Dynamic Purchasing Behavior in Healthcare Consumption

Lin Xu*

See Latest Draft.

This Draft November 19, 2017

Abstract

Existing empirical evidence from Medicare Part D documents significantly lower spending and suboptimal behavior at the "donut hole", a region of low insurance coverage sandwiched between two regions of higher coverage. Current efforts to explain the drop in drug adherence rely heavily on time-discounting models with strong assumptions that recover far lower rates of discounting than the broader literature would predict. This paper first develops an alternative simple heuristic model with intuitions on how enrollees ought to behave by formulating a perceived marginal out-of-pocket price based on objective probabilities that updates as enrollees accumulate spending through their plans and the year. This approach predicts that with continuously updating expectations, a beneficiary's perceived marginal price and spending should be smooth and occur far in advance of region boundaries or "kinks" early in the year. Further, the paper measures actual behavior throughout the entire insurance schedule using dynamic panel regressions that control for individual heterogeneity. Overall, prescription claim frequencies do respond with significant foresight to anticipated changes in coverage generosity, but they also exhibit a drop when entering the coverage gap both early and late in the year. The magnitudes of these changes are ultimately small, reaffirming that many prescription purchases are highly inelastic.

*Please do not cite or circulate without permission. E-mail: lx82@cornell.edu. I am grateful to my graduate committee – Ted O'Donoghue, Levon Barseghyan, Colleen Carey, and Francesca Molinari – for their direction and support. Thanks to Gregory Besharov, Douglas Miller, Daniel Reeves, and numerous seminar participants at Cornell. For more information about my research and teaching, please visit my website: linxu.info.

1 Introduction

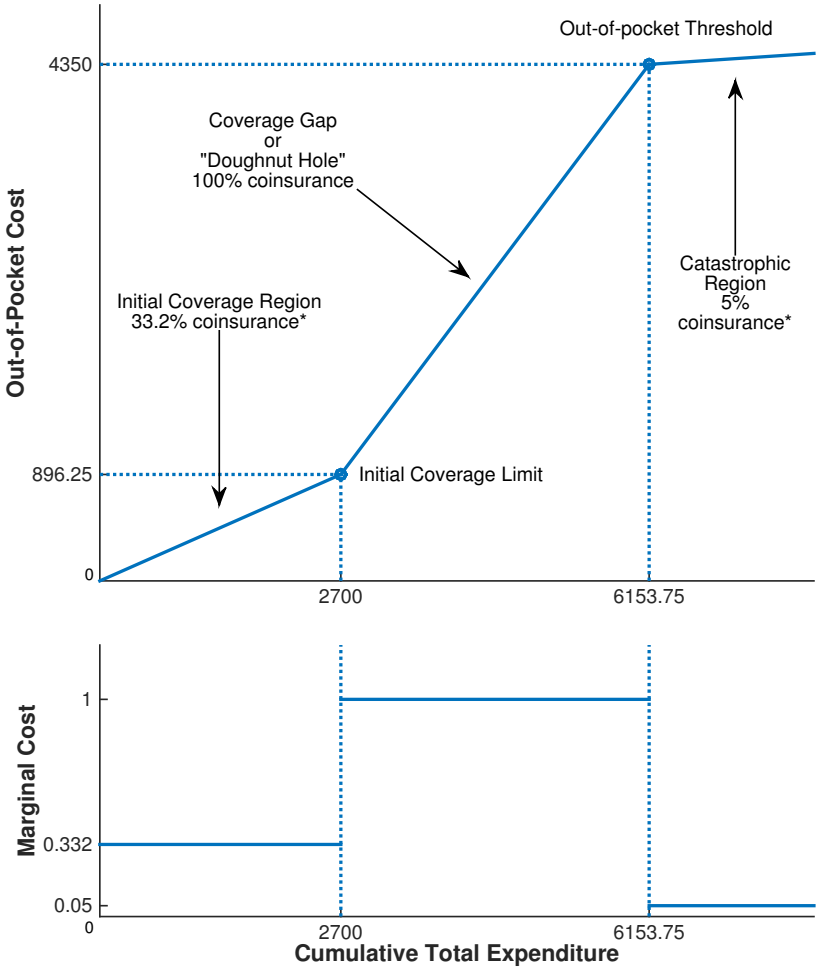
With the rapid growth of the US health care sector and the accompanying financial burden it poses, government and insurers have responded by experimenting with insurance plan characteristics and cost-sharing aimed at reducing their costs. Policy makers and academics have long had an interest in understanding how people's healthcare spending and health outcomes respond under these plans. However, it is only recently that economic research has begun to emphasize the extent of the dynamics and complexity of beneficiaries' decisions when facing multiple levels of cost-sharing within health insurance plans. This paper studies beneficiary behavior in the context of the Medicare Part D market in 2009-2012, where the government has had significant regulatory oversight on imposing plans with these features.

This paper studies beneficiary responses to plans where the marginal out-of-pocket price that the beneficiary is required to pay out of the total cost of a prescription is not constant (or non-linear) between coverage regions. Figure 1 illustrates an example Medicare Part D 2009 plan contract[1]. The generosity of the plan explicitly depends on the cumulative amounts that patients, the insurance company, and Medicare have spent on prescriptions within the plan-year. In this example, there is a region of initial coverage (ICR) where the beneficiary is responsible for 33.2% of the total prescription costs (33.2% coinsurance), followed by a coverage gap or "doughnut hole" where beneficiaries are fully responsible for costs (100% coinsurance), after which beneficiaries reach the Catastrophic region and have a minimal 5% coinsurance. After the plan-year elapses, plans reset or beneficiaries switch to new plans, and the cumulative total spending for the new plan-year is reset to zero. Within a plan-year, beneficiaries face a truly dynamic problem—consuming healthcare today can impact the marginal price of future healthcare consumption.

Much of the literature has focused on beneficiary behavior at or around the controversial coverage gap. What had been introduced as a cost-saving mechanism has been shown in the literature to have a negative effect on drug adherence. Joyce et al. (2013) and Zhang et al. (2009) find that having a coverage gap disrupts the use of prescription drugs, with a higher decline on more expensive medications as compared to cheaper ones. And while Joyce et al. (2013) fails to detect a corresponding substitution from drugs to medical treatment in concurrent Medicare claims, if one believes that adherence to drug treatments is good for patient outcomes, discontinuing the use of these drugs would have a negative welfare effect on patients.

[1] See Section 3 for more information on the types of plans. In general plans are structured similar to this example with either three or four regions.

