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

Liran Einav, Amy Finkelstein, and Paul Schrimpf[†]

November 2014

Abstract. We study the demand response to non-linear price schedules using data on insurance contracts and prescription drug purchases in Medicare Part D. We exploit the kink in individuals' budget set created by the famous "donut hole," where insurance becomes discontinuously much less generous on the margin, to provide descriptive evidence of the drug purchase response to a price increase. We then specify and estimate a simple dynamic model of drug use that allows us to quantify the spending response along the entire non-linear budget set. Counterfactual analysis of the model suggests that "filling" the donut hole, as will be required by 2020 under the Affordable Care Act, will increase annual drug spending by about $150 per beneficiary, or about 8 percent. About one-quarter of this spending increase is "anticipatory," coming from beneficiaries whose spending prior to the policy change would leave them short of reaching the donut hole. About two thirds of the spending increase could represent a decrease in cross-year substitution rather than a net increase in spending measured over a longer time horizon.

JEL classification numbers: D12, G22.

Keywords: Medicare, Moral hazard, Contract Design, Health insurance, Health care.

[*] We thank Larry Katz, Elhanan Helpman, five anonymous referees, Jason Abaluck, Raj Chetty, Ben Handel, Day Manoli, John Friedman, Nathan Hendren, Maria Polyakova, and many seminar participants for useful comments and helpful discussions, and Allyson Barnett, Ray Kluender, and Belinda Tang for 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.

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. In 2013, about 37 million people received Part D coverage (Kaiser Family Foundation 2014). We analyze the response of drug expenditures to insurance contract design using detailed micro data on insurance contracts and prescription drug purchases from a 20% random sample of Medicare Part D beneficiaries from 2007 to 2009. Section 2 describes the data and institutional setting in more detail.

Figure 1 illustrates the highly non-linear nature of the Part D contracts; it shows the 2008 government-defined standard benefit design. 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 sample a beneficiary entering the coverage gap at the "donut hole" experiences, on average, a price increase of almost 60 cents for every dollar of total spending.

Motivated by these contract features, we begin in Section 3 by exploiting the kink in the individual's budget set created by the donut hole to provide descriptive evidence on the nature of the drug purchase response to the drug price increase at the kink. We document significant "excess mass," or "bunching" of annual spending levels around the kink. This is visually apparent in even the basic distribution of annual drug spending in any given year, as shown in Figure 2 for 2008. The behavioral response appears to grow over time, which may reflect a "learning" effect (by individuals or pharmacists) about the presence of the gap in the new program; it also tends to be larger for healthier individuals. Using the detailed data on the timing of claims, we also show a sharp decline in the propensity to claim toward the end of the year for those individuals whose spending is near the kink. This decline is concentrated later in the year but is also visible at earlier months in the year; this is consistent with individuals updating over the course of the year about their expected end-of-year price, and having a positive discount factor. The decline in drug purchases for individuals near the kink is somewhat more pronounced for chronic than acute drugs and substantially more pronounced for branded than generic drugs.

The descriptive results provide qualitative evidence of the extent and nature of the drug purchase response to the insurance contract. To quantify this response and use it for counterfactual analysis of behavior under other contracts requires us to develop and estimate a model of individual behavior. Section 4 presents therefore 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 and statistical objects that determine the expenditure response to the contract. The first key object is the distribution of health-related events, which determine the set of potential prescription drug expenditures, and which we allow to vary across observable and unobservable patient characteristics, including patient health. The second is the "primitive" price elasticity that captures the individual's willingness to trade off health and income, which we also allow to be heterogeneous across observable and unobservable patient characteristics. The third object is the extent to which individuals respond to the dynamic incentives associated with the non-linear contract. We parameterize the model and estimate it using method of moments, relying on the descriptive patterns in the data described earlier, as well as additional features of the data and modeling assumptions. The estimated model fits the data reasonably well.

Section 5 presents the main results. We illustrate the demand response to non-linear contract design through analysis of a specific, policy-relevant counterfactual contract: the requirement in the 2010 Affordable Care Act (ACA) that, effective in 2020, the standard (i.e. minimal) benefit plan eliminate 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, annual drug spending by \$150 per beneficiary (or about 8%), and will increase Medicare drug spending by substantially more (by \$260 per beneficiary, or about 25%). 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 60% of our estimated effect.

The results also illustrate some of the subtle, distributional effects that non-linear contracts can produce. For example, we find that somewhat counter-intuitively, filling the gap provides less coverage on the margin to some individuals, causing them to *decrease* their spending. We also show that filling the gap changes spending behavior for individuals far from the gap; we estimate that about one-quarter of the \$150 per-beneficiary increase in annual 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 final section of the paper, we consider yet another implication of non-linear contracts: potential incentives for cross-year substitution. We provide descriptive evidence that some, but not all of the decline in purchasing for individuals who reach the kink reflects the postponement of purchases to the beginning of the subsequent year. To quantitatively assess the importance of this cross-year substitution for our main results, we provide a simple extension to the baseline model that allows – in a highly stylized way – for individuals to engage in cross-year substitution. The results suggest that up to two thirds of the annual spending increase from filling the gap may be explained by a decline in substitution of purchases to the subsequent year, rather than a net increase in spending over a longer-than-annual horizon.

Our findings have several implications for future empirical work on "moral hazard" (that is, spending) effects of health insurance contracts. Most of this literature has focused on characterizing the spending effect of a health insurance contract with respect to a given single price 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; see Aron-Dine et al. 2013). Our paper is part of a recent flurry of attention to the non-linear nature of typical health insurance contracts (Vera-Hernandez 2003; Bajari et al. 2011; Kowalski 2012; Dalton 2014). It complements our earlier work (Aron-Dine et al. 2014) in which we tested – and rejected – the null hypothesis that individuals do not consider the dynamic incentives in non-linear health insurance contracts when making drug or medical purchase decisions. Our results suggest that the distributional consequences of a change in a non-linear health insurance contracts may be quite subtle, affecting people whose spending is outside the range of where the actual contract change occurs. They also suggest the importance of cross-year substitution, and therefore examining spending effects over horizons longer than one year. With the exception of Cabral's (2013) recent work, it has been the standard approach in the literature to focus on the analysis of each annual contract in isolation. Our findings raise interesting questions regarding the optimal coverage horizon in the presence of non-linear contracts, which almost always cover a single year.[1]

Our paper also relates to the growing literature on the new Part D program. Most of this literature focuses on consumers' choice of plans (e.g. Heiss et al. 2010, 2013; Abaluck and Gruber 2011; Kling et al. 2012; Ketcham et al. 2012; Polyakova 2014). However, some of it also examines the impact of Part D on drug purchases (e.g. Yin et al. 2008; Duggan and Scott Morton 2010; Joyce et al. 2013), including the role of non-linear contracts (Abaluck et al. 2014, Gowrisankaran et al., 2014).

Finally, our analysis of "bunching" at the kink is related to recent studies analyzing bunching of annual earnings in response to the non-linear budget set created by progressive income taxation (Saez 2010; Chetty et al. 2011, 2013). This literature has emphasized that "excess mass" estimates cannot directly translate into an underlying behavioral elasticity since they also are affected by frictions – such as supply-side constraints on the choice of number of hours to work or limited awareness of the budget set in the labor supply context (Chetty et al. 2011; Chetty 2012; Kleven and Waseem 2013). These frictions are likely to be substantially less important in our setting, in which 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, unlike the analysis of bunching in the static labor supply framework developed by Saez (2010), we must account for the fact that, in our

[1] Indeed, Cabral (2013) documents similar cross-year substitution in the context of employer-provided dental insurance, and exploits variation in coverage duration in her data to explore precisely this question.

[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

context, decisions are made sequentially throughout the year and information is obtained gradually as health shocks arrive and individuals move along their non-linear budget set. 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 Setting and Data

Medicare provides medical insurance to the elderly and disabled. Parts A and B provide in-patient hospital and physician coverage respectively; Part D, which was introduced in 2006, provides prescription drug coverage. We have data on a 20% random sample of all Medicare Part D beneficiaries from 2007 through 2009. We observe the cost-sharing characteristics of each beneficiary's plan, as well as detailed, claim-level information on any prescription drugs purchased. We also observe basic demographic information (including age, gender, and eligibility for various programs tailored to low income individuals). In addition, we observe each beneficiary's Part A and B claims; we feed these 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."[3]

Part D enrollees choose among different prescription drug plans offered by private insurers. The different plans have different cost-sharing features and premiums. All plans provide annual coverage for 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.[4]

Sample definitions and characteristics We make a number of sample restrictions to our initial sample of approximately 16 million beneficiary-year observations. We limit our sample to those 65 and older who originally qualify for Medicare through the Old Age and Survivors 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, so that 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

ability to make an essentially continuous choice in that context as well (Ito 2014).

[3] 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 available demographic information. The risk scores are designed (by CMS) to be normalized to the average Part D beneficiary drug spending. For more information see http://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html (last accessed November 24, 2014).

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

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, comprising about 1.7 million unique beneficiaries.[5]

Panel A of Table 1 presents some basic demographic characteristics of the original full sample and our baseline 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.[6]

Prescription drug spending We use the detailed, claim-level information on prescription drug purchases to construct data on annual spending, as well as on the timing of purchases during the year. We also use the National Drug Codes (NDCs) to measure 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 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. Panel C reports the distribution of annual out-of-pocket spending, which ranges from zero to several thousand dollars annually.

Insurance contracts Insurance companies are required to offer a basic plan, which is either the government-defined "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 drops sharply to 0.25. That is, for every additional dollar spent at the pharmacy, the individual pays 25 cents out of pocket and the insurance company pays the remaining 75 cents. This 25%

[5] 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.

[6] 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.

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 $5.60 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.[7]

In analyzing the main cost-sharing features of the actual plans in our sample, we make two simplifying abstractions. First, we summarize cost-sharing in each plan-arm in terms of the percent of the total claim amount 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). 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.[8] 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 shows that these assumptions drive a (small) wedge between the stylized description of the plans and our empirical cost sharing calculations. For example, we estimate average cost sharing in the gap for plans with "no gap coverage" of 0.98, and cost-sharing in the deductible arm for plans with a deductible of 0.88. In principle, both of these numbers "should be" 1.

There are several thousand different plans in our sample, although the differences among them are sometimes minimal. Table 2 summarizes some of the main distinguishing plan features. About three-quarters of our sample chooses 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, reflecting the fact that insurance companies often find it attractive to offer an "actuarially equivalent" plan that, relative

[7] The standard benefit has the same basic structure in all years, although 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).

[8] Since very few individuals reach the catastrophic limit, computing plan-specific cost sharing above this limit is difficult. We therefore calculate 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.

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 0.93. However, it varies substantially based on whether Medicare classifies the plan as one with no or “some” gap coverage.

3 Descriptive patterns

3.1 Bunching at the kink

We examine 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 about 60 cents for every dollar spent in the pharmacy. Standard economic theory suggests that, as long as preferences (for healthcare and income) are convex and smoothly distributed in the population, we should observe individuals bunching at this convex kink point of their budget set; Appendix Figure A1 illustrates this intuition graphically. 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; Saez (2010) provides a formal discussion of this in the context of labor supply.

Figure 2 provides an empirical illustration of this theoretical response to a non-linear budget set. It shows 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 that 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.[9]

Figure 4 pools the analyses across the three years of data 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.

As one way to quantify the amount of excess mass, we follow the approach taken by Chetty et al. (2011) and approximate the counterfactual distribution of spending that would exist near

[9] Appendix Figure A2 further shows that for the small subsample of individuals (outside of 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, Appendix Figure A3 shows no evidence of missing mass at the concave kink created by the price decrease when individuals hit the deductible; in Appendix A we speculate about a potential explanation.

the kink if there were 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), 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 = 0.3%) in the -$200 to +$200 range of the kink, relative to the area under the counterfactual distribution 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 a statistically significant 29.1%. The presence of statistically significant excess mass around the kink allows us to reject the null of no behavioral response to price.

The average 60 cents (per dollar spent) increase in price at the kink masks considerable heterogeneity across plans, reflecting differences in cost-sharing both before and in the gap. Appendix Figure A4 plots plan-specific excess mass estimates (constructed in the same way as before) against the size of the plan-specific price change at the kink. As we would expect, we observe the excess mass at the kink to be increasing in the price increase at the kink.

We also explored heterogeneity in the behavioral response to price – as measured by the size of the excess mass around the kink – across different types of individuals. Table 3 shows the results. Column 3 shows a statistically significant excess mass in each sub-group we examine. The size of the excess mass increases with year of the Part D program, from 9% in the first year (2006) to 29% in the last year we observe (2009). This may reflect a "learning" effect (by individuals or pharmacists) about the presence of the gap.[10] The excess mass is slightly higher for men than for women, and tends to be larger for healthier individuals, as measured by age or the number of hierarchical conditions the individual has.[11] We defer a discussion of column 4 to Section 6 below.

3.2 Timing of purchases

The bunching in response to the kink presumably reflects individuals foregoing or postponing prescriptions that they would otherwise have filled.[12] An attractive feature of our setting is that, unlike in the classic labor supply setting for studying bunching (Saez 2010), we observe within-year

[10] 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.

[11] 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 is likely different than at other ages. The Hierarchical Condition Categories 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.

[12] It is unlikely that the bunching reflects purchases that are made but not claimed (due to the reduced incentives to claim in the gap). 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.

behavior, and thus can explore such timing channels in more detail.

Figure 5 shows the propensity to purchase at least one drug during a given month, as a function of the total annual spending. Because much of the response is at the end of the year, we present results (separately for each month) for the last four months of the year (September through December). As in the earlier graphical analysis of bunching (Figure 4), 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 annual spending, the share of individuals with at least one prescription drug purchase during that month.

Absent a price response, one would expect that the share of individuals with a purchase in a given month would monotonically increase with the level of annual spending (that is, with the overall frequency of purchases) and would approach one for sufficiently sick individuals who visit the pharmacy every month. Indeed, for individuals whose spending is far enough from the kink, this is the pattern that we see in Figure 5. But we also see a slowdown in the probability of end-of-year purchases as individuals get close to the kink. The pattern is extremely sharp in December, and (not surprisingly) less sharp for earlier months. In all months shown, the pronounced decline in purchase frequency is concentrated around the kink; once individuals are above the kink, the pattern reverts to the original (below the gap) monotone pattern, albeit at a lower-than-predicted frequency, presumably reflecting the higher cost-sharing in the gap.[13]

Overall, the figure illustrates that individuals respond to the kink by purchasing less, and that much of this response is concentrated late in the calendar year. This general pattern motivates some of the modeling choices we make in the next section. In particular, the pattern is consistent with uncertainty about future health events: if all claims were fully anticipated at the start of the year, one would not see a larger decline in purchasing later in the year. At the same time, the fact that there is a (albeit smaller) decline in purchasing before the end of the year is consistent with individuals responding to dynamic incentives.[14]

Focusing on the last month of the calendar year, we also examined how the behavioral response to price – as measured by the decline in end-of-year purchases around the kink – varies across types of drugs. Table 4 reports the results. Column 5 shows the estimated percent decline in purchase

[13] The fitted line in the graph illustrates the difference between actual purchase probabilities and what would be predicted in the absence of the kink. To fit the line, we run a simple regression of the logarithm of the share of individuals with no claim (during the corresponding month) 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 claims in the month monotone in the spending bin and asymptote to one as the bin amount approaches infinity.

[14] An alternative explanation for the decline in purchasing in earlier months of the year (e.g. in September; see Figure 5) is that some people have already hit the kink by September and therefore stop purchasing. However, if we restrict the sample to people who are not around our kink window in September, we continue to see a clear decline in purchasing patterns for those who end up around the kink by the end of the year. Therefore, at least some of the decline in earlier months is associated with a response to dynamic incentives.

probability at the kink in December. The first row reports our estimates for the entire sample, which mirrors the graphical analysis presented in the bottom right panel of Figure 5. On average, individuals who reach the gap reduce their probability of a December drug purchase 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; this may reflect greater flexibility regarding the timing of drug purchases for chronic conditions.[15] The probability of a branded drug purchase in December declines much more sharply at the kink than the probability of a generic drug purchase (20% compared to 8.5%).[16] The bottom row of Table 4 shows a greater (12%) decline in purchasing of "inappropriate" drugs compared to the average 8% decline in drug purchasing, providing some evidence that higher prices may lead to a somewhat more careful selection of drugs.[17] We defer a discussion of the right-most column until Section 6 below.

4 Model and estimation

The results in the previous section provided descriptive evidence of a behavioral response of drug purchases to the sharp increase in price as individuals enter the donut hole. The timing of the response points to the importance of dynamic incentives. In order to make counterfactual, quantitative inferences about what behavior would look like under alternative contracts, we develop and estimate a simple, dynamic model of an optimizing agent's prescription drug utilization decisions given a specific, non-linear contract design. The model focuses on several elements of individual choice behavior motivated by the preceding descriptive results and the highly non-linear nature of Medicare Part D prescription drug coverage.

4.1 A model of prescription drug use

We consider a risk-neutral, forward looking individual who faces stochastic health shocks within the coverage period.[18] These health shocks can be treated by filling a prescription. The individual is

[15] 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 analogous to being a drug for a chronic condition or not.

[16] The size of the kink is roughly similar in our sample for branded and generic drugs; we estimate a price increase of 60 cents and 55 cents respectively. However, because branded drugs tend to be much more expensive (on average, in our sample, the price of a branded drug is about $130 compared to about $20 for generics), the per-prescription (rather than per-dollar) price effect of entering the gap is significantly greater for branded drugs.

