]	Elasticity[b]
Cubic	200	-0.024
Cubic	100	-0.015
Cubic	300	-0.029
Uniform	n/a[c]	-0.020
Quadratic	200	-0.022

Table reports estimates of the implied elasticities under alternative assumptions. We limit the analysis to the approximately 80% of our baseline sample who in plans with at least 1,000 beneficiaries whose end the year is within $2,000 of the kink. We analyze the behavior of beneficiaries who are themselves within $2,000 of the kink. For each plan we use equation (30) plus the plan's cost sharing rules to translate it into an estimate of the spending elasticity with respect to the coinsurance rate c, evaluated at the plan's pre-kink cost sharing rate c. The right-most column reports the average estimates across the plans, weighted by their enrollment. The different rows report results from different approaches to calculating the counterfactual distribution of spending that would exist in the absence of the kink. The first row shows the baseline approach (used in Figure 4 and Table 9), in which the counterfactual distribution was calculated by fitting a CDF difference function, constrained to integrate to the sample size included in the graph, using only the points to the left of -$200 (see notes to Figure 4 for details) and using the "exclusion window" of $200 around the kink to estimate the response to the kink. In the next two rows we repeat the estimation but instead use omit observations where spending is greater than -$100 and -$300 (relative to the kink) in fitting the counterfactual distribution and use the "exclusion window" of $100 or $300 around the kink respectively to estimate the response to the kink. In the final rows, we return to the -$200 " exclusion window" but approximate the counterfactual distribution using a locally uniform approximation (following Saez 2010) and a quadratic CDF difference approximation. N=1,985,676.

[a] Exclusion window refers to the distance from the kink location within which we calculate the response to the kink. The counterfactual density is fit using points only to the left of the exclusion window.

[b] Elasticity of spending is calculated with respect to the cost-sharing rate c; and is evaluated at each plan's pre-kink cost sharing rate.

[c] For the locally uniform counterfactual distribution we use the observed distribution of spending that is -$400 to -$200 from the kink to approximate the counterfactual spending distribution within -$200 to $0 of the kink; likewise we use the observed distribution that is $200 to $400 from the kink to approximate the counterfactual distribution within $0 to $200 of the kink.