Figure 1: Example Medicare Part D 2009 Contract Design



The figure depicts the nonlinear structure of an example Medicare Part D benefit contract with no deductible from 2009. The plan depicted here is actuarially equivalent (with no deductible) to the government-defined contract depicted in Appendix Figure A.1. The premium or the amount the patient pays out-of-pocket for the benefit package is not displayed. The Cumulative Total Expenditure is the year-to-date cumulative sum of the beneficiary's total expenditures, which include the drug expenditure between the patient, insurance company, and Medicare. The Out-of-Pocket (OOP) Cost only includes the patient's drug expenditure in 2009. The initial coverage region's (ICR's) 33.2% coinsurance coverage is approximately the actuarially equivalent to the 2009 deductible plus an ICR coinsurance level of 25%. The 5% coinsurance coverage in the catastrophic region is also simplified for the figure. The actual 2009 coverage benefit requires beneficiaries to pay the maximum of either 5% the cost of the drug or \$2.40 and \$6.00 for a one-month supply of generic and branded drugs respectively. This means that patients may pay either the copay dollar amount or a percentage share of the drug price where the remainder is covered by insurance or the government. The bottom panel displays the marginal cost in each of the coverage regions, or the proportion that the beneficiary is responsible to pay of the total expenditure cost.

In order to fully understand the ways in which changes to the nonlinear plans can impact beneficiary behavior and welfare, researches must also determine whether beneficiaries respond sub-optimally to them, but the current consensus in the literature is somewhat mixed. There are papers where beneficiaries appear to be somewhat optimal and fully forward-looking (Aron-

Dine et al., 2015; Einav et al., 2015), and there are papers that support the conclusion that beneficiaries overly respond to the "spot" or current price (Dalton et al., 2015; Abaluck et al., 2015). As evidence of suboptimal behavior, Dalton et al. (2015) also present the stylized fact that even among individuals in their sample who were very likely to end the year in the coverage gap or beyond, there is a sharp drop in average spending on prescription purchases at the coverage gap.

Ignoring liquidity constraints, fully forward-looking optimal behavior suggests that beneficiaries should use their expected end-of-year marginal price for each purchasing decision throughout the year. For example, a Medicare Part D beneficiary who fully expects to end the year in the catastrophic phase of their insurance coverage should not respond to temporary changes in their plan coverage and spot prices as they spend through earlier benefit phases. Under uncertainty about the end-of-year region and price, however, the beneficiaries may adjust their expected marginal prices as risks are realized. Assuming standard geometric discounting ($\delta >> 0$), these transitions should be smooth, especially early in the year.

Because of the beneficiary's complicated optimization problem under uncertainty, this paper first approximates optimal behavior with a heuristic for constructing a beneficiary's perceived marginal out-of-pocket coinsurance rate. This rate is generated from the entire population's probabilities of ending the year in each region and is a useful tool to visualize how the average beneficiary's expected coinsurance rate evolves across the different weeks of the year and cumulative spending levels. It differs from the optimal beneficiary's expected prices because it uses the ex-post population outcomes, which may not be representative of rational agents. Further, because beneficiaries may themselves have difficulties anticipating their cumulative total end-of-year spending and prices, the heuristic approach may provide intuitions on behavior as well. The heuristic approach predicts, that because probability distributions are quite smooth outside of the last weeks of the year, the expected marginal prices and thus spending should also smooth with any pricing updating occurring prior to the kink.

In addition to the heuristic predictions, this paper also uses a unique regression approach to provide a graphic representation of the empirically observed shape of spending patterns through the entirety of a beneficiary's plan year and across a wide range of cumulative total spending levels. In part, the goal of this empirical exercise is to generate a low assumption view of the extent of a beneficiary's anticipatory response to the different pricing regions as a function of both the time of the year and the beneficiary's cumulative total spending. The regression approach uses a fixed effect regression in a dynamic panel that includes four years of claims

data.[2]

Unlike the majority of the literature, this paper mainly focuses on the frequency of beneficiary claims. Suppose a beneficiary has a prescription to be filled, they have two broad decision options: wait to fill the prescription, or switch the prescription either from branded to generic or with more effort acquire an alternative prescription. The decision to stop taking a course of chronic treatment entirely likely has more of a negative effect on beneficiary health than a decision to switch medication. Thus, this paper measures changes in claims frequency which should capture the beneficiary's first decision to postpone or stop (postpone indefinitely) taking a course of treatment.

This paper expands on the reduced form fixed effects analysis in Dalton et al. (2015) of beneficiaries who were likely to end the year in the coverage gap. Their paper analyzed a 2008 subset of employer-sponsored Medicare Part D individuals and only had indicators to measure the level of spending response (and other dependent variables including prescription occurrence) in four cumulative total spending zones ($310 before, between $310-$110 before, $110 before, and after the coverage gap) rather than a continuous response to cumulative total spending. They found no economic or statistically significant evidence of spending or claims frequency decreases in $310-$110 leading up to the coverage gap, but a sharp decrease in spending and claims frequency in the $110 right before the coverage gap.

This paper's estimates show instead that across four quarters of the year and across cumulative total spending amounts, that while beneficiaries possibly exhibit a drop in claims frequency at the beginning of the coverage gap, they also have a statistically significant anticipatory response far in advance of the gap. However these responses in spending frequency may not be economically significant. As expected, the reduction in spending in advance the coverage gap occurs at higher cumulative total expenditure values (closer to the ICL) in later parts of the year. The discontinuity in the frequency of beneficiary spending directly at the coverage gap appears especially in the first quarter of the year, but that estimate is not stable across different specifications and can possibly be due to small sample sizes of individuals who accumulate such large spending early in the year. A small discontinuity exists at the coverage gap in the last quarter of the year as expected.

This paper's reduced form estimates provide a simpler alternative to studying beneficiary behavior under nonlinear contracts as compared to the structural estimates of Einav et al. (2015) and Dalton et al. (2015). And, the reduced form estimates provide a graphical explanation of

[2]See Section 5.2 for a discussion regarding dynamic panel bias and why it is less of a concern in this setting.

why these papers found both evidence of forward-looking behavior and over response to spot prices respectively. Einav et al. (2015) estimate a model with standard geometric discounting that allows for five types of individuals with different risk levels and sensitivities to the coverage gap, but they recover a weekly discount factor[3] δ equal to 0.96, which roughly translates to a yearly discount factor of only 11%. In order to explain observed drops in spending just prior to the coverage gap, Dalton et al. (2015) expand their model to allow for beta-delta time-inconsistency or present bias, the tendency to overweigh the "present" and have self-control problems. The discounting rates they estimate are indistinguishable between $\beta = \delta = 0$ indicating that beneficiaries only consider the spot price.

Further, intuitions from this paper's heuristic approach indicate that errors in time discounting such as present bias may not be an appropriate model to represent beneficiary behavior that resemble a sharp spending drop at discrete changes in spot prices. A key takeaway from the heuristic model is that outside of the last weeks of the year, only sharp discontinuities in a beneficiary's perceived marginal price or coinsurance rate should result in sharp discontinuities in beneficiary spending. As long as beneficiaries update their expectations weekly, only zero discounting would predict sharp changes at the coverage gap.