[17] Following Zhang et al. (2010), we proxy for inappropriate drug 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).

[18] Risk neutrality simplifies the intuition 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. We find this has little effect

covered by a non-linear prescription drug insurance contract j over a coverage period of T weeks.[19] In our setting, as in virtually all health insurance contracts, the coverage period is annual (so that, typically, $T = 52$). 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

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

When health events arrive they are drawn independently from a distribution $G(\theta, \omega)$. It is also convenient to define $G(\theta, \omega) \equiv G_2(\omega|\theta)G_1(\theta)$.

Health events arrive with a weekly probability λ', which is drawn from $H(\lambda'|\lambda)$ where λ is the weekly arrival probability from the previous week. We allow for serial correlation in health by assuming that λ' follows a Markov process, and that $H(\lambda'|\lambda)$ is (weakly) monotone in λ in a first order stochastic dominance sense.

The only choice individuals make is whether to fill each prescription. Optimal behavior can be characterized by a simple finite horizon dynamic problem. The three state variables are the number of weeks left until the end of the coverage period, which we denote by t, the total amount spent so far, denoted by x, and the health state, summarized by λ, which denotes the arrival probability in the previous week.[20]

The value function $v(x, t, \lambda)$ represents the present discounted value of expected utility along the optimal path and is given by the solution to the following Bellman equation:

$$v(x,t,\lambda) = \int \left[(1-\lambda')\delta v(x, t-1, \lambda') + \lambda' \int \max \left\{ \begin{array}{l} -c_j(\theta, x) + \delta v(x+\theta, t-1, \lambda'), \\ -\omega + \delta v(x, t-1, \lambda') \end{array} \right\} dG(\theta, \omega) \right] dH(\lambda'|\lambda). \tag{2}$$

Optimal behavior is straightforward to characterize: if a prescription arrives, the individual fills it if the value from doing so, $-c_j(\theta, x) + \delta v(x+\theta, t-1, \lambda')$, exceeds the value obtained from not filling the prescription, $-\omega + \delta v(x, t-1, \lambda')$. In our baseline model we assume that terminal conditions are given by $v(x, 0, \lambda) = 0$ for all x; in Section 6 we extend the model to allow for cross-year substitution by making the terminal values a function of unfilled prescriptions.

on our main counterfactual estimates.

[19] Aggregating to the weekly level reduces the computational cost of estimating the model. This seems a reasonable approximation given that many prescriptions may arrive as a "bundle" that needs to be consumed together.

[20] The Markov process for λ creates the typical initial condition problem. Throughout, we assume that in the first week of the coverage period, the health state is drawn from the steady-state distribution of λ.

To summarize, the model boils down to a statistical description that describes the individual's health – the arrival rate of prescriptions λ', its transition over time $H(\lambda'|\lambda)$, and the associated (marginal) distribution of cost $G_1(\theta)$ – and two important economic objects. The first economic object, summarized by $G_2(\omega|\theta)$, can be thought of as the "primitive" price elasticity that captures substitution between health and income. The presence of a non-zero price elasticity was suggested by the descriptive evidence of bunching at the kink. 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 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 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. 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.

The second 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 of subsequent out-of-pocket prices relative to the immediate out-of-pocket price for the current utilization decision.[21] 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 other work (Aron-Dine et al. 2014), we presented evidence that individuals do not respond only to the current, "spot" price of medical care in making utilization decisions (i.e. we reject $\delta = 0$) both in the Medicare Part D context and in the context of employer-provided health insurance. Likewise, the claim timing patterns (Figure 5) showing a drop in the probability of monthly claiming by people in the kink, which is much more pronounced toward the end of the year, is also consistent with within-year uncertainty about the distribution of health-related events (as captured by λ and $G_1(\theta)$) and a positive discount factor (captured by $\delta > 0$).

[21] 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.

4.2 Parameterization

To estimate the model, we need to make three 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 assumption is the nature of the Markov process for the weekly probability of health event λ. We assume that, for each individual, λ can take one of two values, λ^L and λ^H (with $\lambda^L \leq \lambda^H$), and that $\Pr(\lambda_t = \lambda^L|\lambda_{t+1} = \lambda^L) = \kappa^L \geq 0.5$ and $\Pr(\lambda_t = \lambda^H|\lambda_{t+1} = \lambda^H) = \kappa^H \geq 0.5$, so there is (weakly) positive serial correlation. This simple parameterization of the Markov process for λ is computationally attractive, as we only need to consider two possible values for the third state variable.

The third type of assumption is about parameterization of heterogeneity across individuals in a given year. Since individuals' health is likely serially correlated, even conditional on the serial correlation introduced above, 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). The parameters μ_m, σ_m, and p_m are all allowed to flexibly vary across types. The discount factor δ, which is more difficult to identify in the data, is assumed to be the same for all types. For λ, we allow λ_m^L to vary flexibly by type, but then impose the restriction that κ^L, κ^H, and the ratio $\lambda^L/\lambda^H \leq 1$ is the same for all types. Our parameterization allows for heterogeneity in both individual health (λ, μ, σ), and in the responsiveness of individual spending to cost-sharing (p). Overall, the model has $4M$ parameters that define the M quadruplets $(\mu_m, \sigma_m, \lambda_m^L, p_m)$, the single parameter δ, three parameters for κ^L, κ^H, and the ratio λ^L/λ^H, 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 36 parameters to estimate.

Our choice of parameterization imposes a number of limitations that deserve further discussion. First, 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 ω above θ does not affect the model's predictions, and therefore has no effect on our counterfactual exercises. Second, 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. We also report in the robustness section below an alternative specification that allows the distribution of ω to depend on θ. Third, although we allow for heterogeneity across individuals in the distribution of health events and in their behavioral response to the contract (as shown in Table 3), we do not directly model the choice of drug (e.g. brand vs. generic, or type of molecule), and relatedly abstract from heterogeneity across drugs in the behavioral response to the contract (as shown in Table 4). To the extent that drug type varies across people, this may be partially and indirectly captured by allowing for heterogeneity in response rates and the distribution of health events across individuals, and we explore related parametric assumptions in the robustness section. This simplification, however, allows us to think about drug expenditure in monetary terms, and to have a unified model for all spending decision. A more detailed model of health conditions and potential cures would make it very difficult to uniformly treat all possible drugs, and would make it less consistent to uniformly treat all plans in a similar way by converting co-pay plans to co-insurance plans (as mentioned in Section 2).

4.3 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 3) allows us to identify the spending response to price where the spot and future price are the same (as in a linear contract). Second, timing of purchases (also shown in Section 3) helps in identifying the discount factor δ. The larger δ is, the more current purchases would respond to expected total spending, and hence the greater the decline we would see in earlier months in Figure 5.[22] Finally, observing weekly claims made by the same individual over the entire year, along with our assumption that the (unobserved) type is constant throughout the year and health status follows a Markov process, 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, $H_m(\lambda_t|\lambda_{t+1})$ and $G_1(\theta|m)$, the substitutability between income and health, $G_2(\omega|\theta, m)$, and the parameter δ. The results of Hu and Shum (2012) show the nonparametric identification of the distribution of types, π_m, the conditional (on type) distribution of θ, $G_1(\theta|m)$, conditional claim probabilities, $P(\text{claim}|m, \theta, x, t)$, and distribution of λ, $H_m(\lambda_t|\lambda_{t+1})$. Given the distribution of health status and conditional claim probabilities, the non-linearity of the contract generates variation in incentives that traces out the distribution of ω.

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

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

where $\overline{G}_2\left(\cdot|m,\theta\right)$ 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

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

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

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

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

[22] To assess the importance of these moments for the actual identification, we follow the procedure recently proposed by Gentzkow and Shapiro (2014). We find that the estimation moments (described below) that are associated with the timing of drug purchases at the end of the year account for approximately 20% 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 48-72% of the contribution of all the moments to the estimation of the different p's.

so we can identify $\overline{G}_2\left(\cdot|m,\theta\right)$ 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 $\overline{G}_2\left(\cdot|m,\theta\right)$ 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 $\overline{G}_2\left(\cdot|m,\theta\right)$, 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.

4.4 Estimation

We estimate the model using simulated minimum distance on a slightly modified "baseline"sample.[23] 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

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

where W_n is an estimate of the inverse of the asymptotic variance of the sample statistics. Appendix B describes in detail how we solve for the value function and simulate our model.

We use several “types” of moment conditions, with our choices motivated by some of the key descriptive patterns in Section 3. One type of moments summarizes the distribution of annual spending that is shown in Figure 2, with particular emphasis on the bunching pattern around the kink. Specifically, 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 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 the frequency of each bin as a moment we try to match.

A second type of moments focuses on the claim timing pattern around the kink, as shown in Figure 5. Specifically, we construct twelve such timing moments. For each month from July to December, we use two moments: the share of individuals with at least one claim in the month conditional on the individuals' total spending in the year being within $150 of the kink, and an analogous share conditional on the individuals' total spending being between $800 and $500 below the kink. Finally, a third type of moments captures the persistence of individual spending over time, for which we use the covariance between spending in the first half and second half of the year.

It may be useful to highlight some computational 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

[23] To reduce computational cost, we make two inconsequential restrictions to our baseline sample. We restrict to the 500 most common plans; this represents about 10% of plans but about 90% of beneficiary-years. From this modified baseline sample we use a 10% random sample.

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 2013).

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.

5 Results

5.1 Parameter estimates and model fit

Table 5 presents the parameter estimates. We find δ to be relatively close to one, at 0.96. 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 5). The second, third, and fifth types are the most common and together account for about 93% of the individuals. As should be expected, increases in risk score are associated with increased probability of the highest spending types (types 4 and 5) and decreased probability of lower spending types (types 2 and 3). It is interesting to note that these (predictable) average correlations between risk score (and age) and spending type are not estimated to be monotone across types. For example, individuals with higher risk scores are more likely to be of types 4 and 5 and less likely to be of types 2 and 3, but their probability of being the lowest spending type (type 1) – which only accounts for about 5% of the beneficiaries – is not affected much. This pattern may point to multi-dimensional heterogeneity, and to important (unobserved) individual traits that are associated with low drug expenditure for reasons that are less related to health.

The parameter estimates suggest quite modest positive serial correlation in health on a week-to-week basis. The event probability in the "sicker" state (λ^H) is estimated to be only 13% higher than the event probability in the "healthier" state (λ^L) and the probability of each state does not change much with the health state in the earlier week. These results suggest that conditional on allowing for unobserved heterogeneity across individuals in their ("permanent") health state for the year, the remaining week-to-week correlation is not very important.[24]

A type's health is characterized by the rates of arrival of prescription drug events (the λ's) and the distribution of their size (θ). The first type is fairly healthy, with relatively low event probabilities and small claim amounts when there is a (potential) claim (i.e. low $E(\theta)$). It is interesting to observe that the two highest spending types (types 4 and 5) exhibit different patterns. While type 4 has a very high event probabilities (0.78 and 0.88), the expected cost of a (potential) claim conditional on an event is relatively low (82 dollars). In contrast, type 5 is half as likely to experience a health event, but once he has one, the expected cost of the potential claim is more than doubled. Heuristically, one could think of type 4 as an individual with chronic conditions, while type 5 has fewer chronic problems, but is generally sicker and experiences frequent acute diseases.

Annual spending depends not only on health but also on the propensity to purchase (i.e. to fill a prescription in response to a health event). This purchase propensity depends on the parameter p, which likewise determines how responsive drug purchasing may potentially be to the cost-sharing features of the contract. As it turns out, the estimated parameter is highly correlated with how sick individuals are. The heathiest types (types 1 and 2), who have the lowest expected spending, are also the types that have the highest estimates of p (0.86 and 0.9), and are thus the most responsive to the coverage features. In contrast, the sickest type (type 5) has the lowest estimate of p (0.37) and in two out of three drug events, he would fill the drug regardless of the coverage.

Overall, the model fits the data quite 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. Figure 6 presents the distribution of spending, for both the observed and the predicted data; we show both the

[24] Indeed, the raw data is generally consistent with this finding. To see this, we look at the raw correlation between an individual's spending in one week and the week before; we also performed the same exercise at the monthly level. When we do not account for heterogeneity, this correlation is low (0.02) at the weekly level but high (0.50) at the monthly level, presumably reflecting the lumpiness of (often 30-day) prescriptions. However, if we first subtract individual fixed effects, and then compute the correlation for the within-individual residual, we actually obtain small and negative serial correlation at both the weekly and monthly level (-0.15 and -0.10, respectively).

fit of the overall spending distribution and the "zoomed in" fit within $1,000 of the kink. The patterns are extremely similar. Figure 7 shows the observed and predicted patterns of monthly claim probabilities, separately for individuals who are below and around the kink. The model predicts well the sharp drop in claim probabilities toward the end of the year for individuals who end up around the kink (dark lines) and the lack of such a drop for individuals who end up below the kink (grey lines); however, the overall fit is not as striking as in Figure 6, presumably reflecting the fact that monthly claim probabilities are much more noisy than annual spending measures.

Figures 6 and 7 relate to the moments we explicitly try to fit. Appendix Figures A5-A7 present the fit of our model in three out-of-sample cases: for another 10% random sample from the estimation sample, for a set of small plans excluded from the estimation sample,[25] and for the 2010 calendar year, which is not part of the original data, which was limited to 2007-2009. The out-of-sample predictions are reassuring and fit quite well. For the case of 2010, however, while the model does predict the right direction out-of-sample by which spending changes relative to the baseline sample, the fit is not as good as the other cases. This may not be surprising: simple macroeconomic time trends in, say, drug prices may generate differences between the model prediction and the data, and our predictions are not designed to capture such trends.

5.2 Spending response to counterfactual contract designs

The 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. To do so, we generate model predictions in precisely the same way that we assessed goodness of fit in Figures 6 and 7, except that we now also simulate 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.

We focus on a policy-relevant counterfactual. As part of the 2010 Affordable Care Act (ACA), there will no longer be a gap in the government-defined standard benefit contract by 2020: 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."

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

[25] Recall (see footnote 23) that to reduce computation time for model estimation, we limit the baseline sample to the 500 most common plans and then use only a 10% random subsample.

focus on a single contract is useful for illustrating particular aspects of the spending response to alternative contract designs.

Row 1 of Table 6 shows spending under the 2008 standard contract, and row 2 shows the results of filling the gap. On average, total spending increases by $204, or about 12%, from $1,760 to $1,964. This increase in total spending reflects the combined effect of a $154 decline in average out-of-pocket spending and about $358 increase in insurer spending (rightmost 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 6, we see that the median increase in total spending is only about $40, while the 90th percentile change is about $820. The top panel of 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 the biggest change in spending from filling the gap is for individuals whose predicted spending under the standard contract would be in the gap. However, it 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. Consistent with this "anticipatory effect," we see an increase in spending 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 "anticipatory" effect accounts for about 25% of the average $204 increase in annual drug spending.[26]

A second, somewhat counter-intuitive finding in the top panel of Figure 8 is that filling the gap causes some previously high spending 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. 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.

The bottom panel of Figure 8 illustrates how the filled gap affects the change in spending through the lens of calendar time. The figure plots average weekly spending for individuals under

[26] The increase in spending among people more than $200 below the kink location under the standard plan is $74 on average, and this portion of the annual spending distribution accounts for 70% of the people.

the 2008 standard benefit design (gray line) and under the filled gap contract (black line). Under the filled gap contract – which provides more coverage and a lower expected end-of-year price than the standard contract – spending is higher in every week throughout the year. The fact that spending is higher even at the very beginning of the year – when no one has hit the gap yet – reflects the positive δ we have estimated. However, as the end of the year gets closer, and the realization of health events reduces residual uncertainty, a share of the individuals see their expected end-of-year price rise sharply under the standard contract (but not under the "filled gap" contract), leading to greater divergence between predicted spending under the standard contract and the filled gap contract. Under the standard benefit contract, many individuals slow down their spending quite a lot over the last few months of the year (and especially in the last few weeks), while under the "filled gap" contract this effect does not exist and average weekly spending does not decline toward the end of the year.

Filling the gap in the observed 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. In rows 3 and 4 of Table 6, therefore, we examine the impact of filling the gap in the observed distribution of contracts in our data.[27]

Given the observed distribution of plans in the data, we estimate that filling the gap will raise total annual drug spending by $148 per beneficiary, or about 8%, from $1,768 to $1,916. Aggregating across affected beneficiaries, this suggests that filling the gap will raise total prescription drug spending by about $1.9 billion per year.[28] About one quarter of this increase comes from increases in spending by individuals who are predicted to spend $200 below the kink location or less under their original plan, suggesting a quantitatively important role for "anticipatory" behavior.[29] Medicare spending increases substantially more than total spending, by about $253 per beneficiary (right most column); by contrast, ignoring the behavioral response to the contract, we calculate that "filling the gap" would increase average Medicare spending by only $150.[30]

[27] 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.

[28] As described in Section 2, our 3 year, 20 percent sample of Medicare beneficiaries includes 7.4 million beneficiary-years who would be affected by the requirement to "fill the gap" (we exclude individuals who are dual eligibles or have low income subsidies for whom the gap is already "filled"). This implies about 12.3 million beneficiaries per year affected by the "filled gap" counterfactual, or – multiplied by $150 per beneficiary increase in annual drug spending – about a $1.8 billion increase in annual prescription drug spending.