The heuristic approach presented here is most similar to the average price model presented in Abaluck et al. (2015), which was used to study prescription spending when plan coinsurance levels changed between years. They estimate the weights that beneficiaries place on the coinsurance rates in either the initial coverage or coverage gap regions. However unlike this paper, they restricted their sample to individuals who are unlikely to have uncertainty about their end-of-year region and cannot make conclusions about anticipatory responses around the pricing kinks.

In the following sections, this paper lays out in Section 2 the intuition on how beneficiaries may behave when they are faced with dynamic nonlinear health insurance contracts such as what is available in Medicare Part D. In a simplified two-region model, this section also introduces the paper's heuristic approach and its implications for beneficiary behavior. Section 3 provides more background on the institutional details of Medicare Part D, and Section 3.1 details the specific sample in this paper—a 2009-2012 sample of individuals who have the government-defined initial coverage limit and out-of-pocket-threshold. Section 4 applies the heuristic approach to this paper's data sample. Section 5 lays out the fixed effects estimation procedure to recover

[3]Einav et al. (2015) refer to the estimate as a "'behavioral' parameter that also reflects individual's understanding of the insurance coverage contract, in particular the salience of the (future) nonlinearities of the contract", as part of the reason why δ is so low.

beneficiary's spending function as a function of time and the cumulative year-to-date spending totals. Section 6 concludes.

2 Intuition

In the Medicare Part D setting, patients make a combination of periodic and unexpected purchases of prescription drugs within a year. Within the decision to make each prescription purchase, as with all economic decision-making, beneficiaries employ some cost-benefit analysis comparing the perceived cost with the perceived benefits of purchasing and then consuming drugs. While beneficiary's beliefs on the medical benefits of consuming certain drugs are important, this paper focuses on the beneficiary's beliefs on his or her monetary costs. This analysis makes the assumption that the benefits of drug purchase and consumption do not depend directly on the arbitrary contract design or the time of the year.

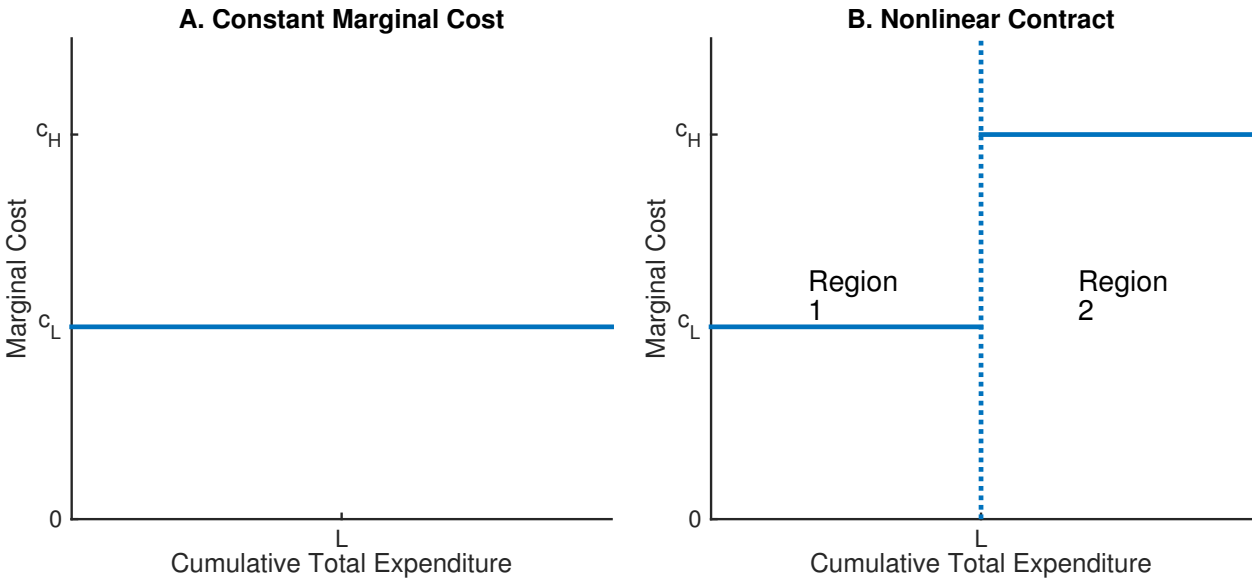
Suppose a beneficiary with observables X faces a decision whether to fill a prescription that costs s in week w of the year.

$$B(s|X) > PMC(Z_w, w|X) * s$$

To build intuition, consider the two simple pricing contracts in Panel A and B of Figure 2. If the beneficiary were in a plan with contract A, she should always expect that her marginal cost of purchasing a drug would be equal to c_L, as it is the coinsurance rate applied to all purchases. It doesn't matter if she has to make a purchasing decision in the beginning of the year or the end, her $PMC(Z_w, w|X) = c_L$.

In contract, Plan B is a nonlinear cost structure, and thus a beneficiary's perceived marginal cost for purchasing a drug is less clear. Under this plan, her spot price in a given week $MC(Z_w) = c_L$ if $Z_w < L$ and $MC(Z_w) = c_H$ if $Z_w \geq L$. At the two extremes of behavior, a beneficiary may only respond to the spot marginal price for care determined by the current insurance region, or a beneficiary may be a "fully forward-looking", perfectly rational economic agent. If she only responds to the spot price, then her $PMC(Z_w, w|X) = MC(Z_w)$. Her PMC is just her coinsurance rate in the region she is in in week w, and she does not take into consideration how her spending can impact her future marginal costs. If a beneficiary is "fully forward-looking", in each purchasing period decision, she optimizes her decision by discounting the future stream of expected benefits and costs that result from her current decision, including any changes to her expected marginal costs due to the non-linear pricing. Ultimately if a beneficiary is fully forward-

Figure 2: Marginal Costs



looking, her perceived marginal price in a period should be her expected year-end marginal price (Einav et al., 2015; Abaluck et al., 2015; Dalton et al., 2015). Thus, her expectation of ending in Region 1 or 2 matters. If she is entirely confident in week w that her end-of-year (week W) cumulative total expenditure will be below the limit L, $Pr(Z_W < L|Z_w) = 1$, then her perceived marginal cost should always be $PMC_w = c_L$. Similarly if she is confident that her end-of-year cumulative total expenditure is greater than L, $P(Z_W \geq L|Z_w) = 1$, then her end-of-year marginal cost should be the cost in Region 2, or $PMC_w = c_H$. Depending on the beneficiary's perceived uncertainty about spending past L, and conditional on not passing L when evaluating her problem in week w, her perceived marginal cost in w may be somewhere between c_L and c_H.

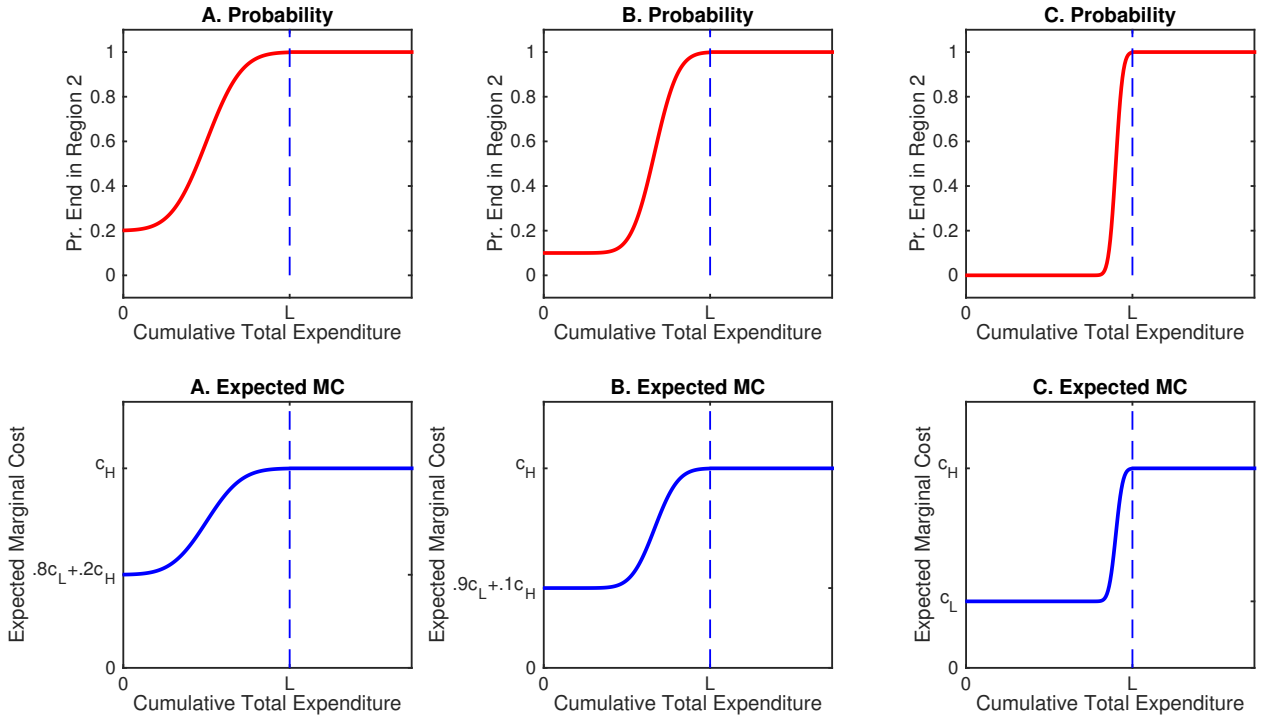
Within the "in-between" response, other papers particularly Dalton et al. (2015) has leaned on present-bias or inconsistent time-discounting to explain beneficiaries' behaviors in their data on 2008 Medicare Part D claims. However, while their structural model that allowed for present bias was a better fit than a standard discounting model, their estimation result of discount factors that were indistinguishable from zero $\beta = \delta = 0$ is indicative that present bias may not be an appropriate model. Abaluck et al. (2015) also allow for an "in-between" response to inter-year changes in the coinsurance rates by estimating the weights beneficiaries in Medicare Part D place on changes in coinsurance rates in the initial coverage region or coverage gap. While Abaluck et al. (2015)'s heuristic is somewhat similar to mine, in part due to the nature of their

empirical approach, they limit their study to individuals who the researchers were confident to end the year in either the initial coverage region or coverage gap.

This paper proposes an alternative "in-between" response that a beneficiary may use to make the prescription purchasing deciding. She may use a heuristic mental shortcut to calculate her objective expected year-end marginal price. Rather than comparing the net present value of the costs and benefits to conceive of an optimal expected year-end price, she may estimate a perceived marginal cost based on her beliefs of the population objective probability of ending the year in any contract region. Using the objective population probability of being in any contract region, she can then infer the price in that region. Note that in the two region case, $Pr_w(\text{in Region 2 in W}|Z_w) = 1 - Pr_w(\text{in Region 1 in W}|Z_w)$.

If a beneficiary were to make a decision to spend on prescription on the last period of the year, no-matter her method of evaluating her "in-between" response, her perceived marginal cost is just her actual marginal costs in whichever region she is in. There is no uncertainty as to what her marginal cost is. However, at earlier parts of the year $w < W$, her beliefs on her probability of ending the year in any particular region, and the associated expected marginal cost can impact her spending patterns.

Figure 3: Example Probabilities of Ending the Year in Region 2 and Expected Marginal Cost



The simulation assumes an individual faces example marginal costs conditional on the depicted in the top graphs.

Consider the example probabilities of ending the year in Region 2 presented in Figure 3. In the Panel A graphs, suppose that early in the year, the beneficiary who has $0 cumulative total spending believes that there is only a 20% chance of ending the year in Region 2 (and necessarily a 1-0.2=0.8 chance of ending the year in Region 1). As the cumulative total expenditure increases prior to $Z_w = L$, her probability of ending the year in Region 2 increases. Even before she reaches $Z_w = L$, because there are still many weeks left in the year, she anticipates an almost 100% chance of ending the year in Region 2.

Suppose she employs the heuristic with no discounting to generate her marginal cost, then her objective expected marginal cost (HMC) is below.

$$HMC_w = c_L * Pr_w(\text{in Region 1 in W } |Z_w) + c_H * Pr_w(\text{in Region 2 in W}|Z_w)$$

. Her overall expected marginal cost would then be the function depicted in in Figure 3 A. Her perceived marginal cost transitions from $.8 * c_L + .2 * c_H$ to c_H as the cumulative total expenditure increases. Her spending is expected to be a monotonically decreasing function of her perceived marginal cost and should also adjust far prior to the limit L.

Suppose Panel B represents a week in the middle of the year and Panel C represents a week at the end of the year. As time progresses, the beneficiary's probability of ending the year in Region 2 become more certain and further resemble a piecewise graph with 0 probability of ending the year if $Z_w < L$ and 1 otherwise. In Panel B, when the beneficiary has accumulated zero or low levels of prescription spending, she expects a lowered 10% chance of ending the year in Region 2, because there is less time left in the year for a negative health shock as compared to Panel A. The trend continues in Panel C, and there is little uncertainty to the region beneficiaries end the year in. Those who have spent 0 are certain to end the year in Region 1 and should spend using the c_L rate, while those who have crossed the limit L spend according to the c_H rate. There is only a small range of cumulative total expenditure amounts where a beneficiary may have positive but not certain probability of ending the year in Region 2. It is really only in Panel C and the end of the year that there may be sharp changes in the probability of ending the year in Region 1 or 2 and thus expected marginal costs and spending patterns. It is natural to see how these expectations could generate spending that may appear to be discontinuous at the limit L, especially if there are few individuals observed in the transition region.