[29] We estimate the increase in spending among people more than $200 below the kink location is $57 on average, and this portion of the annual spending distribution accounts for 71% of the people.

[30] 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

Reducing cost-sharing on all arms Naturally we can use the model and its results to predict counterfactual spending under other counterfactual contract designs as well. To try to move beyond policy-specific counterfactuals to a more general economic object that could potentially be applied to other budget sets or compared 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. Appendix Table A1 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.[31] The implied elasticity of drug spending with respect to the price ranges from about -0.5 (for a 1% reduction in cost-sharing throughout the budget set) to -0.3 (for 7.5% and higher reductions in cost-sharing throughout the budget set).

5.3 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. For each robustness check, it reports the implied effect on "filling the gap" on total and insurer spending. Overall, the results appear reasonably stable across specifications. Specifically, across the specifications (discussed below), the estimated increase in total annual drug expenditures from filling the gap ranges from 6% to 11%, which is qualitatively similar to our baseline estimate (of 8.4%). Similarly, the estimated increase in insurer expenditure ranges between 22% and 29%, while our baseline estimate was 26%. 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 two of the five types in our baseline specification was relatively small (3% and 5%; see Table 5), it may not be surprising that adding a sixth type or eliminating two types have relatively small affect on the results. 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

on how Medicare reimburses insurers.

[31] We report elasticity estimates in Table A1 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).

65 years old as the two covariates (see Table 5). 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 attempt to capture potential plan selection on unobservables, capturing, for example, that unobservably 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, 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"). 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.

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 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 4 and specify a utility model that allows for a concave utility function. More details are provided in Appendix C.1. To summarize, 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.[32] The limit, as α approaches zero, is equivalent 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.

Finally, in row 8 of Table 7 we explore the robustness of our results to the distributional assumption of ω. Our baseline specification assumes that moral hazard is captured by a (type-specific) parameter p_m: ω is assumed to be drawn from a uniform distribution over $[0, \theta]$ with probability p_m, and ω is greater than θ (and therefore inconsequential) with probability $1 - p_m$.

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

One concern about this assumption is that it assumes that prescriptions that are more expensive to fill have similar price responsiveness as prescriptions that are cheaper. One could imagine that expensive prescriptions are more likely to be associated with more serious conditions, which in turn are associated with less consumer discretion and lower moral hazard. Another possibility is that ω increases less than proportionally in the cost of filling a prescription, and moral hazard is therefore more prevalent for higher values of θ. To explore the sensitivity of the results to this possibility, we enrich the parameterization by allowing p_m to depend on θ. That is, we still assume that ω is a mixture of a draw from a uniform distribution over $[0, \theta]$ or that it is greater than θ, but the probability for the former is now a function of θ, $p_m(\theta)$. Our specific parameterization is

$$p_m(\theta) = \frac{\exp(\rho\theta)}{\frac{1}{p_m} - 1 + \exp(\rho\theta)}, \tag{12}$$

where p_m is a type-specific moral hazard parameter, as in the baseline specification, and ρ is a new parameter that we estimate (identical across types) that allows for a correlation between ω and θ. This specification nests our baseline model (when $\rho = 0$), but could also allow for positive correlation between p_m and θ (when $\rho > 0$) or negative correlation (when $\rho < 0$).

The results from this specification gives rise to an estimate of $\rho = 0.000073$, which implies a small positive relationship between p_m and θ. To get a sense of magnitude, consider $p_m = 0.6$ (our (weighted) average estimate of p_m across types; see Table 5) and two prescription of \$10 and \$1,000. The above estimate implies that the more expensive prescription is associated with a $p_m(\theta)$ that is 3% greater. Given this small correlation, it is not surprising that this additional flexibility in the model makes little difference to the overall results of filling the gap (Table 7, row 8).

6 Cross-year substitution

With the exception of Cabral's (2013) recent work, empirical analyses of the spending effects of health insurance contracts have typically focused on annual spending effects. Our results thus far have followed this tradition. However, our focus on the non-linear nature of annual health insurance contracts raises the possibility that they may induce individuals to engage in cross-year substitution, and that changes to contract design may therefore have different impacts over a longer-than-annual period.

Since non-linear contracts are the norm in health insurance, this raises the possibility that the traditional annual analysis of behavioral effects of insurance contracts may not give an accurate portrayal of their net spending effects. In our context, individuals who end up in the gap may defer some late-in-the-year drug purchases that do not require immediate attention to the beginning of the next year when the coverage schedule "resets" and the expected end-of-year price (as well as the spot price for the many individuals in no-deductible plans) is lower. When the gap is "filled," such cross-year substitution incentives are reduced. If a large part of the annual spending response to filling the gap reflects the end of such deferrals of drug purchases to the following year, the budgetary implications of filling the gap may be very different if considered over a longer time horizon than the annual analysis that is traditionally done, and that we have engaged in thus far. In this section, we explore this possibility empirically.

6.1 Descriptive evidence

We saw in Figure 5 that individuals slow their propensity to purchase drugs late in the year once they are near the kink. This raises the question of whether they never purchase these prescriptions or simply shift the purchase to the beginning of the next calendar year. To explore this possibility of cross-year substitution, we examine whether there is a relationship between January spending in the following year (year $y+1$) and total annual expenditures relative to the kink in year y.

We define an individual's relative spending in January $y+1$ as the ratio of her January spending in year $y+1$ to her average monthly spending in March to June of $y+1$.[33] The top panel of Figure 9 graphs this measure of "relative January spending" in year $y+1$ against total annual expenditures in year y. If the slowdown in purchasing propensity toward the end of the calendar year as individuals approach the gap simply reflects a decline in drug purchases, there should be no systematic relationship between relative January spending in the subsequent year and prior year's spending. However, if some of the slowdown in purchasing reflects cross-year substitution toward filling the same prescriptions after coverage resets, we should expect to see higher relative January spending in year $y+1$ for individuals who approach (or enter) the gap in year y.

The results in the top panel of Figure 9 strongly suggest that such cross-year substitution 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 considerably relative to 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. We define "excess" January spending in year $y+1$ as the ratio of average relative January spending in year $y+1$ for individuals whose annual spending in year y is between the kink and \$500 above it, to average relative January spending in year $y+1$ for individuals whose spending is \$500 to \$2,000 below the kink in year y. We estimate excess January spending of 25%. In other words, average relative January spending in year $y+1$ for individuals near the gap in year y is 25% higher than average relative January spending in year $y+1$ for individuals further below the gap in year y.

In Appendix Figure A8 we show that, as one would expect, cross-year substitution is greater for beneficiaries who are covered by a no-deductible plan. Approximately one-quarter of the sample has a deductible plan in both years, so that if they end up in the gap in year y they face little change in spot price on January 1 of year $y+1$, although their expected end-of-year price likely drops. For this population, the January effect does not disappear, but its magnitude is noticeably

[33] We omit February in case some of the January effect "spills" to February, and we omit July to December that are likely affected by potential substituting between years $y+1$ and year $y+2$. The qualitative results are not particularly sensitive to the set of months used to define an "average month."

lower (19% excess January spending, compared to 38% for those in a no-deductible plan in both years).

We also explored heterogeneity in the extent of cross-year substitution across different types of people and drugs. Consistent with cross-year substitution being one channel through which behavioral responses operate, we find that drugs and individuals that are associated with a relatively high response to the kink also tend to be associated with a relatively high degree of cross-year substitution. Column 4 of Table 3 shows excess January spending – our measure of cross-year substitution – for different groups of individuals. We see, for example, that, like excess mass in column 3, excess January spending in column 4 is much greater for younger beneficiaries than for older ones. Column 6 of Table 4 shows excess January spending for different types of drugs; in general, drugs with a larger decline in the probability of a December purchase in column 5 also have a greater excess January spending in column 6.

The bottom panel of Figure 9 provides evidence that bunching at the kink in year y cannot be entirely explained by cross-year substitution. We follow a strategy used by Chetty et al. (2011) and present the density of "adjusted annual spending," which is the sum of year y's spending and the average "excess" January spending (in dollars).[34] 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 y spending relative to the kink) remains when the density is plotted as a function of this "adjusted" year y spending: it would have been eliminated if the entire response was due to shifting purchasing to January.[35] We can therefore reject the null that cross-year substitution can account for all of the spending effect at the kink.

[34] Analogously to the top panel of Figure 9, the excess January spending is measured as the difference, for each spending bin in year y, between the average January spending in year $y+1$ and the average monthly spending in March to June of year $y+1$. We then add the bin-specific (but not beneficiary-specific) excess January spending to each beneficiary's year y spending.

[35] Specifically, the adjustment shrinks the excess mass from our baseline estimate of 0.291 (standard error = 0.003), to 0.242 (standard error = 0.004). We note that the size of the adjusted excess mass gets smaller 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. (Indeed, consistent with this, an alternative, "placebo" exercise, which adjusts for the difference between the individual's year $y+1$ *July* spending and the average monthly spending in March-June of year $y+1$, also make most of the bunching disappear.). Yet, under the null that the entire response is driven by cross-year substitution to January, the adjusted bunching would be eliminated for any bin size within the range we examine excess mass. Therefore, the fact that it appears large and significant for the bin size plotted in the bottom panel of Figure 9 is sufficient to reject the possibility that the entire response at the kink is driven by shifting claims to January.

6.2 Extending the model to allow for purchase delays

The above descriptive evidence indicates the existence of cross-year substitution, but rejects the hypothesis that it can explain all of the spending response to the kink. This raises the question of how much of the increase in annual spending that we estimated would come from filling the gap represents a decrease in cross-year substitution, rather than a net increase in spending measured over a time horizon of more than one year. One crude, back-of-the-envelope calculation would be to assume that the 25% excess January spending we estimated for those who spend just beyond gap (Table 3, column 4, row 1) applies to the entire population (even those who are not close to the gap). Given mean annual spending of about \$1,900 (Table 1), a 25% increase in January spending corresponds to about \$40 per beneficiary in cross-year substitution. Assuming that all of this \$40 in cross-year substitution is eliminated when the gap is filled, this suggests that cross-year substitution may account for about a quarter of our \$150 estimate of the annual spending increase from filling the gap.

A problem with this exercise, however, is that our baseline model – from which we estimated that filling the gap would increase annual spending by \$150 – does not itself allow for cross-year substitution. It assumes that terminal values are given by $v(x, 0, \lambda) = 0$ for all x. That is, every January everything resets, and every beneficiary starts the new calendar year with a a "clean slate," regardless of his earlier drug purchase decisions. This is, of course, a simplification, designed to make the model easier to estimate, identify, and understand. However, we have just shown that this may be an important abstraction for the spending implications of filling the gap over horizons of more than a year.

We therefore examine a stylized extension of the baseline model designed to assess the quantitative importance of cross-year substitution for our baseline estimates. The baseline model assumes that individuals make a binary decision: either to fill a prescription or to never fill it. We now allow a third, intermediate decision: to defer treatment. We make the simplifying assumption that all the delayed treatments over the year are collected together, and are treated at one later point, in January of the subsequent year; this greatly reduces the estimation and computational burden. Moreover, we assume that individuals make this decision to defer filling a prescription under the assumption that the out-of-pocket price they will have to pay for these deferred prescriptions is known at the time of the decision to defer and is given by q_i, which we assume is a function of the individual's risk score and plan.[36] This strong assumption provides intuitive predictions: individuals who hit the catastrophic level or those who hit the deductible but don't expect to hit the kink may decide to not fill a prescription, but would never defer filling a prescription: the price they would have to pay in January is higher than what they have to pay when the prescription arrives. In contrast, individuals who are at the gap (and expect to stay there) are most prone to prefer to defer filling the prescription until January of the next year.

The rest of the model is similar to the baseline model, with minor modifications that allow us

[36] Specifically, we divide the risk score distribution into 3 equally-sized bins (lowest third, middle third, and highest third) and assume that q_i is given by the average end-of-year price in each plan and risk score bin.

to let the utility from deferring filling a prescription differ from the utility from never filling it. As before, a drug event is described by a pair (θ, ω), except that we now think of ω as the weekly flow of disutility of not treating the event (rather than the present value of not treating it, as in the baseline model). We denote by $\delta_h \in [0,1]$ a "new," additional depreciation parameter associated with deferred treatments, capturing the possibility that the disutility associated with not taking the drug may go away over time.

Given these assumptions, the revised Bellman equation becomes

$$v_i(x,t,\lambda) = \int \left[\begin{array}{c} (1-\lambda')\delta v_i(x,t-1,\lambda') + \\ \lambda' \int \max \left\{ \begin{array}{c} -c_j(\theta,x) + \delta v_i(x+\theta, t-1, \lambda'), \\ -\omega \frac{1-(\delta\delta_h)^t}{1-\delta\delta_h} - \delta^t \delta_h^t q_i \theta + \delta v_i(x,t-1,\lambda'), \\ -\frac{\omega}{1-\delta\delta_h} + \delta v_i(x,t-1,\lambda') \end{array} \right\} dG(\theta,\omega) \end{array} \right] dH(\lambda'|\lambda), \tag{13}$$

where the maximum operator reflects the three possible decisions. The first option is to fill the current prescription, pay the corresponding out-of-pocket price $c_j(\theta, x)$, and update the first state variable. The second option is to defer filling the prescription until January of next year: $\omega\frac{1-(\delta\delta_h)^t}{1-\delta\delta_h}$ is the present value of the utility loss from not treating the health event between the current week until next January, and $\delta^t\delta_h^t q_i\theta$ is the present value of the out-of-pocket cost of filling the prescription in January. The final option is to never treat the event, and the net present value of the disutility of this is given by $\frac{\omega}{1-\delta\delta_h}$. It is thus easy to see that without the option to defer, the extended model is reduced to our baseline model, except that we have to renormalize ω by $1-\delta\delta_h$. The problem remains a finite horizon dynamic problem, and the terminal value remains $v_i(x, 0, \lambda) = 0$ as a normalization (we could have equivalently defined them to be a function of the deferred prescriptions, but because there is full certainty of the January price, we already account for the associated costs of deferred prescriptions inside the Bellman equation).

Estimation and identification of the model is generally analogous to that of the baseline model. In Appendix C.2 we provide more details, but the key difference to highlight is that in addition to the moments we use for the estimation of the baseline model, we also add a moment, which attempts to capture the extent of cross-year substitution, and thus helps in the identification of the key "new" parameters that drive cross-year substitution (δ_θ and δ_ω).

The bottom panel of Table 6 reports the main results.[37] We report the sum of both spending during the coverage year and any additional January spending that was deferred. That is, the estimate accounts for the total spending effect of filling the gap, including the amount that was deferred until January, which is likely to be higher for the observed contracts relative to contracts in which the gap is filled. Consistent with the descriptive evidence in the bottom panel of Figure 9, we find that cross-year substitution cannot explain the entire spending effect of the gap. However, our estimates suggest that accounting for cross-year substitution is quantitatively important: it

[37] Appendix Table A2 reports all the parameter estimates from this model. We note, however, that the interpretation of some of the parameters have changed – in particular, ω now describes flow utility rather than a stock – so their values are not directly comparable to those in the baseline model estimate (Table 5).

reduces the effect of filling the gap by about two-thirds, from our baseline estimate of $148 per beneficiary to $44.

7 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. We provide descriptive evidence of a behavioral response to the kink in the individual's budget set created by the famous "donut hole." We then specify and estimate a simple dynamic model of prescription drug use that allows us to analyze the response to the entire non-linear budget set. We focus our counterfactual analysis on the impact on spending of the ACA-legislated "filling of the donut hole" by 2020. We estimate that this will increase total annual drug spending by $150 per beneficiary (or about 8%), and Medicare drug spending by much more ($260 per beneficiary, or about 25%). However, we also estimate that about two-thirds of this annual spending increase may be explained by a decline in substitution of purchases to the subsequent year, rather than a net increase in spending. Recognition of the incentives created by non-linear contracts thus suggests the importance of analyzing the behavioral response to health insurance contracts over a time horizon of longer than one year, contrary to the current practice in the empirical literature.

Our analysis illustrates several other 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 about one-quarter of the increase in annual drug spending from filling the gap.

Our paper has not explored several areas that might be of interest in future work. One is to consider how mandated changes in contract design (such as "filling of the donut hole") may affect other margins: the contracts offered by insurers, the plan-selection response of beneficiaries, and the pricing of drugs by pharmaceutical companies. Another is to consider the normative implications of our positive analysis. Some of our findings may be useful inputs here, including our findings 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, concerns that incomplete information or potential failures of rationality may lead individuals to under-consume drugs in the absence of insurance raises the possibility that insurance-induced increases in drug consumption may be efficiency enhancing (Baicker et al. 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. 2014. "Prescription Drug Use under Medicare Part D: A Linear Model of Nonlinear Budget Sets." Unpublished mimeo.

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

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.

Aron-Dine, Aviva, Liran Einav, Amy Finkelstein, and Mark Cullen. 2014. "Moral Hazard in Health Insurance: Do Dynamic Incentives Matter?" Mimeo. Available at http://web.stanford.edu/~leinav/Forward_Looking.pdf

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2012. "Behavioral Hazard in Health Insurance." NBER Working Paper No. 18468.

Bajari, Patrick, Han Hong, Minjung Park, and Robert Town. 2011. "Regression Discontinuity Designs with an Endogenous Forcing Variable and an Application to Contracting in Health Care." NBER Working Paper No. 17643.