While the examples in Figure 3 did not include any explicit discontinuities in the probabilities or expected marginal costs, an alternative scenario could exist. For example, suppose that in a period w early in the year beneficiaries believe that there is a 20% chance of ending the year in

Region 2 for all $Z_w < L$ and by necessity if they are aware of entering Region 2, the probability becomes 1 if $Z_w >= L$. The explanation for why a beneficiary near the limit may not upwardly revise her 20% probability of ending the year in Region 2 may include her inattention to being so close to Region 2, or completely ignoring her potential future health shocks (having a zero discount factor).

This paper argues that non-zero geometric discounting and present bias models should not generate such discontinuities. Consider the adaptation to the heuristic model that allows for present bias. Assume the decision in the heuristic approach only considers today's medical impact (or benefit of the drugs) compared to the total cost today multiplied by the future perceived marginal cost. The beneficiary would necessarily be making her spending decisions by discounting her entire heuristic marginal price $PMC_w = \beta\delta^{W-w} * HMC_w$. So she would purchase her prescriptions if

$$B(s|X) > \beta\delta^{W-w} * HMC(Z_w, w|X) * s$$

where β is the "present bias" discounting factor that represents the difference between the present t and all future outcomes and δ^{W-w} is the geometric or standard discounting factor that is the product of discounting in every week from the current week w to the end of the year W.

Assume a beneficiary has Panel A probability beliefs that generate a continuous heuristic marginal cost. Denote $\beta'_w = \beta\delta^{W-w}$ for a given week w. Then, introducing a $\beta > 0$ and $\delta > 0$ implies $\beta'_w > 0$ and will not generate a significant discontinuity in her perceived marginal costs. Instead her perceived marginal cost would fall somewhere between $\beta'_w(.8c_L + .2c_H)$ and $\beta'_w c_H$ in week w.

Further, introducing present bias as a explanation of suboptimal behavior observed in the empirical data is not necessarily appropriate. Present bias models of behavior only generate suboptimal behavior when decisions are made between "the present" and "the future" and not when considering different time points within the future. The assumption that the benefits of prescription coverage are incurred "today" and in the present, while prevalent in the literature is not necessarily accurate. When beneficiaries fill prescriptions, they are often not for immediate consumption with 30 and 90 day supplies. Further, even if consumption was immediate, as Baicker et al. (2015) explain, the effects of many prescriptions such as statins to treat high cholesterol have far delayed benefits rather than any immediate symptomatic changes.

In Section 4, this paper applies the heuristic approach to the empirical setting of Medicare

Part D. For plans with three regions, the actual objective probabilities of ending the year in each of the regions are presented along with the objective expected coinsurance rates.

3 Background on Medicare Part D and Data

Before discussing the exact data that is used in this study, this section covers the institutional details about the Medicare Part D program and the specific plan types that are part of the program.

In the United States, Medicare is a health insurance program for the elderly that covered approximately 46 to 51 million individuals from 2009 to 2012.[4] It is structured in four parts: A, B, C and D. Parts A and B include hospital and medical insurance for in- and outpatient care. Patients who are enrolled or eligible for Parts A and B, are also eligible to enroll in Medicare Part D, which provides insurance for the prescription drug purchases, covering mostly self-administered drugs. Unlike Parts A and B, which are administered by the government, the Part D plans are administered by private insurers who are subject to the rules and regulations laid out by the government. Part C, also known as Medicare Advantage, is also administered by private insurers and is an all-inclusive alternative to Parts A, B, and D. While enrollment in Medicare is voluntary, individuals face significant penalties within the program if they choose not to sign up when first eligible (usually at 65) or have creditable (similar) health insurance coverage.

This paper focuses on beneficiary purchasing behavior of enrollees in Medicare Part D with stand-alone Prescription Drug Plans (PDP). In 2009, there were almost 18 million enrollees, and there were almost 20 million by 2012. The program began with the Medicare Modernization Act of 2003, which was enacted in 2006. The government regulates a "standard" plan with a baseline minimum amount of coverage, and private insurers can offer a variety of plans that on an actuarially equivalent basis meet or exceed the generosity of the standard plan. This results in a significant variety in prescription coverage plans with 1,689 stand-alone PDP plans in 2009 ("The Medicare Part D", 2016).

In the context of studying beneficiary behavior when faced with nonlinear prices, the main advantage of studying Medicare Part D is the highly nonlinear structure of the coverage regions in the stand-alone Prescription Drug Plans (PDPs). While this paper will focus on plans that do not have a deductible, because the majority of beneficiaries do not choose plans with deductibles,

[4]Program statistics from the Centers for Medicare & Medicaid Services' Statistical Supplement https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Archives/MMSS/index.html

this section will discuss the full variety of plan structures. The 2009 standard Part D Plan included a deductible of $295, an initial coverage limit (ICL) of $2,700, and an out-of-pocket threshold (OOPT) of $4,350. While the specific deductibles, ICLs, and OOPTs differ every year, the structure of the standard plans and thus plans in the market are similar to that depicted in Figure A.1.

In the standard plan in all years, the patient is responsible for 100% of the cost of prescriptions until their cumulative spending reaches the deductible amount, after which they are in the initial coverage region (ICR). Note that both patient out-of-pocket and "total" spending (the total amount spent through a combination of patient, insurance company, government through Medicare, and drug companies) are the same up to the deductible. Patients in the Initial Coverage Region (ICR) are then responsible for 25% or less of the total price of prescription purchases. Once the patient's cumulative total spending amount reaches the ICL amount, patients enter the phase often referred to as the coverage gap or "doughnut hole" where they are again responsible for 100% of the total spending amount in 2009. Once the patient's cumulative out-of-pocket costs reaches the OOPT, they reach the "catastrophic" phase and are responsible for paying a greatly reduced share of the total costs of drugs. Specifically, they pay either the maximum of 5% of the total price of prescriptions or a $2.40 and $6.00 copay for a one-month supply of generic and branded drugs respectively. The pricing schedule resets at the end of the calendar year, and at the beginning of the next year beneficiaries begin anew with a total cumulative spending of 0 and the associated marginal costs.

The standard plan up until the catastrophic region has an exact mapping of out-of-pocket and total payments—an ICL of $2,700 corresponds exactly to cumulative OOP costs of $896.25, and an OOPT of $4,350 corresponds to $6,153.75 in total expenditures. In practice, an exact mapping between the two cumulative spending measures is difficult to ascertain in all plans. Plans only have to meet (or exceed) the coinsurance generosity of the standard plan on an "actuarially equivalent" basis through coinsurance, copays, or a combination of the two. Because of the actuarially equivalent clause, insurance companies have tremendous flexibility in structuring plans. Often, the cost sharing can be specific to drug tier or whether it is branded or generic. Regulation just requires that on average, the plan is expected to be similarly or more generous than the standard plan.