Cabral, Marika. 2013. "Claim Timing and Ex Post Adverse Selection." Mimeo. Available at http://www.marikacabral.com/Cabral_ExPostAdverseSelection.pdf

Chetty, Raj. 2012. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica* 80(3): 969-1018.

Chetty, Raj, John N. 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 N. Friedman, and Emmanuel Saez. 2013. "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Revie*w 103(7): 2683-2721.

Dalton, Christina. 2014. "Estimating Demand Elasticities using Nonlinear Pricing." *International Journal of Industrial Organization.* November. Volume 37: 178-191.

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 Effect of Medicare Part D on Pharmaceutical Prices and Utilization.” *American Economic Review* 100(1): 590-607.

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 M. Shapiro. 2014. “Measuring the Sensitivity of Parameter Estimates to Sample Statistics." NBER Working Paper No. 20673.

Gowrisankaran, Gautam, Christina Marsh, and Robert Town. 2014. “Myopic and Complex Dynamic Incentives: Evidence from Medicare Part D.” Mimeo. Available at http://joris.pinkse.org/Gowrisankaran.pdf

Handel, Benjamin R. 2013. “Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts.” *American Economic Review* 103(7): 2643-2682.

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). 2010. Washington, DC: National Committee for Quality Assurance. http://www.ncqa.org/tabid/59/default.aspx.

Heiss, Florian, Daniel McFadden, and Joachim Winter. 2010. “Mind the Gap! Consumer Perceptions and Choices of Medicare Part D Prescription Drug Plans.” in David Wise (ed.), *Research Findings in the Economics of Aging*, University of Chicago Press: Chicago, IL.

Heiss, Florian, Adam Leive, Daniel McFadden, and Joachim Winter. 2013. “Plan Selection in Medicare Part D: Evidence from Administrative Data.” *Journal of Health Economics* 32(6): 1325-1344.

Hu, Yingyao, and Matthew Shum. 2012. “Nonparametric identification of dynamic models with unobserved state variables.” *Journal of Econometrics* 171(1), 32-44.

Ito, Koichiro. 2014. “Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing.” *American Economic Review* 104(2): 537-563.

Joyce, Geoffrey, Julie Zissimopoulos, and Dana Goldman. 2013. “Digesting the Doughnut Hole.” *Journal of Health Economics* 32(6): 1345-1355.

Kaiser Family Foundation. 2014. "Medicare at a Glance." Available at http://kff.org/medicare/fact-sheet/medicare-at-a-glance-fact-sheet

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): 2639-2673.

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(2): 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): 199-235.

Kowalski, Amanda. 2012. "Estimating the Tradeoff Between Risk Protection and Moral Hazard with a Nonlinear Budget Set Model of Health Insurance." NBER Working Paper No. 18108.

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

Polyakova, Maria. 2014. "Regulation of Insurance with Adverse Selection and Switching Costs: Evidence from Medicare Part D." Mimeo. Available at http://web.stanford.edu/~mpolyak/Polyakova_PartD_inertia.pdf

Rios, Luis Miguel, and Nikolaos V. Sahinidis. 2013. "Derivative-Free Optimization: A Review of Algorithms and Comparison of Software Implementations." *Journal of Global Optimization* 56(3): 1247-1293.

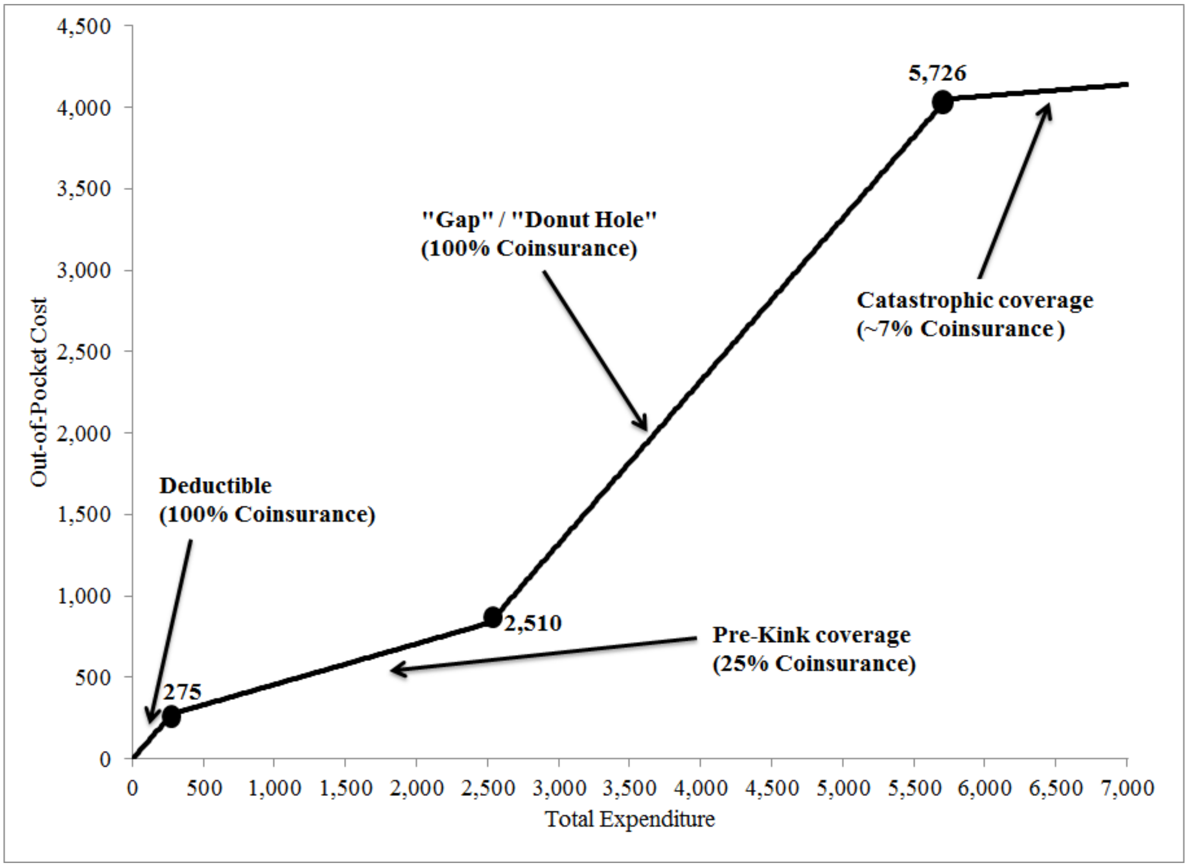
Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2(3): 180-212.

Vera-Hernandez, Marcos. 2003. "Structural Estimation of a Principal-Agent Model: Moral Hazard in Medical Insurance." *RAND Journal of Economics* 34(4): 670-693.

Yin, Wesley, Anirban Basu, James Zhang, Atonu 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 (3): 169-177.

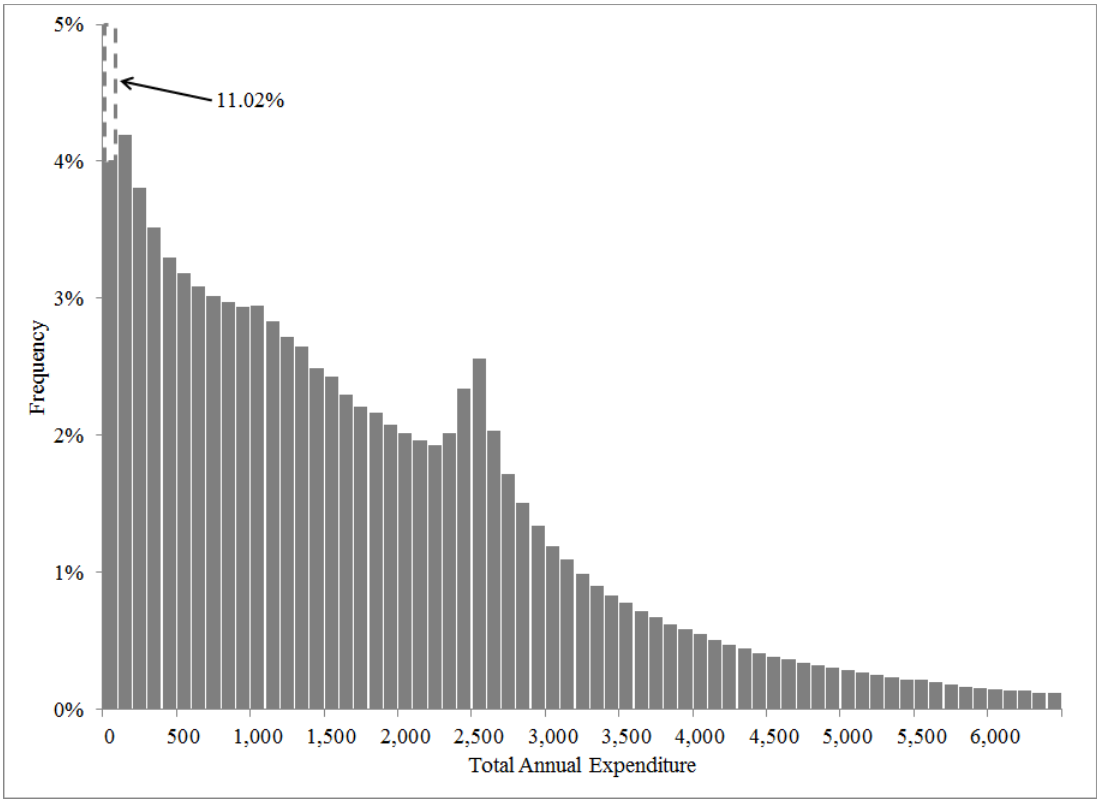
Zhang, Yuting, Katherine Baicker, and Joseph P. Newhouse. 2010. "Geographic Variation in the Quality of Prescribing." *New England Journal of Medicine* 363: 1985-1988.

Figure 1: Standard benefit design (in 2008)



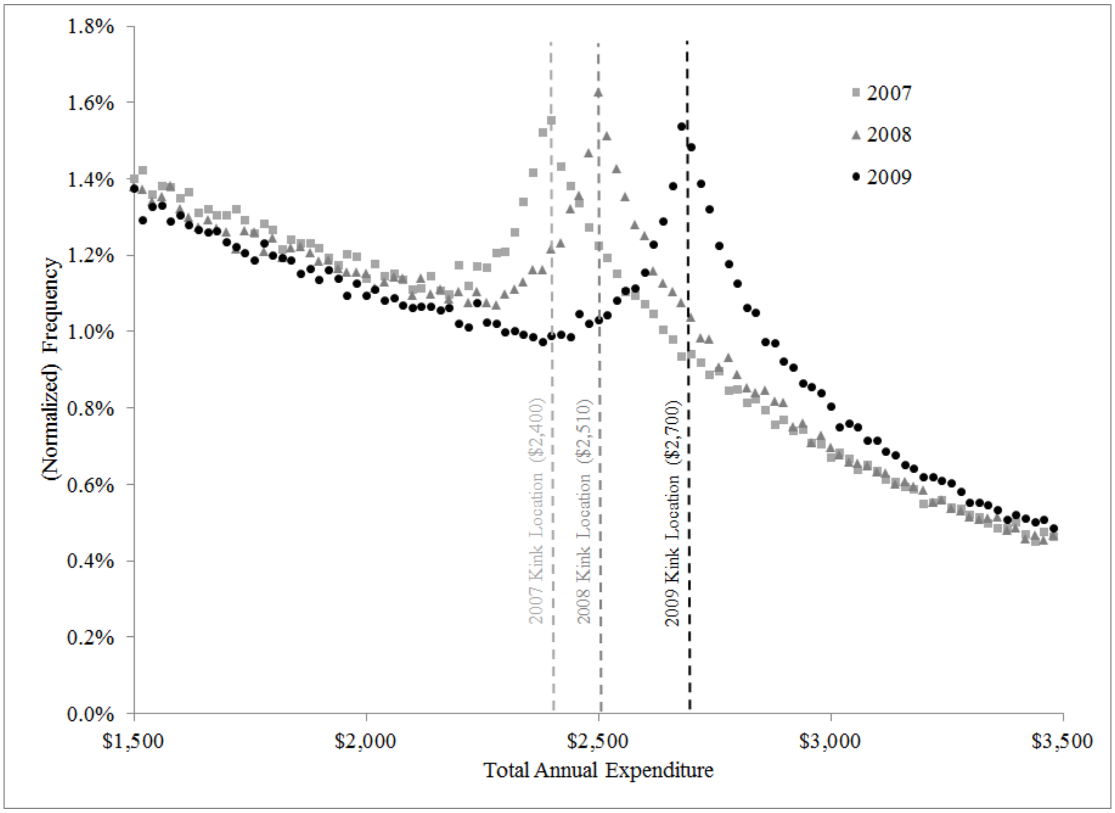
The figure shows the standard benefit design in 2008. "Pre-Kink coverage" refers to coverage prior to the Initial Coverage Limit (ICL) which is where there is a kink in the budget set and the gap, or donut hole, begins. As described in the text, the actual level at which the catastrophic coverage kicks 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)



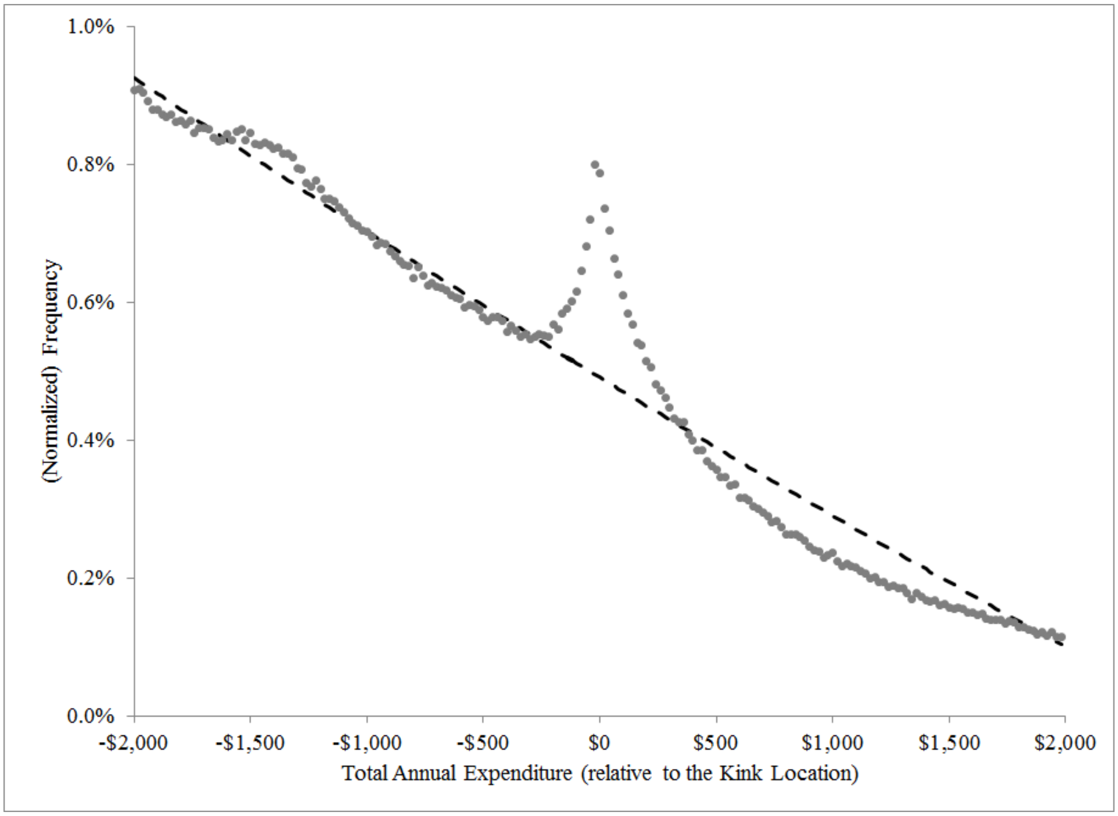
The figure displays the distribution of total annual prescription drug spending in 2008 for our baseline sample. Each bar represents the set of people that spent up to $100 above the value that is on the x-axis, so that the first bar represents individuals who spent less than $100 during the year, the 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



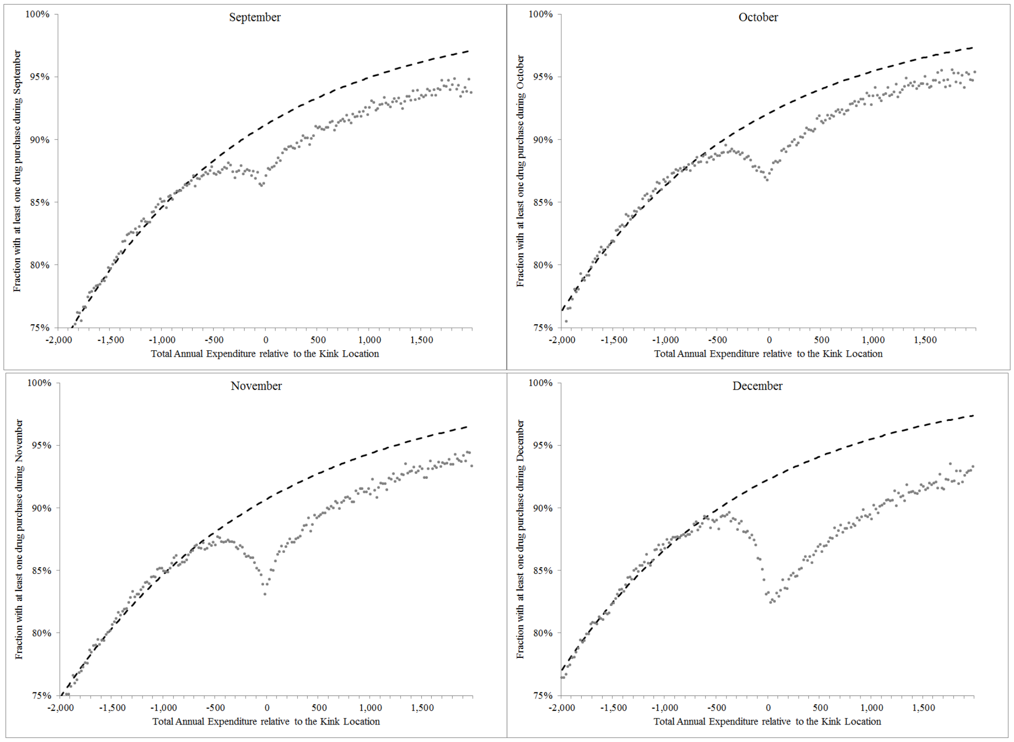
The figure displays the distribution of total annual prescription drug spending, separately by year, for individuals in our baseline sample whose annual spending in a given year was between $1,500 and $3,500 (N=1,332,733 overall; by year it is 447,006 (2007), 443,317 (2008), and 442,410 (2009)). Each 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: Timing of drug purchases