The inclusion of the Part D coverage gap or "doughnut hole" (and a main component of the plan nonlinear structure) has been widely criticized and analyzed in the healthcare literature. The coverage gap was initially included as a cost-saving measure for the government similar to

a deductible but positioned in the middle of the patient benefit schedule (Baker, 2006). The health policy literature has both criticized the arbitrary location of the doughnut hole and the impact it has on drug adherence.

Every year, the standard plan's spending limits are updated to adjust for rising costs. And under the Patient Protection and Affordable Care Act (ACA), the government began to fill in the coverage gap and will continue to increase the plan benefits in this region through 2020. The phase out began in 2010 when the standard plan included an automatic $250 rebate for beneficiaries who reached the coverage gap. Further coverage in the doughnut hole increased in 2011 and 2012, when instead of a rebate, standard plans included a 50% discount on brand name prescriptions that was paid by the drug manufacturer. While it might be considered unintuitive, in most cases the 50% discount while not actually paid by the patient, did contribute towards their cumulative out-of-pocket threshold. This means that in 2011 and 2012, patients in the doughnut hole only paid 50% of the cost of branded drugs. Also because the drug manufacturer discount counted towards patient out-of-pocket thresholds, the discount did not significantly change the total expenditure amount it took for patients to get out of the doughnut hole. Table 1 shows the changes in the standard plan from 2009-2012.

Table 1: Medicare Part D Benefit Parameters for Defined Standard Benefit 2009-2012

Plan Characteristics	2009	2010	2011	2012
Deductible	295	310	310	320
Initial Coverage Limit (ICL)	2700	2830	2840	2930
Out-of-Pocket Threshold (OOPT)	4350	4550	4550	4700
Total Expenditure equivalent OOPT	6153.8	6440	6447.5	6657.5
Rebate (1)		250		
Brand discount(2)			50%	50%
Generic copay (3)	2.40	2.50	2.50	2.60
Branded copay (3)	6.00	6.30	6.30	6.50

(1) The rebate begins when patients reach their out-of-pocket threshold (OOPT).
(2) The brand discount only applies when patients are in the coverage gap, i.e. when their cumulative total spending is above the ICL, and their cumulative non-insurer spending (patient payments, any subsidies, brand discounts paid by the drug manufacturers) is below the OOPT.
(3) In the catastrophic region, beneficiaries pay the maximum of the copay or 5% the total cost of the prescription.

This paper will focus on plans that have the government-defined ICL and OOPT, but not plans with the government-defined deductible. Enrollees have choices over a wide variety of Part D plans, with a majority of patients opting for plans with more generous plan benefits than the standard plans including no deductible plans. In 2006, fewer than 10% of beneficiaries were in plans with the standard design (Abaluck and Gruber, 2016), and our data support this finding

as well. Even prior to 2010 when the ACA started phasing out the coverage gap, many plans offered some type of gap coverage, though these were typically on generic prescriptions. See Section 3.1 for a deeper discussion of plan types.

For the purpose of understand beneficiary behavior in the face of nonlinear contracts, there are other advantages to studying the Part D plans instead of other nonlinear employer-sponsored health insurance plans. These advantages include the widespread frequency of claims and large percentages of beneficiaries experiencing different coverage phases year-over-year. Hoadley et al. (2011) indicate that 16% of Medicare beneficiaries ended the year in the coverage gap, with 3% of beneficiaries reaching the gap and passing it to end the year in the catastrophic region. Across 2008-2009, almost 30% of patients experienced the gap. Further, they document that reaching the coverage gap is consistent, as 71 percent of enrollees who reached the gap in 2008 did so again in 2009. This recurrence of reaching the coverage gap is due to the fact that many patients are taking medications for chronic conditions rather than for acute, short-term medical needs.

Because many of the medications patients take in Part D are for chronic conditions, enrollees in the nonlinear Part D setting may have a better ability to forecast their yearly spending on prescription drugs than enrollees do in employer sponsored health insurance plans. In a MedPac report on Medicare Part D, they state that the list of top 15 therapeutic classes of drugs by spending and volume has remained relatively consistent since 2007. The values from 2013 indicate that drugs in the diabetic, asthma/COPD antihyperlipidemics,[5] antipsychotics, antihypertensive,[6] and peptic ulcer therapeutic classes are responsible for approximately 40% of drug spending (MedPAC, 2016). These drugs are all used to treat ongoing chronic conditions.

Further, in studying beneficiary behavior, it is also an advantage that Part D claims only cover self-administered drugs and do not cover drugs administered at the doctor's office or in the hospital. Unlike hospital claims, there is an increased likelihood that the prescription purchases are the beneficiary's decision rather than decisions made by a medical professional. However, it is still a concern for older patients that drug purchases could be done by a proxy.

One of the challenges of the Medicare Part D data for studying spending in the non-linear pricing schedule across years is that the pricing schedules change over years. As mentioned, each year the standard plan adjusts and it is highly likely that the individual private plans adjust. Patients also have choices to switch between plans with different insurers across years and to switch across into Medicare Part C. However, while plans change every year, the schedule remains similar with mostly minor increases in the exact limits of each coverage region. The

[5] Used to treat high cholesterol

[6] Used to treat high blood pressure.

literature also document a significant amount of inertia and inattention in patient choice of plans, indicating that approximately 10% of Part D patients switching their plans between every two years (Abaluck and Gruber, 2016; Abaluck et al., 2015; Ho et al., 2015).

3.1 Data Description

The primary dataset includes the prescription drug and medical claims of a random 5% subsample of Medicare beneficiaries enrolled in Part D in the years between 2009 and 2012.[7] The data comes from the Center for Medicare and Medicaid Services. Broadly, the data cover the beneficiary demographics, their Part D prescription claims, and the plan characteristics of the specific Part D plan that each beneficiary chose. The prescription drug claims include the exact drug purchased, days supply, purchase date, the proportion paid by both the patients and insurance companies and the benefit phase each claim occurs in. The plan characteristics supplement this information on the contracts that patients face with full details on the plan premiums and the exact cost-sharing characteristics: deductibles, coinsurance, copays for specific drug types and tiers.

Basic demographic information of the beneficiaries (gender, age, race) along with hospitalization and doctor claims information from Medicare Part A and B that are used to determine patient health conditions are also observed. With these data, I use CMS-provided risk model to calculate a "risk score" or summary estimate of the expected average drug spending implied by patients' demographics and health conditions.[8]

The original data come from a 5% sample of Medicare/Medicaid enrollees over 2009-2012 and contains approximately 2 million individuals per year.[9] The data used for the analysis is limited to a far more homogenous group of only Medicare beneficiaries with Medicare Part D plans in these years.