Each panel of the figure shows the fraction of individuals who have at least one drug purchase during the corresponding month (which appears in the panel title) 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. The dashed line in each panel is generated by regressing the logarithm of the share of individuals with no purchase (in the panel-specific month) 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 6: Model fit: spending distribution

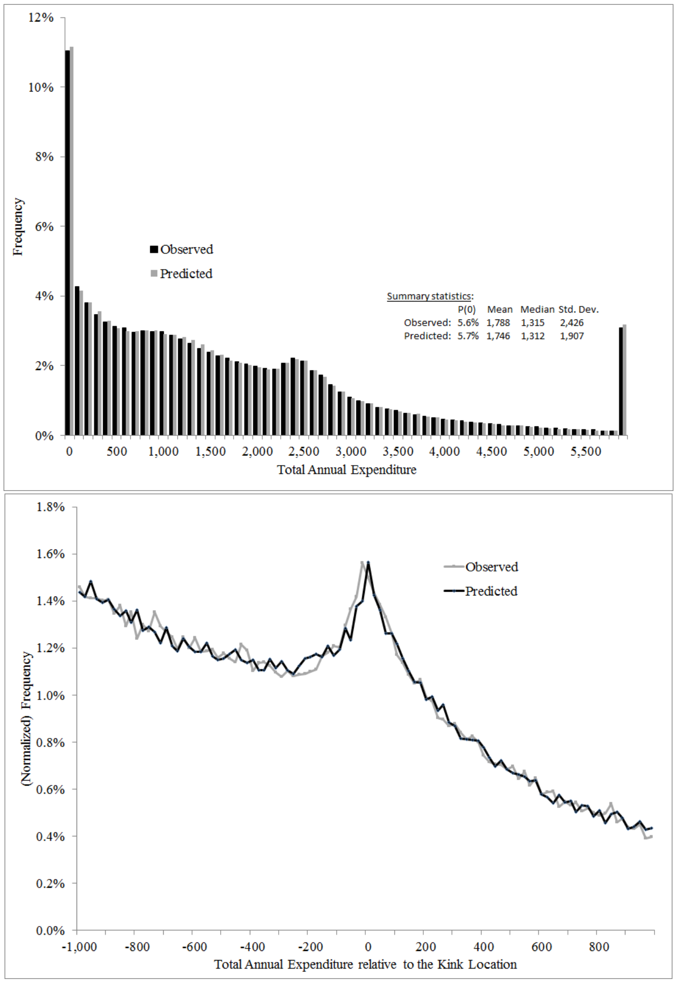


Figure shows the distribution of observed and predicted total annual drug spending. The top panel shows the results for the whole distribution, where each bar represents a $100 spending bin above the value on the x-axis (except for the last bar, which includes all spending above $5,900). The bottom panel "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. We note that the figure is based on the estimation sample rather than the baseline sample (see footnote 23), so the summary statistics do not perfectly match those presented in Table 1.

Figure 7: Model fit: claim timing

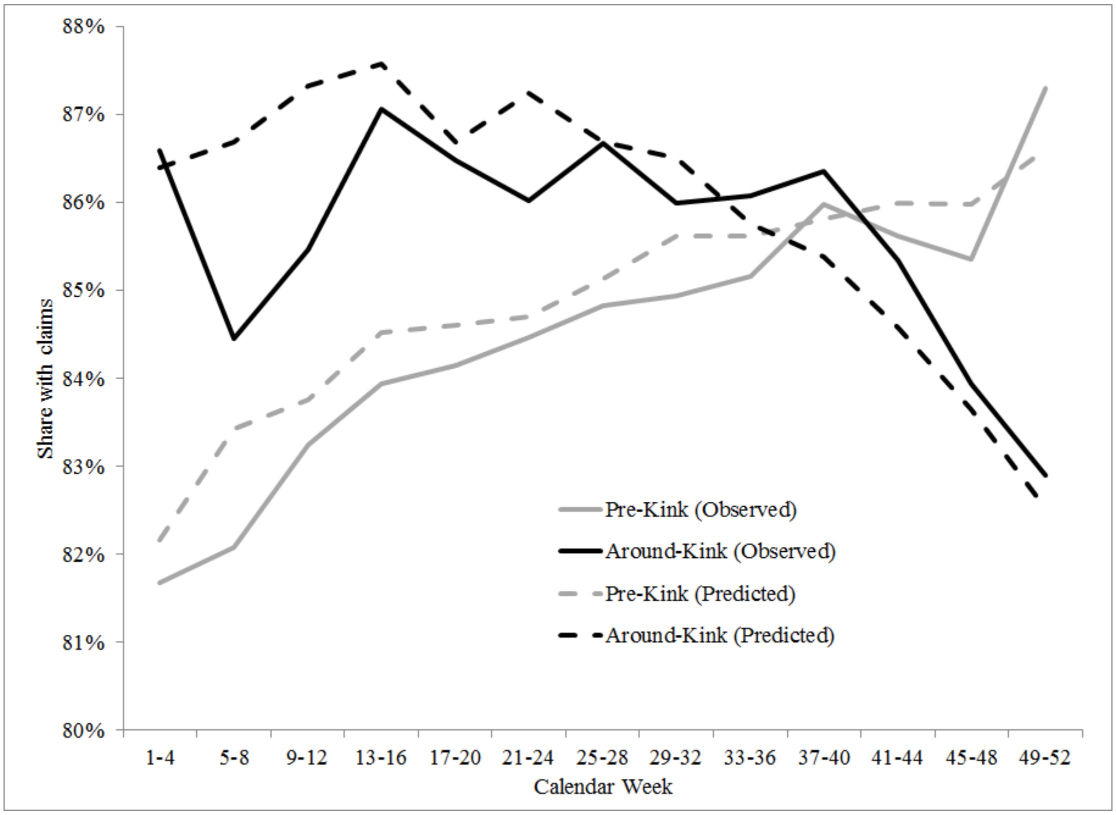


Figure shows the claim timing pattern of the observed and predicted spending. The x-axis represents calendar "months" (where a month is a group of four weeks, so we have 13 such "months" in a coverage year), and the y-axis reports the observed and predicted share of individuals who have at least one claim during that "month." The gray lines show this pattern for individuals who (by the end of the year) are well below the kink: their overall spending is between 800 and 500 dollars below the (year-specific) kink. These individuals tend to increase their monthly spending over the calendar year as a greater fraction of them hits the deductible and enter the pre-kink coverage arm. The black lines show this pattern for individual who end up around the kink (within 150 dollars of it, on both sides). These individuals decrease their propensity to claim toward the end of the year, as shown (in more details) in Figure 5.

Figure 8: Change in spending from “filling the gap”

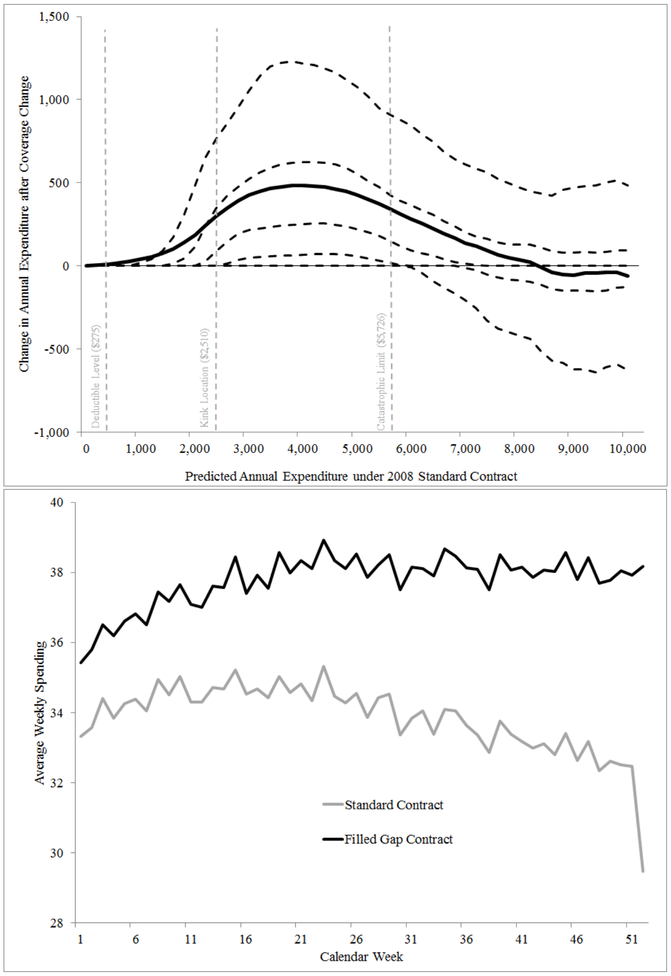
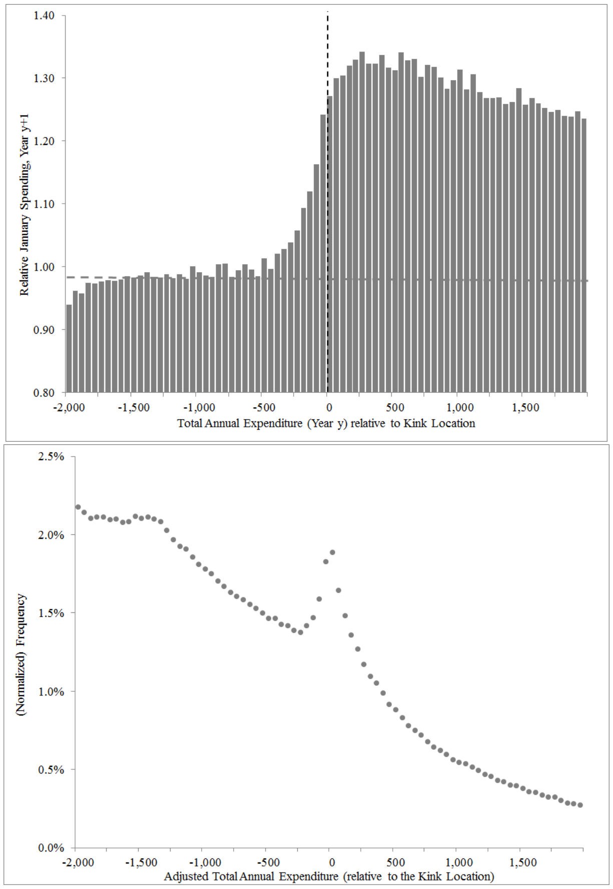


Figure shows the change in spending from “filling the gap” (i.e. providing 25% cost-sharing in the gap) for the 2008 standard benefit (which provides no coverage in the gap). In the top panel, the x-axis shows predicted spending under the 2008 standard benefit. The solid black line shows the mean change 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. In the bottom panel, we show the average predicted weekly spending, by calendar week, for the 2008 standard benefit (gray) and for the “filled gap” contract (black).

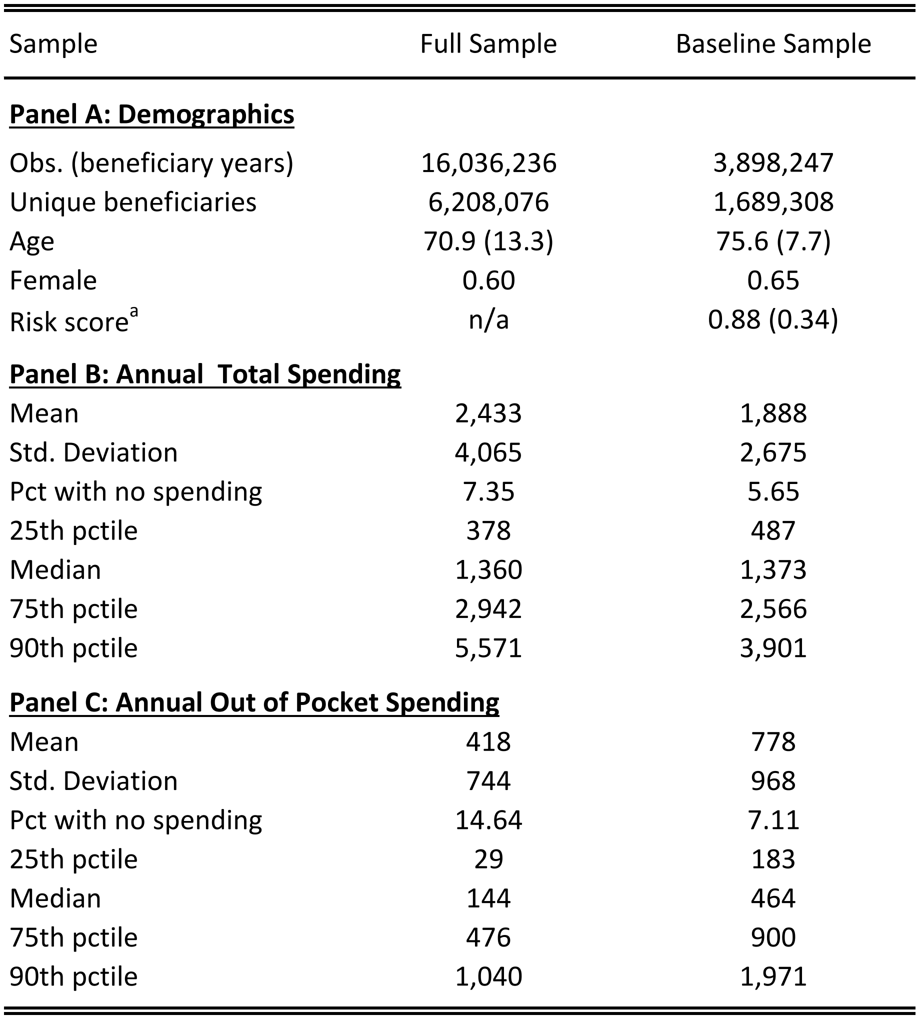
Figure 9: Cross-year substitution



The top panel shows the individual's relative January spending in year $y+1$ as a function of her total annual spending (relative to the kink location, which is normalized to 0) in the prior year (year y). "Relative" January spending in year $y+1$ is defined as the ratio of January spending in year $y+1$ to average monthly spending in March through June (of year $y+1$). 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 y spending bin, of the "relative January spending" measure. The dashed, horizontal "counterfactual" relative January spending is calculated as the average relative January spending for people -500 to -2000 below the kink in year y. 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 $y+1$ spending. N=1,534,952.

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 y) and, for each $50 bin of annual spending in year y, adding the average dollar difference between January spending in year $y+1$ and the average monthly spending in March to June of year $y+1$. "Adjusted" annual spending is reported relative to the kink location in year y (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=1,534,952).

Table 1: Summary statistics



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

Table shows summary statistics for the full 20% random sample of Medicare Part D beneficiaries (first column) and for our baseline sample (second column). 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. See text for more details on sample restrictions.

[a] Risk scores are predictions of Part D annual spending using CMS's 2012 RxHCC risk adjustment model (see text for details). They are normalized 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 use a different method and are only a function of a few crude demographics like gender.

Table 2: Cost-sharing features

	Deductible plans	No Ded. plans
Obs. (beneficiary years)	1,036,824	2,861,423
Avg. Deductible Amount	265.9	0
Avg. Deductible Coins. Rate	0.88	--
Avg kink location[a]	2,523.0	2,541.7
Avg. pre-kink Coins. Rate	0.26	0.37
Pct w/ Some Gap Coverage	0.01	0.17
Avg. Gap Coins. Rate (no gap Coverage)	0.88	0.98
Avg Gap Coins. Rate (some gap coverage)	0.71	0.77
Avg catastrophic limit (out of pocket)[a]	4,060.0	4,091.8
Catastrophic Coins. Rate	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: Heterogeneity in response across individuals

Population	Share of Sample	Share of Spending	Excess Mass	Excess January Spending
	(1)	(2)	(3)	(4)
All	100.0	100.0	0.291 (0.003)	1.238
Year[a]				
2006	n/a	n/a	0.088 (0.008)	n/a
2007	n/a	n/a	0.150 (0.008)	1.248
2008	n/a	n/a	0.213 (0.009)	1.228
2009	n/a	n/a	0.293 (0.010)	n/a
Gender				
Male	35.0	34.3	0.348 (0.005)	1.269
Female	65.0	65.7	0.262 (0.004)	1.221
Age group				
66	5.4	4.6	0.519 (0.016)	1.347
67	5.7	4.9	0.426 (0.015)	1.318
68-69	11.0	9.8	0.383 (0.009)	1.299
70-74	24.0	23.0	0.334 (0.006)	1.264
75-79	19.2	20.2	0.255 (0.006)	1.218
80-84	15.4	17.6	0.194 (0.006)	1.185
85+	14.8	17.8	0.136 (0.007)	1.149
Number of Hierarchical Condition Categories[b]				
0	15.5	6.6	0.837 (0.020)	1.259
1	8.9	5.0	0.494 (0.018)	1.260
2	14.9	10.7	0.191 (0.008)	1.227
3	17.6	16.2	0.197 (0.006)	1.247
4	15.0	16.9	0.236 (0.007)	1.240
5+	28.1	44.6	0.316 (0.005)	1.227
Risk Score Quartile				
1 (healthiest)	25.1	11.9	0.448 (0.011)	1.252
2	25.1	19.3	0.155 (0.005)	1.238
3	24.9	26.7	0.250 (0.005)	1.245
4 (least healthy)	25.0	42.2	0.346 (0.005)	1.221

Table reports excess mass at the kink and the excess January spending separately for different populations. Excess mass and excess January spending are calculated separately for each group. For the excess mass calculation for a given group, 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. Excess January spending is calculated for each sub-sample as the ratio of year $y+1$'s relative January spending (which is defined as January $y+1$ spending relative to the monthly average of year $y+1$'s March to June spending) for individuals between $0 and $500 of the kink in year y compared to this same calculation for those between -$500 and -$2000 of the kink in year $y+1$.

[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.

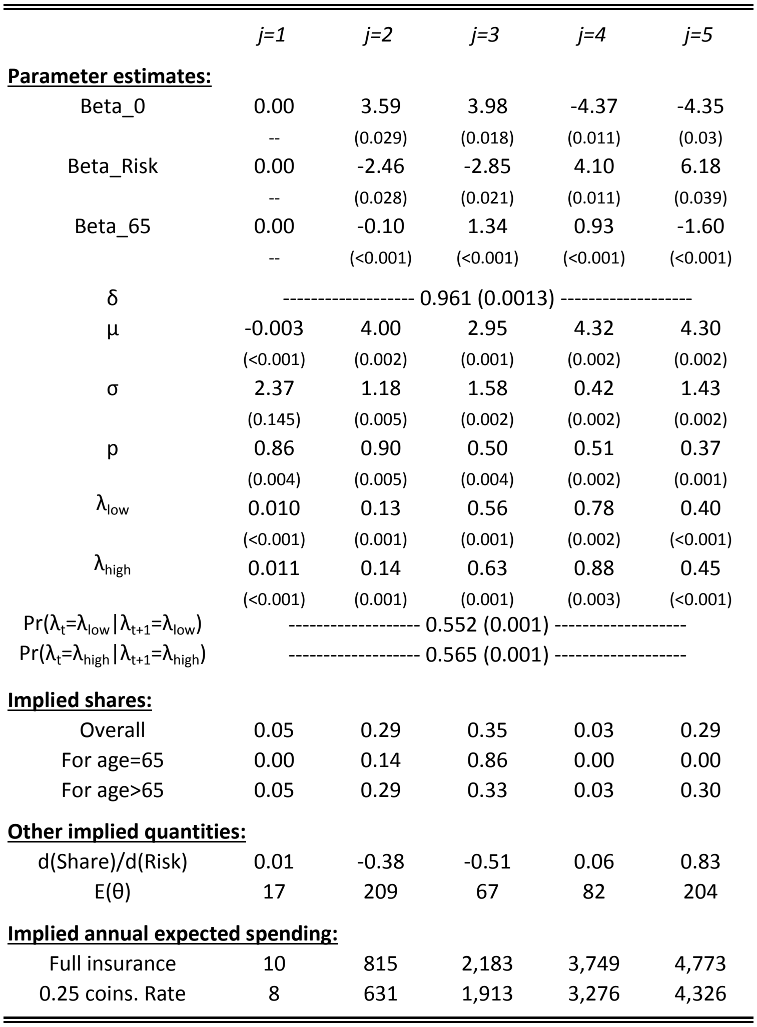
Table 4: Heterogeneity in response across drug types

Drug Type	Percent of purchases (1)	Percent of spending ($) (2)	Actual P(Dec. Purchase) (3)	Predicted P(Dec. Purchase) (4)	Percent decrease in purchase probability (5)	Excess January spending (6)
All	100.0	100.0	0.846	0.923	0.083 (0.001)	1.238
Chronic	69.6	77.1	0.762	0.867	0.121 (0.001)	1.251
Acute	30.4	22.9	0.532	0.590	0.099 (0.002)	1.206
Maintenance	85.1	90.1	0.807	0.899	0.103 (0.001)	1.142
Non-Maintenance	14.9	9.9	0.303	0.330	0.080 (0.004)	1.270
Brand	33.4	74.8	0.624	0.780	0.199 (0.001)	1.247
Generic	66.6	25.2	0.731	0.798	0.085 (0.001)	1.133
"Inappropriate" [a]	2.7	1.3	0.075	0.085	0.119 (0.009)	1.202

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 5 (N=2,589,420). The "actual" probability of a December purchase (column (3)) 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 (column (4)) 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 the dashed line in Figure 5) 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. Column (5) 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. In column (6) we report excess January spending (in year $y+1$) for each drug type. Excess January spending is calculated for each sub-sample as the ratio of year $y+1$'s relative January spending (which is defined as January $y+1$ spending relative to the monthly average of year $y+1$'s March to June spending) for individuals between $0 and $500 of the kink in year y compared to this same calculation for those between -$500 and -$2000 of the kink in year $y+1$.

[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 5: Parameter estimates



	j=1	*j=2*	*j=3*	*j=4*	*j=5*
Parameter estimates:					
Beta_0	0.00	3.59	3.98	-4.37	-4.35
	--	(0.029)	(0.018)	(0.011)	(0.03)
Beta_Risk	0.00	-2.46	-2.85	4.10	6.18
	--	(0.028)	(0.021)	(0.011)	(0.039)
Beta_65	0.00	-0.10	1.34	0.93	-1.60
	--	(<0.001)	(<0.001)	(<0.001)	(<0.001)
δ	------------------- 0.961 (0.0013) -------------------				
μ	-0.003	4.00	2.95	4.32	4.30
	(<0.001)	(0.002)	(0.001)	(0.002)	(0.002)
σ	2.37	1.18	1.58	0.42	1.43
	(0.145)	(0.005)	(0.002)	(0.002)	(0.002)
p	0.86	0.90	0.50	0.51	0.37
	(0.004)	(0.005)	(0.004)	(0.002)	(0.001)
λ_{low}	0.010	0.13	0.56	0.78	0.40
	(<0.001)	(0.001)	(0.001)	(0.002)	(<0.001)
λ_{high}	0.011	0.14	0.63	0.88	0.45
	(<0.001)	(0.001)	(0.001)	(0.003)	(<0.001)
$Pr(\lambda_t=\lambda_{low} \mid \lambda_{t+1}=\lambda_{low})$	------------------- 0.552 (0.001) -------------------				
$Pr(\lambda_t=\lambda_{high} \mid \lambda_{t+1}=\lambda_{high})$	------------------- 0.565 (0.001) -------------------				
Implied shares:					
Overall	0.05	0.29	0.35	0.03	0.29
For age=65	0.00	0.14	0.86	0.00	0.00
For age>65	0.05	0.29	0.33	0.03	0.30
Other implied quantities:					
d(Share)/d(Risk)	0.01	-0.38	-0.51	0.06	0.83
E(θ)	17	209	67	82	204
Implied annual expected spending:					
Full insurance	10	815	2,183	3,749	4,773
0.25 coins. Rate	8	631	1,913	3,276	4,326

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 spending depends on the arrival rate of drug events (λ), the distribution of event size (θ), as well as on the decision to claim, which is affected by the features of the contract and the parameter p.

Table 6: Spending effect from filling the gap

	Mean	Std. Dev.	25th pctile	Median	90th pctile	Mean OOP	Mean Insurer
Assign everyone to Standard 2008 contract:							
1 Baseline	1,760	1,924	402	1,413	3,632	809	951
2 "Filled" gap[a]	1,964	2,127	407	1,455	4,450	655	1,309
Assign everyone observed (chosen) contract:							
3 Baseline	1,768	1,909	499	1,342	3,675	796	973
4 "Filled" gap[a]	1,916	2,051	502	1,371	4,287	690	1,226
Assign everyone observed (chosen) contract, plus allow cross-year substitution:							
5 Baseline	1,770	1,912	501	1,337	3,749	807	963
6 "Filled" gap[a]	1,814	1,964	501	1,341	3,956	658	1,155

Table reports the predicted annual drug spending under various observed and counterfactual contracts. All columns report total annual drug spending except the rightmost two which separately report out-of-pocket and insurer spending. Rows 1 and 2 report predicted spending under the standard contract in 2008, which was illustrated in Figure 1, and counterfactual changes to it. Rows 3 and 4 report predicted spending for the observed contracts in our sample, and counterfactual changes to them. Rows 5 and 6 repeat the same exercise (as in rows 3 and 4), but use the extension of the model that accounts of cross-year substitution. For all of the simulations, we assume individuals are in the contract for a full 12 months. (Predicted mean spending for observed contracts - row 3 - is slightly higher than the estimate reported in the top panel of Figure 6 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.

Table 7: Robustness

	Total Spending			Insurer Spending		
	Baseline	"Filled" Gap	Change	Baseline	"Filled" Gap	Change
1 Baseline model	1,768	1,916	8.4%	973	1,226	26.0%
Number of types:						
2 Three types	1,778	1,878	5.6%	986	1,208	22.5%
3 Six types	1,760	1,937	10.1%	967	1,240	28.2%
Different sets of covariates:						
4 Remove all covariates	1,783	1,892	6.1%	976	1,209	23.9%
5 Add "no gap" covariate	1,758	1,946	10.7%	968	1,246	28.7%
Concave (risk averse) utility function:						
6 CARA = exp(-6.5)	1,737	1,873	7.8%	945	1,193	26.2%
7 CARA = exp(-9.5)	1,738	1,862	7.1%	940	1,183	25.9%
Distribution of ω:						
8 ω correlated with θ	1,781	1,918	7.7%	983	1,230	25.1%

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 3 and 4 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 covariates 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). Finally, row 8 reports an extension of the model that allows the moral hazard parameter p to vary with θ, as described in the main text.

NOT FOR PUBLICATION APPENDICES

Appendix A: 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.[38] Appendix Figure A3, however, shows no evidence of such missing mass around the deductible level for individuals in plans with the standard deductibles. 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.

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 decreasing 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: Estimation details

Simulation We estimate our model using simulated minimum distance. As described in Section 4.4:

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

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 $\exp(z_i\beta_m)/\left(\sum_{k=1}^{M}\exp(z_i\beta_k)\right)$. Then, starting from the first week of the year

[38] Analysis of the spending distribution around the catastrophic limit is noisy for two reasons. First, only 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.

($t = 51$) and going until the final week of the year ($t = 0$), we simulate a claim history.[39]

Cumulative spending begins with $x_{is,51} = 0$. The initial health state, λ_{ist}, is drawn from its type specific stationary distribution. Each week there is an event with probability λ_{ist}. 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

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

In this case, $x_{ist-1} = x_{ist} + \theta_{ist}$. Otherwise, $x_{ist-1} = x_{ist}$. Finally, λ_{ist-1} is drawn from $H_m(\cdot|\lambda_{ist})$.

We repeat this simulation until $t = 0$. We then use the simulated data to calculate the statistics described in Section 4.4. 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 (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 2013). Andreasen (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 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

$$v_{jm}(x, t, \lambda) = E_m \left[(1 - \lambda')\delta v_{jm}(x, t-1, \lambda') + \lambda' \left(\max \left\{ \begin{matrix} -c_j(\theta, x) + \delta v_{jm}(x + \theta, t-1, \lambda'), \\ -\omega + \delta v_{jm}(x, t-1, \lambda') \end{matrix} \right\} \right) \middle| \lambda \right],$$

[39] For 65 year olds 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.

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, ω, and λ'. Given that $v_{jm}(x,0,\lambda)=0$, we can compute an approximation to v_{jm} sequentially. First, we approximate $v_{jm}(x,1,\lambda)$. Then, we use that approximation to compute $v_{jm}(x,2,\lambda)$, 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 $v_{jm}(x,t-1,\lambda)$, say $\tilde{v}_{jm}(x,t-1,\lambda)$, we compute

$$v_{k,jm\lambda}=(1-\lambda_m)\delta\tilde{v}_{jm}(x_{k,j},t-1)+\lambda_m E_m\left[\max\left\{\begin{matrix}-c_j(\theta,x_{k,j})+\delta\tilde{v}_{jm}(x_{k,j}+\theta,t-1,\lambda'),\\ -\omega+\delta\tilde{v}_{jm}(x_{k,j},t-1,\lambda')\end{matrix}\right\}\middle|\lambda_m\right].$$

We then calculate $\tilde{v}_{jm}(x,t,\lambda)$ using linear interpolation between the $\{(x_{k,j},v_{k,jm\lambda})\}$ values.[40] 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 $v_{k,jm\lambda}$, we must integrate over θ, ω, and λ'. λ' is discrete, so integrating over its distribution is straightforward. For θ and ω, we must compute

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

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

$$E_m\left[\max\left\{\begin{matrix}-c_j(\theta,x_{k,j})+\delta\tilde{v}_{jm}(x_{k,j}+\theta,t-1,\lambda),\\ -\omega+\delta\tilde{v}_{jm}(x_{k,j},t-1,\lambda)\end{matrix}\right\}\right]=$$
$$=E_m\left[\begin{matrix}P\left(\frac{c_j(\theta,x_{k,j})-\delta\tilde{v}_{jm}(x_{k,j}+\theta,t-1,\lambda)+\delta\tilde{v}_{jm}(x_{k,j},t-1,\lambda)}{\theta}\le\frac{\omega}{\theta}\middle|\theta\right)\left(-c_j(\theta,x_{k,j})+\delta\tilde{v}_{jm}(x_{k,j}+\theta,t-1,\lambda)\right)+\\ +\left(\begin{matrix}P\left(\frac{c_j(\theta,x_{k,j})-\delta\tilde{v}_{jm}(x_{k,j}+\theta,t-1,\lambda)+\delta\tilde{v}_{jm}(x_{k,j},t-1,\lambda)}{\theta}>\frac{\omega}{\theta}\middle|\theta\right)\cdot\\ \cdot\left(E\left[-\omega\middle|\frac{c_j(\theta,x_{k,j})-\delta\tilde{v}_{jm}(x_{k,j}+\theta,t-1,\lambda)+\delta\tilde{v}_{jm}(x_{k,j},t-1,\lambda)}{\theta}>\frac{\omega}{\theta}\right]+\delta\tilde{v}_{jm}(x_{k,j},t-1,\lambda)\right)\end{matrix}\right)\end{matrix}\right],$$

where

$$P\left(C\le\frac{\omega}{\theta}\middle|\theta\right)=\begin{cases}0 & \text{if } C\le 0\\ p_m C & \text{if } C\in(0,1)\\ 1 & \text{if } C\ge 1\end{cases}$$

and

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

[40] 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.

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 C: More details about model extensions

In the main text we report results from various variants and extensions to the baseline model. Some of the variations, like changing the number of types, are mechanical. Others require some explanation. This section describes the two less trivial variations of the model, and how the value function computation is altered for each.

C.1 Allowing for risk aversion

As stated in the main text, we introduce constant absolute risk aversion while maintaining perfect intertemporal substitution by specifying recursive preferences as in Kreps and Porteus (1978) or Epstein and Zin (1989). Individual preferences over a stochastic sequence of flow utilities, $\{u_t\}$, are defined recursively as

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

where α is the coefficient of absolute risk aversion. Using the form of u_t in our model, this becomes

$$V_t = -\ell_t d_t c_j(\theta_t, x_t) + \ell_t(1-d_t)(-\omega_t) + \delta\frac{-1}{\alpha}\begin{Bmatrix} d_t\ell_t \log E[\exp(-\alpha V_{t-1})|x_{t-1} = x_t + \theta_t, \lambda_t = \lambda]+ \\ +(1-d_t\ell_t)\log E[\exp(-\alpha V_{t-1})|x_{t-1} = x_t, \lambda_t = \lambda] \end{Bmatrix},$$

where $\ell_t = 1$ if there was a prescription to potentially fill and $d_t = 1$ if the prescription was filled. The expected value function is

$$\tilde{v}(x,t,\lambda) = \sum_{\lambda'} P(\lambda'|\lambda)\left\{\begin{array}{c} \lambda' E\left[\exp\left(-\alpha\max\begin{Bmatrix} -c_j(\theta,x) + \delta\frac{-1}{\alpha}\log\tilde{v}(x+\theta,t-1,\lambda'), \\ -\omega + \delta\frac{-1}{\alpha}\log\tilde{v}(x,\lambda',t-1)\end{Bmatrix}\right)\right] \\ +(1-\lambda')\tilde{v}(x,t-1,\lambda'^{\delta}) \end{array}\right\}.$$

Let

$$v(x,t,\lambda) = \frac{-1}{\alpha}\log\tilde{v}(x,t,\lambda).$$

The Bellman equation for v is then

$$v(x,t,\lambda) = \frac{-1}{\alpha}\log\left(\sum_{\lambda'} P(\lambda'|\lambda)\left\{\begin{array}{c} \lambda' E\left[\exp\left(-\alpha\max\begin{Bmatrix} -c_j(\theta,x) + \delta v(x+\theta,t-1,\lambda'), \\ -\omega + \delta v(x,t-1,\lambda')\end{Bmatrix}\right)\right] \\ +(1-\lambda')\exp(-\alpha\delta v(x,t-1,\lambda')) \end{array}\right\}\right)$$

The expectation of the maximum is calculated in a similar way as in the risk neutral case.