The "baseline" sample is a balanced panel of beneficiaries who are in the Medicare Part D system from 2009 through 2012, have plans with no deductible but the standard ICL and OOPT limits, and have at least one claim in each of these years. A set of reasons why beneficiaries are omitted from the sample is listed in detail in the Appendix Section A and Appendix Table A.1

[7] This analysis does not consider patients enrolled in Medicare Advantage (Part C), because while the Medicare Part D claims data include Part C prescription claims, the dataset does not include the doctor and hospitalization claims, which are used to control for heterogeneity.

[8] CMS use Hierarchical Conditional Codes (HCC) and RxHCC (for prescriptions) to adjust the reimbursement payments to insurance companies that offer plans in Medicare Advantage and other programs. The HCC and RxHCC scores are scaled reimburse the plans for managing patients with illnesses with expected increased medical and prescription medication costs respectively. http://setma.com/EPM-Tools/tutorial-hcc-rxhcc-risk

[9] Summary statistics for the entire 5% sample are in Table A.2.

with the percentage of the entire Medicare/Medicaid 5% sample they encompass. In general, sample contains beneficiaries who are 65 or older and are enrolled in Medicare PDPs from 2009-2012 through the Old Age and Survivors Insurance (OASI) and not for disability insurance or other qualifiers for Medicare. The sample also excludes individuals who are dual eligible for Medicaid financial assistance or receive other types of low-income subsidies (LIS) for premiums or cost-sharing. These individuals are excluded because they face very low cost-sharing and minor changes in their marginal costs. Even individuals who only receive premium subsidies are omitted, because they are more likely lower income and are more likely to be influenced by budget constraints. Further, the analysis of the paper also excludes individuals whose Medicare Part B claims indicate beneficiaries were in long-term care institutions (LTI) such as nursing homes in the prior year. This leaves approximately 300k beneficiaries in each of the 2009-2012 samples respectively (Appendix Table A.3).

The variety of plans and the beneficiary's choices to switch plans between years poses a challenge for studying beneficiary behavior, since there may be an endogenous relationship between beneficiary's choice of plans and their spending patterns. That plan variety can be seen in Table 2. Very few enrollees over the years have chosen plans with deductibles; 65-75% of enrollees have plans without deductibles. Overall only 15-20% of enrollees have plans with the standard government-defined plan limits in the deductible, Initial Coverage Region, and Out-of-Pocket Thresholds. However, a majority of patients still have the same ICL and OOPT limits with 76% without a deductible. But despite the pervasiveness of the ICL and OOPT spending limits, less than 2% of beneficiaries have plans that use the exact 25% coinsurance rate suggested by the government for the initial coverage region. Because many beneficiaries do have plans with the Medicare-Defined spending limits for their ICL and OOPT, this paper will focus on the beneficiary's response to approaching these limits rather than her response to the specific coinsurance rates.

The baseline "No Deductible" sample then focuses on the balanced panel of individuals who qualify under the above sample restrictions throughout 2009-2012, had the government prescribed limits for the ICL and OOPT, and who did not have deductibles. Individuals were also omitted if they were observed to have either zero spending in any year or had claims in every week of the year. Keeping only individuals with these plans limits, the sample contains 89,354 beneficiaries or about 1/3 of the full sample. Part of the reason for this restriction is to help standardize the spending limits for the analysis in later sections. This restriction has the negative effect of decreasing the sample size and reducing the generalizability of these results.

Further, because these individuals choose plans without deductibles, the selected sample may be simultaneously more risky, more risk adverse, or richer. Also because the sample requires the beneficiary have the same plan structure in all four yeas, they have higher inertia and may have higher costs of switching. An additional "Standard" sample is also created from the 9,178 individuals who signed up for plans with the government-defined deductible, ICL, and OOPT. See the Appendix for the analysis of the sample of individuals who had standard plan limits "Standard".

Notice that this dataset of 2009-2012 claims differs from all of the papers previously mentioned in the literature. In all cases, this sample draws from a later sample of individuals with Medicare Part D than individuals who are in Joyce et al. (2013)'s 2006, Dalton et al. (2015)'s 2008, Einav et al. (2015)'s 2007-2009, and Abaluck et al. (2015)'s 2006-2009 sample. It is also a longer sample than most other papers covering four years. Because this sample involved individuals who retained similar plan structures through all four years, they are mechanically more likely to have experience with Medicare Part D[10], their plan structure, and their prescription needs than individuals described in these other papers. These differences possibly translate to different findings in the empirical section.

Table 3 displays the demographics of the baseline sample of individuals without deductibles. The average age of the population in 2009 is about 75, which is slightly older than the average Medicaid 5% sample population, but is consistent with the ages of beneficiaries enrolled in Medicare Part D. The vast majority (approximately 95%) of beneficiaries are white. They are a sicker population, which may reflect both the higher age and the optional nature of joining Medicare Part D for prescription purchases. A majority of individuals have chronic conditions with 66% and 73% of individuals having a prior year condition code of hypertension and high cholesterol in 2009. Also, almost a quarter of the sample has diabetes and 10% has had cancer treatment of some kind in 2009. These are all conditions that often require constant prescription refills and spending using the Medicare Part D benefit. The Kaiser foundation documents that patients who took drugs to treat some of these conditions are far more likely to reach the coverage gap hole and catastrophic regions (Hoadley et al., 2011). They observed that the average Part D enrollee's probability of reaching either of the two regions as 19% in 2009 but 56% for patients on breast cancer treatment drugs, 40% for those taking oral anti-diabetics, 32% for those on statins etc. Also notice that as the beneficiaries progress through the years, they get older and sicker. The increase in sickness is reflected in the claims data with an increasing

[10]Remember also Medicare Part D began in 2006.

trend in spending amount and frequency through 2009-2012.

Table 2: Summary of Full Sample of Medicare Beneficiary Plans 2009-2012

	2009 mean	2010 mean	2011 mean	2012 mean
Deductible: None	0.7627	0.6828	0.6258	0.6675
Deductible: Other	0.0391	0.1690	0.1901	0.1483
hasStandardDed	0.1970	0.1481	0.1841	0.1838
ICR: Standard Coinsurance	0.0101	0.0095	0.0119	0.0041
ICR: Cost Share Tiers	0.9899	0.9905	0.9881	0.9959
ICL: Standard	0.9897	0.9262	1	1
OOPT: Standard	1	1	1	1
Standard Plan Limits	0.1970	0.1481	0.1841	0.1838
No Ded, Standard ICL & OOPT	0.7528	0.6090	0.6258	0.6675
Main 4 Year Sample	0.3358	0.3216	0.3082	0.2938
Observations	291,550	304,477	317,670	333,309

The full sample includes individuals in each year who satisfy the criteria for the "Full 12 Month Sample" from Table A.1.