C.2 Allowing for the delay of purchasing to subsequent year

As described in the main text, we assume that each prescription must be filled either immediately, at the start of next year, or never. A potential prescription comes with a monetary cost θ and a utility flow cost of not filling ω. If a potential prescription is not filled, then each period θ depreciates at rate $\delta\delta_h$ and ω depreciates at rate δ_h. Unfilled prescriptions may be filled at the start of the next year at a (known) expected price q_i. We assume that q_i is known and taken as given. To calculate it, we calculate $E[p|\text{risk score}, \text{plan}]$ and assume that people use their current year risk score and plan to predict next year's end-of-year price. We compute $q_i = E[p|\text{risk score}, \text{plan}]$ by dividing risk score into 3 bins (lowest third, middle third, and highest third) and taking the average observed end-of-year price in each plan and bin.

With these assumptions, the dynamic optimization is different for each plan and risk score bin, so we subscript the value function by i to capture the idea that it varies with q_i, which as described varies by plan and risk score tercile. Then, the value functions can be written as

$$v_i(x,t,\lambda) = \int \left[\begin{array}{c} (1-\lambda')\delta v_i(x,t-1,\lambda')+ \\ +\lambda' \int \max \left\{ \begin{array}{c} -c_j(\theta,x) + \delta v_i(x+\theta,t-1,\lambda'), \\ -\omega\frac{1-(\delta\delta_h)^t}{1-\delta\delta_h} - \delta^t\delta_h^t q_i\theta + \delta_i v_i(x,t-1,\lambda') \\ -\frac{\omega}{1-\delta\delta_h} + \delta v_i(x,t-1,\lambda') \end{array} \right\} dG(\theta,\omega) \end{array} \right] dH(\lambda'|\lambda),$$

To calculate the value function we must compute,

$$\int_\Theta \int_\Omega \max \left\{ \begin{array}{c} -c_j(\theta,x) + \delta v_i(x+\theta,t-1,\lambda'), \\ -\omega\frac{1-(\delta\delta_h)^t}{1-\delta\delta_h} - \delta^t\delta_h^t q_i\theta + \delta v_i(x,t-1,\lambda') \\ -\frac{\omega}{1-\delta\delta_h} + \delta v_i(x,t-1,\lambda') \end{array} \right\} dG(\omega|\theta)dG(\theta).$$

We calculate the inner integral analytically using the assumption that $\frac{\omega}{\theta} \sim U(0,1)$, and calculate the outer integral using quadrature. The inner integral can be written as

$$B = \theta \int_0^1 \max \left\{ \begin{array}{c} \frac{-c_j(\theta,x)+\delta v_i(x+\theta,t-1,\lambda')}{\theta}, \\ -\frac{r}{1-\delta\delta_h} + \frac{\delta v_i(x,t-1,\lambda')}{\theta}, \\ -r\frac{1-(\delta\delta_h)^t}{1-\delta\delta_h} - (\delta\delta_h)^t q_i + \frac{\delta v_i(x,t-1,\lambda')}{\theta} \end{array} \right\} dr.$$

The values of r where we switch from one of the three terms in the max to another are

$$\begin{aligned} r_1 &= \left(\frac{\delta_h}{\delta_h}\right)^t q_i(1-\delta\delta_h) \\ r_2 &= \frac{1-\delta\delta_h}{\theta}\left(c_j(\theta,x) - \delta v_i(x+\theta,,t-1,\lambda') + \delta v_i(x,t-1,\lambda')\right) \\ r_3 &= \frac{1-\delta\delta_h}{1-(\delta\delta_h)^t}\frac{1}{\theta}\left(c_j(\theta,x) - \delta v_i(x+\theta,t-1,\lambda') + \delta v_i(x,t-1,\lambda') - (\delta\delta_h)^t q_i\right) \end{aligned}$$

If $0 \leq r_1 \leq r_2 \leq r_3 \leq 1$, then our expression for the inner integral becomes

$$\begin{aligned} B &= \theta \left(\begin{array}{l} \int_0^{r_1} -\frac{r}{1-\delta\delta_h} + \frac{\delta v_i(x,t-1,\lambda')}{\theta} dr + \\ + \int_{r_1}^{r_3} -r\frac{1-(\delta\delta_h)^t}{1-\delta\delta_h} - (\delta\delta_h)^t q_i + \frac{\delta v_i(x,t-1,\lambda')}{\theta} dr + \\ + \int_{r_3}^{1} -\frac{c_j(\theta,x)+\delta v_i(x+\theta,,t-1,\lambda')}{\theta} dr \end{array} \right) \\ &= \left(\begin{array}{l} -\theta\frac{r_1^2/2}{1-\delta\delta_h} + r_1 \delta v_i(x,t-1,\lambda') + \\ -\frac{1}{2}(r_3^2 - r_1^2)\frac{1-(\delta\delta_h)^t}{1-\delta\delta_h} + \left[-(\delta\delta_h)^t q_i \theta + \delta v_i(x,t-1,\lambda')\right](r_3 - r_1) + \\ +(1-r_3)\left[-c_j(\theta,x) + \delta v_i(x+\theta,,t-1,\lambda')\right] \end{array} \right). \end{aligned}$$

It will always be true that $0 \leq r_1 \leq 1$. However, the rest of these inequalities need not hold. If $0 \leq r_2 \leq r_1 \leq r_3$, then the integral is

$$\begin{aligned} B &= \theta \left(\begin{array}{l} \int_0^{r_2} -\frac{r}{1-\delta\delta_h} + \frac{\delta v_i(x,t-1,\lambda')}{\theta} dr + \\ + \int_{r_2}^{1} -\frac{c_j(\theta,x)+\delta v_i(x+\theta,,t-1,\lambda')}{\theta} dr \end{array} \right) \\ &= \left(\begin{array}{l} -\theta\frac{r_2^2/2}{1-\delta\delta_h} + r_2 \delta v_i(x,t-1,\lambda') + \\ +(1-r_2)\left[-c_j(\theta,x) + \delta v_i(x+\theta,,t-1,\lambda')\right] \end{array} \right). \end{aligned}$$

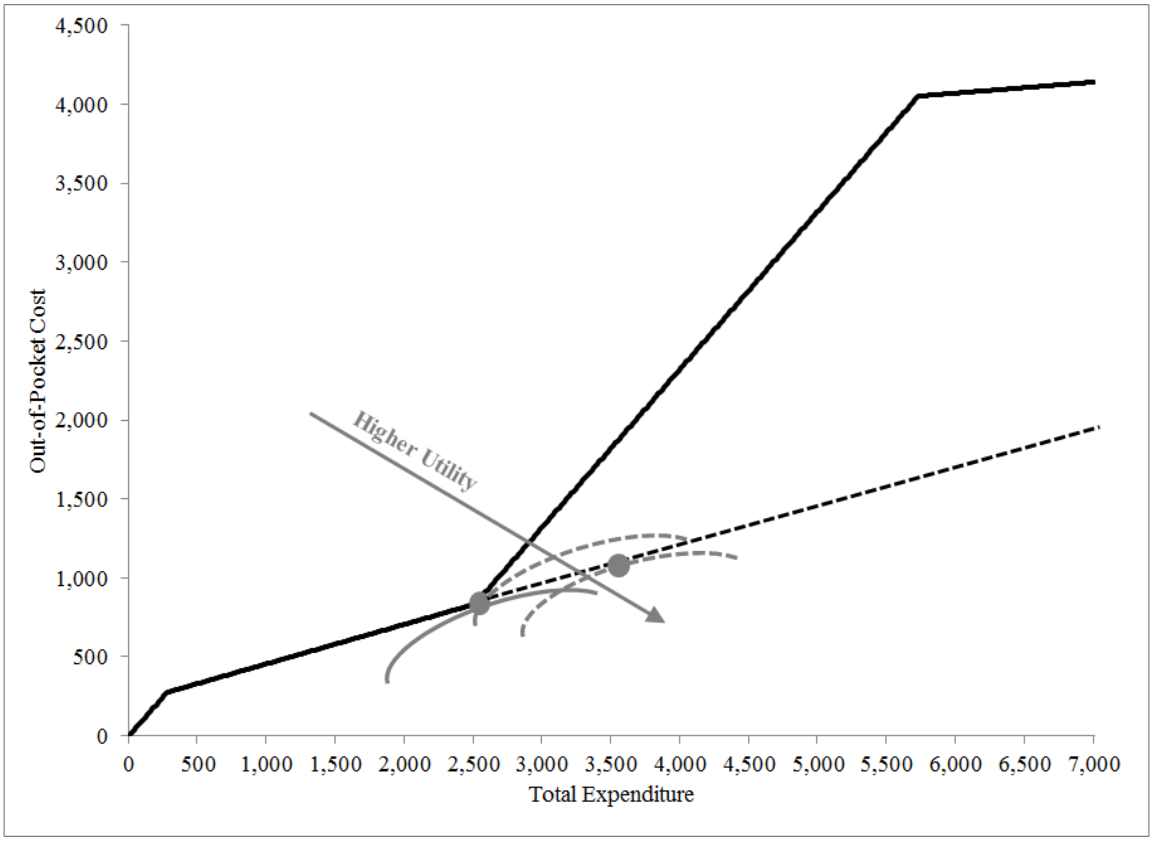
Other cases are treated similarly.

Additional references mentioned only in the appendix

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

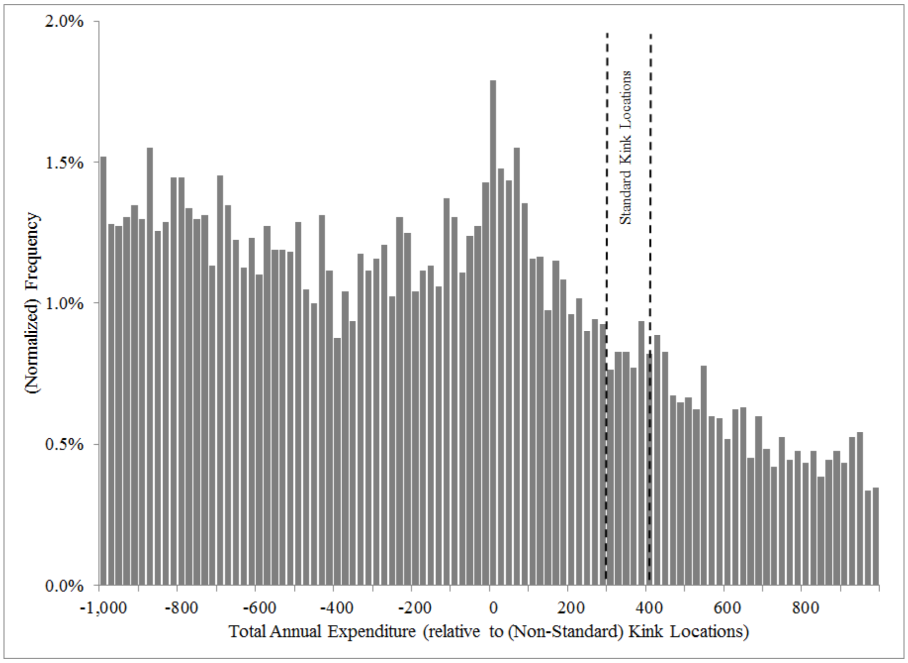
Lo Sasso, Anthony, Lorens Helmchen and Robert Kaestner. 2010. "The Effects of Consumer Directed Health Plans on Health Care Spending." *Journal of Risk and Insurance* 77(1): 85-103.

Appendix Figure A1: Rationale for bunching



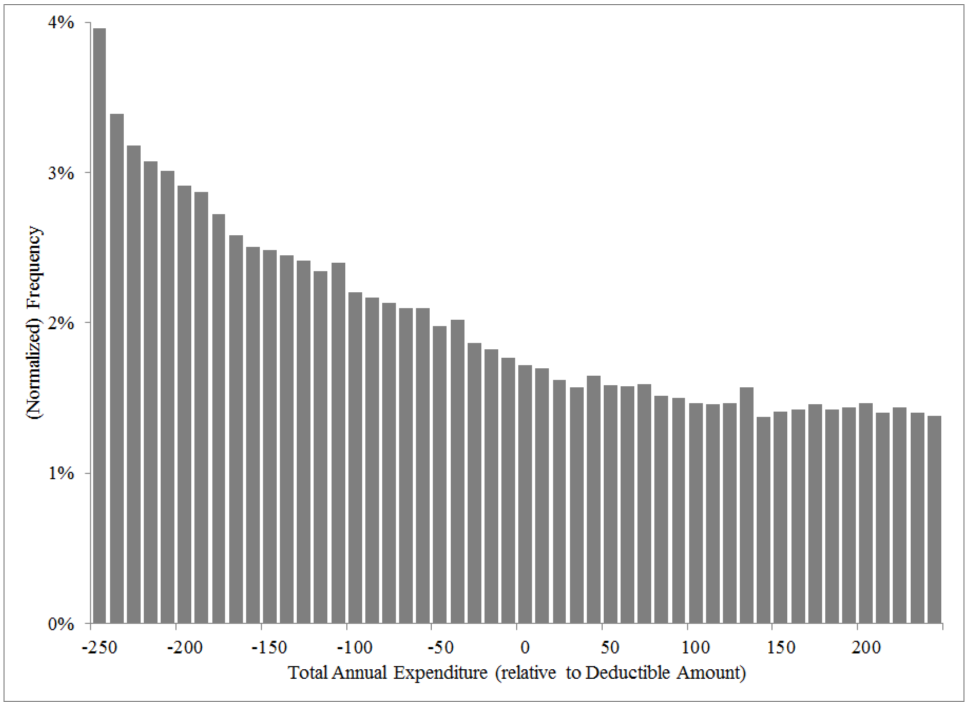
This figure illustrates graphically the theoretical prediction that individuals will bunch at the convex kink point in their budget set. The solid line illustrates the budget set of the same standard benefit design as in Figure 1; the standard budget set has a kink (price increase) at $2,510 in total spending. By contrast, the dashed line considers an alternative budget set with a linear budget (above the deductible) at the co-insurance arm's cost sharing rate. The solid and dashed indifference curves represent two individuals with different healthcare needs who would have different total drug spending under the linear contract. The (healthier) 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 (sicker) 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.

Appendix Figure A2: Distribution of spending for those individuals with non-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. 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. 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. The figure shows 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.

Appendix 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.

Appendix Figure A4: Excess mass by plan

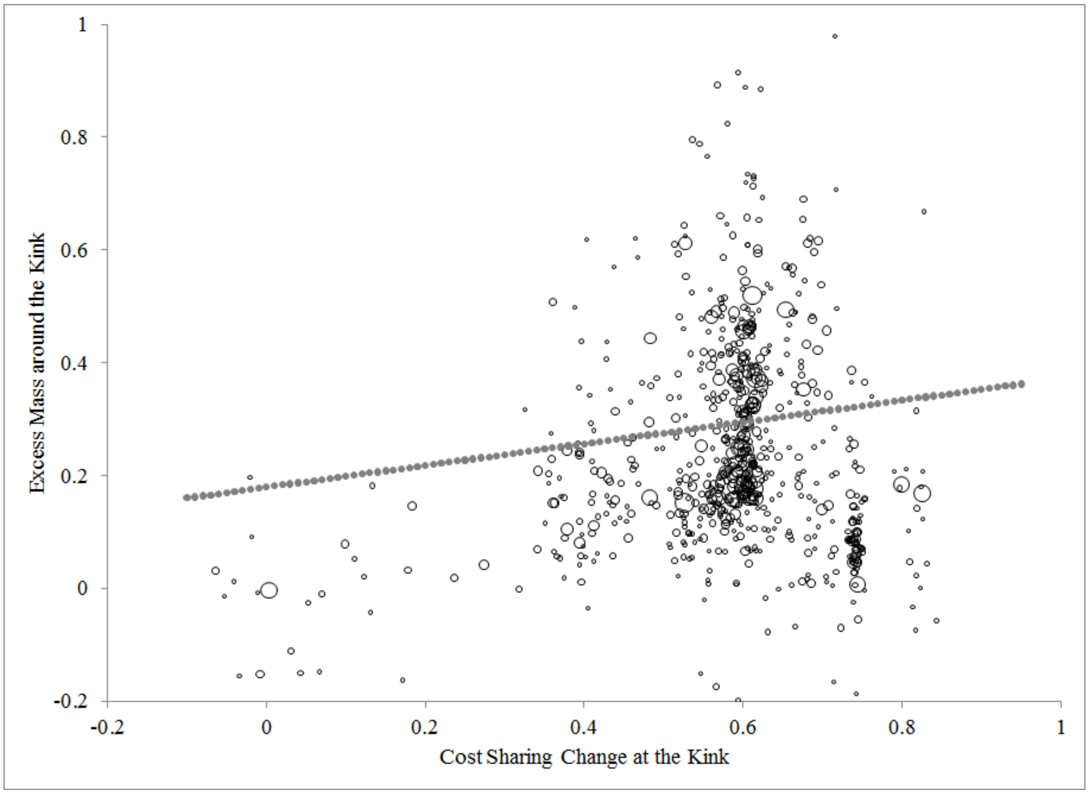
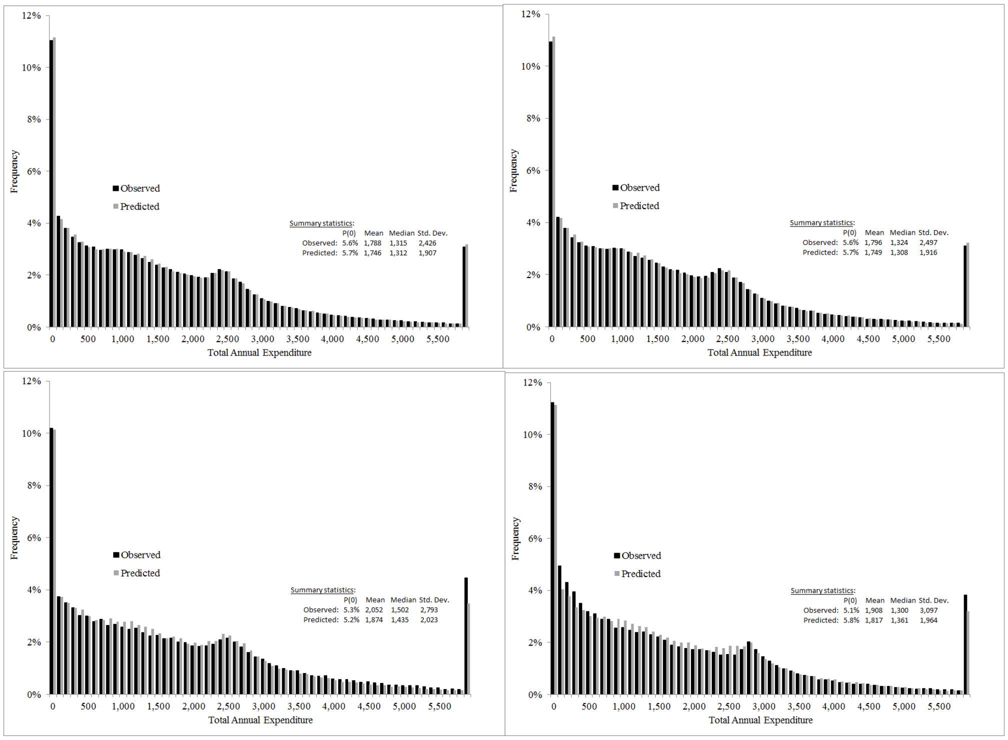


Figure graphs the excess mass in different plans against the size of the kink (i.e. the size of the price increase faced by the consumer as she moves into the gap). The size of the circles is proportional to the number of beneficiaries in the plan. Analysis is limited to the approximately 80% of our 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 for Figure 4. The dashed line in the figure represents the enrollee-weighted regression line of the relationship between excess mass and kink size; the slope of the line is 0.19 (standard error = 0.08) and the regression has an R-square of 0.011. N =1,985,676.

Appendix Figure A5: Out-of-sample fit: spending distribution



Summary statistics:

	P(0)	Mean	Median	Std. Dev.
Observed:	5.6%	1,788	1,315	2,426
Predicted:	5.7%	1,746	1,312	1,907

Summary statistics:

	P(0)	Mean	Median	Std. Dev.
Observed:	5.6%	1,796	1,324	2,497
Predicted:	5.7%	1,749	1,308	1,916

Summary statistics:

	P(0)	Mean	Median	Std. Dev.
Observed:	5.3%	2,052	1,502	2,793
Predicted:	5.2%	1,874	1,435	2,023

Summary statistics:

	P(0)	Mean	Median	Std. Dev.
Observed:	5.1%	1,908	1,300	3,097
Predicted:	5.8%	1,817	1,361	1,964

Figure presents the out-of-sample fit of the model. Top left panel replicates the in-sample fit of the model as in the top panel of Figure 6 in the main text. Recall (see footnote 23) that for estimation we limit the baseline sample to the 500 most common plans, and then only use a 10% random subsample to reduce the computational time. The top right panel presents the fit of the model predictions against a different 10% random subsample. The bottom left presents the model prediction for other plans (those that are not the 500 most common), taking these plans' features into account when generating predictions. Finally, the bottom right panel presents the model prediction for 2010 spending (recall that our baseline data covers 2007-2009 only). Here the fit is not as striking, presumably due to macroeconomic changes (e.g., in drug prices) that change over time. Still, the model's prediction change (relative to the 2009 predictions) in the right direction.

Appendix Figure A6: Out-of-sample fit: bunching

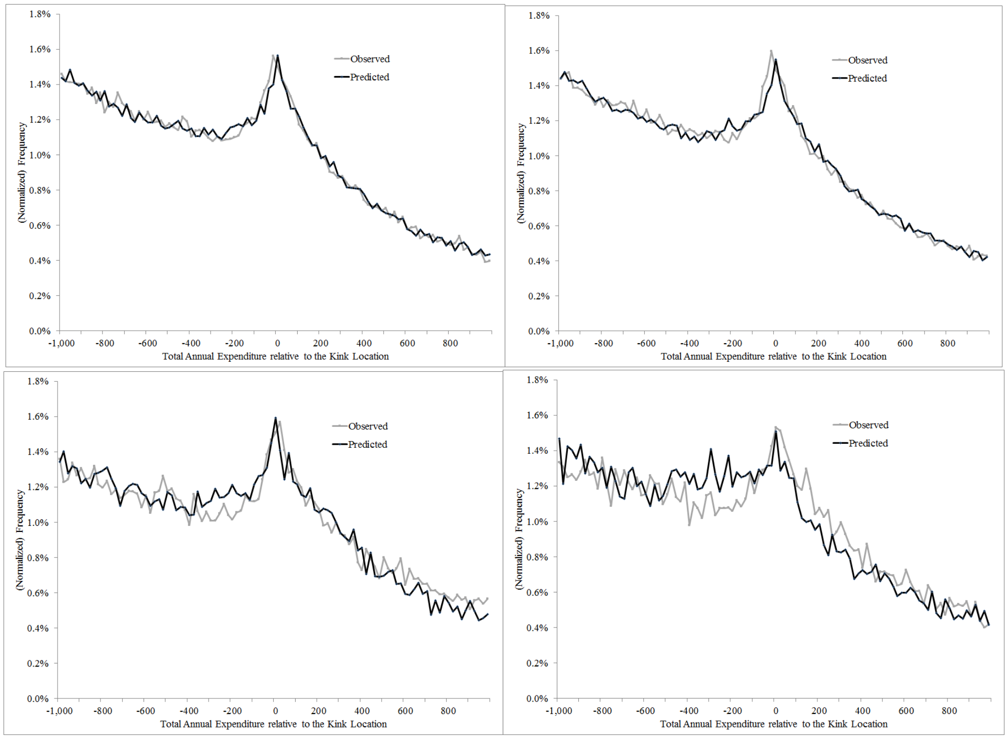


Figure presents the out-of-sample fit of the model. Top left panel replicates the in-sample fit of the model as in the bottom panel of Figure 6 in the main text. Recall (see footnote 23) that for estimation we limit the baseline sample to the 500 most common plans, and then only use a 10% random subsample to reduce the computational time. The top right panel presents the fit of the model predictions against a different 10% random subsample. The bottom left presents the model prediction for other plans (those that are not the 500 most common), taking these plans' features into account when generating predictions. Finally, the bottom right panel presents the model prediction for 2010 spending (recall that our baseline data covers 2007-2009 only). Here the fit is not as striking, presumably due to macroeconomic changes (e.g., in drug prices) that change over time. Still, the model's prediction change (relative to the 2009 predictions) in the right direction.

Appendix Figure A7: Out-of-sample fit: claim timing

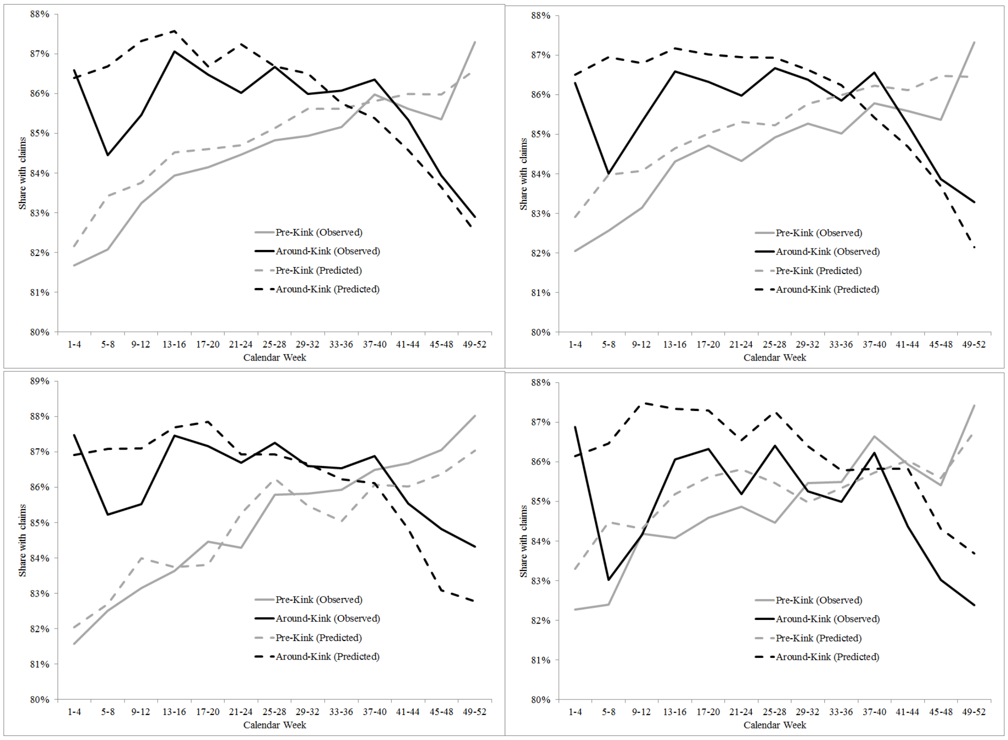


Figure presents the out-of-sample fit of the model. Top left panel replicates the in-sample fit of the model as in Figure 7 of the main text. Recall (see footnote 23) that for estimation we limit the baseline sample to the 500 most common plans, and then only use a 10% random subsample to reduce the computational time. The top right panel presents the fit of the model predictions against a different 10% random subsample. The bottom left presents the model prediction for other plans (those that are not the 500 most common), taking these plans' features into account when generating predictions. Finally, the bottom right panel presents the model prediction for 2010 spending (recall that our baseline data covers 2007-2009 only). Here the fit is not as striking, presumably due to macroeconomic changes (e.g., in drug prices) that change over time. Still, the model's prediction change (relative to the 2009 predictions) in the right direction.

Appendix Figure A8: Relative January spending by plan type

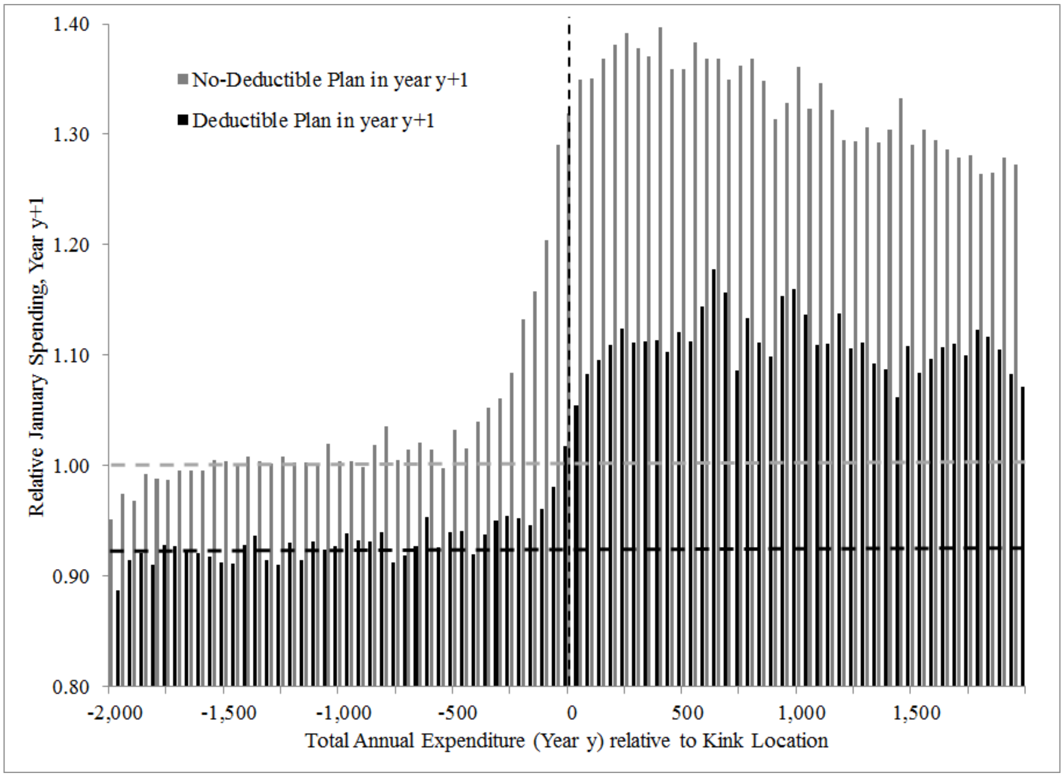


Figure replicates the top panel of Figure 9, but does it separately for beneficiaries who are enrolled in a deductible plan (black bars, $N = 305,437$) and no-deductible plan (gray bars, $N = 1,148,102$) in year $y+1$. Figure shows the individual's relative January spending in year $y+1$ as a function of her total annual spending (relative to the kink location, which is normalized to 0) in the prior year (year y). "Relative" January spending in year $y+1$ is defined as the ratio of January spending in year $y+1$ to average monthly spending in March through June (of year $y+1$). 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 "relative January spending" measure. The dashed, horizontal "counterfactual" relative January spending is calculated as the average relative January spending for people -500 to -2000 below the kink in year y.

Appendix Table A1: Implied price elasticities

(Uniform) Price Reduction[a]	Average Annual Spending	Implied "Elasticity"[b]
0% (Baseline)	1,760	
1.0%	1,769	-0.54
2.5%	1,776	-0.38
3.0%	1,779	-0.36
3.5%	1,781	-0.35
5.0%	1,789	-0.33
7.5%	1,801	-0.31
10.0%	1,813	-0.30
15.0%	1,837	-0.29
25.0%	1,887	-0.29
50.0%	2,018	-0.29
75.0%	2,163	-0.31

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).

Appendix Table A2: Parameter estimates from the extension of the model

	j=1	*j=2*	*j=3*	*j=4*	*j=5*
Parameter estimates:					
Beta_0	0.00	3.45	3.94	-4.51	-4.43
	--	(0.0006)	(0.0003)	(0.0004)	(0.0003)
Beta_Risk	0.00	-2.49	-2.85	4.07	6.06
	--	(0.0006)	(0.0005)	(0.0003)	(0.0003)
Beta_65	0.00	-0.10	1.19	1.08	-1.52
	--	(0.0009)	(0.0001)	(0.0003)	(0.0001)
δ	------------------- 0.661 (0.003) -------------------				
δ_ω (=δ_θ)	------------------- 0.612 (0.003) -------------------				
μ	-0.04	3.98	3.00	4.30	4.25
	(0.0001)	(0.0044)	(0.0001)	(0.0045)	(0.0046)
σ	2.46	1.29	1.64	0.28	1.46
	(0.0001)	(0.0040)	(0.0040)	(0.0025)	(0.0060)
p	0.87	0.94	0.58	0.59	0.29
	(0.0001)	(0.0007)	(0.0050)	(0.0071)	(0.0013)
λ_{low}	0.005	0.09	0.44	0.64	0.31
	(0.0001)	(0.0004)	(0.0017)	(0.0035)	(0.0012)
λ_{high}	0.006	0.12	0.58	0.84	0.41
	(<0.0001)	(0.0001)	(0.0018)	(0.0049)	(0.0009)
$Pr(\lambda_t=\lambda_{low} \mid \lambda_{t+1}=\lambda_{low})$	------------------- 0.549 (0.002) -------------------				
$Pr(\lambda_t=\lambda_{high} \mid \lambda_{t+1}=\lambda_{high})$	------------------- 0.566 (0.002) -------------------				
Implied shares:					
Overall	0.06	0.27	0.37	0.03	0.28
For age=65	0.01	0.14	0.85	0.00	0.00
For age>65	0.06	0.27	0.35	0.03	0.29
Other implied quantities:					
d(Share)/d(Risk)	0.01	-0.35	-0.53	0.06	0.80
E(θ)	20	123	76	77	204
Implied annual expected spending:					
Full insurance	6	753	2,280	3,338	4,295
0.25 coins. Rate	5	577	1,951	2,845	3,988

Top panel reports parameter estimates, with standard errors in parentheses, from the extension of the model that allows for cross-year substitution. 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 spending depends on the arrival rate of drug events (λ), the distribution of event size (θ), as well as on the decision to claim, which is affected by the features of the contract and the parameter p.