Table 3: Demographics of Baseline No Deductible Sample of Medicare Beneficiaries in 2009-2012

	mean	sd
Age at End of 2009	74.9412	6.58
Female	0.6461	0.48
Race: White	0.9519	0.21
Race: Black	0.0240	0.15
Race: Other	0.0121	0.11
Race: Asian	0.0071	0.08
Race: Hispanic	0.0035	0.06
Observations	89,354	

	2009		2010		2011		2012	
	mean	sd	mean	sd	mean	sd	mean	sd
2011 RxHCC weight	0.4753	0.29	0.5063	0.29	0.5265	0.30	0.5487	0.31
2011 RxHCC demographic weight	0.4196	0.01	0.4196	0.01	0.4196	0.01	0.4197	0.01
hasDiabetes	0.2441	0.43	0.2591	0.44	0.2708	0.44	0.2800	0.45
hasHypertension	0.6664	0.47	0.6918	0.46	0.7009	0.46	0.7072	0.46
hasCancer	0.1016	0.30	0.1075	0.31	0.1107	0.31	0.1153	0.32
highCholesterol	0.7325	0.44	0.7571	0.43	0.7650	0.42	0.7670	0.42

The Baseline No Deductible sample includes individuals in each year who satisfy the criteria from Table A.1 and also were in a plan with standard ICL and OOPT limits with no deductible from 2009-2012. Risk scores are normalized to 2011 RXHCC scores for consistency across years.

Even though the majority of beneficiary plans do not follow the standard Medicare-Defined coinsurance amounts, the coinsurance levels between the different coverage regions are on average still economically and significantly different from each other. Table 4 displays the average coinsurance amount that beneficiaries in the baseline No Deductible Sample face. Because the

plans patients in these regions choose do not have deductibles, their effective coinsurance rate in the initial coverage region (ICR) is higher than the standard plan at 47% to 53% of the total cost of care. The coinsurance rate in the doughnut hole in 2009 and 2010 prior to the ACA legislation to filling in the doughnut hole was not quite 100% but still significantly high at approximately 91%. The beneficiary responsible portion of the coinsurance rate during the coverage gap in 2011-2012 was significantly lowered to effectively 63-64% with the addition of the 50% discount on branded drugs. While the difference between the ICR and Coverage Gap coinsurance rates in the first two years of the sample is higher than latter two years, there is still a difference in latter two years. This means that beneficiaries should qualitatively still respond to these plan characteristics in the way laid out in Section 2. It is expected that any beneficiary response to the coverage gap in 2009-2010 may be muted in 2011-2012, because the marginal price difference between the two regions is smaller. The average coinsurance levels for patients with standard plan limits follow closer to the government recommended plan and are shown in the Appendix Table A.4.

Table 4: Average (person-week) Coinsurance in Phases in Baseline Sample No-Deductible Plans 2009-2012

	2009		2010		2011		2012	
	mean	count	mean	count	mean	count	mean	count
ICR	0.3863	4,285,602	0.4069	4,300,522	0.3975	4,277,143	0.4071	4,293,291
Coverage Gap	0.9223	319,117	0.9155	306,331	0.5508	325,492	0.5428	309,185
Catastrophic	0.0587	32,100	0.0580	30,742	0.0595	36,657	0.0597	38,708
All	0.4209	4,636,819	0.4382	4,637,595	0.4056	4,639,292	0.4133	4,641,184

Table is generated from the baseline sample individuals with no deductible with standard ICL and OOPT limits. The coinsurance rates are averaged over the amount the patient pays (doesn't include the drug manufacture discounts in 2011 and 2012) divided by the total expenditure cost in the person-week observation where spending occurs. This rate is effectively weighted by the time individuals spend in each phase. The count reflects the fact that there are more person-week observations in the ICR region than others. These sums do not reflect the counterfactual coinsurance rates that beneficiaries with low spending would have faced if they had reached higher spending. While the data contain the actual structure of the beneficiary plans with exact coinsurance and copay rates for drug tiers, it is difficult to summarize in a specific coinsurance rate without knowing the mix of drugs that patients may consume.

Since the baseline sample is used to study beneficiary behavior as they cross spending phases, the sample does include individuals who are likely to reach the coverage gap and beyond. Table 5 shows that while the majority of the beneficiaries with and without deductibles end the year in the initial coverage region, over 25% of enrollees end the year past the coverage gap with approximately 3-4% reaching the catastrophic coverage region. While the heterogeneity among beneficiaries mean that many would not have realistic expectations of reaching the higher spending coverage regions, there is certainly a significant sub-set of enrollees who might expect

to end the year in these regions.

Table 5: Proportion of Beneficiaries in Each Phase at the End of the Year in Baseline Sample 2009-2012

	No Deductible			
	2009	2010	2011	2012
ICR	76.47	77.70	76.93	78.81
Gap	21.03	19.97	20.28	18.46
Catastrophic	2.49	2.32	2.79	2.73
Observations	89,354	89,354	89,354	89,354

Table is generated from the baseline sample individuals with no deductible with standard ICL and OOPT limits. The proportion of beneficiaries that end the year in each phase is averaged over the individual beneficiary.

The beneficiary's raw probability of spending (i.e. the probability of submitting a claim) in each of the Medicare Part D insurance coverage regions is depicted in Table 6. Across the four years, while the overage probability of making a prescription claim in a week is 32-34%, the raw probabilities do differ significantly within the insurance regions. Of the beneficiaries who are in the initial coverage region, their average probability of spending is in the 31-33% range. The weekly claim probability for the observations in the coverage gap is higher at 43-47% and highest in the catastrophic region at 56-57%. Given the number of individuals who end the benefit year in each region, and the large number of person-week observations in the ICR relative to the catastrophic region, it makes sense that the ICR probability is closer to the total average probability. The differences in the probabilities across regions is a possible indication of heterogeneity in the probability of spending, where individuals with higher probabilities of claims, who may also have higher average spending amounts, are more likely to have observations in the coverage gap and catastrophic regions. Any analysis on the actual probably of spending for an individual beneficiary as they cross the spending regions needs to take into account the vast amount of between-subject heterogeneity. The approach that is taken in this paper to handle this heterogeneity is through the use of fixed effects in Section 5.

The limitations of these data include the fact that they do not capture prescription purchases outside of Part D such as large retailer generics since those purchases are outside the scope of the Medicare system. Further, this paper has limited data on beneficiary incomes. One potential explanation for non-standard behavior could be patients reaching budget constraints, and this paper is not able to directly measure individual liquid wealth.

Table 6: Average Probability of Weekly Spending in Coverage Regions in No Deductible Plans 2009-2012

	2009	2010	2011	2012
ICR	31.48	31.90	32.15	32.71
Coverage Gap	42.65	43.63	45.01	46.93
Catastrophic	56.22	56.84	55.87	56.84
All	32.44	32.86	33.26	33.87

Table is generated from the baseline sample individuals with no deductible but standard limits for the ICL and OOPT. Table displays the raw probability of spending in a week in each coverage region and year. Counts of the number of person-week observations in each phase can be found in Table 4.

4 Heuristic Approach Applied to Medicare Part D

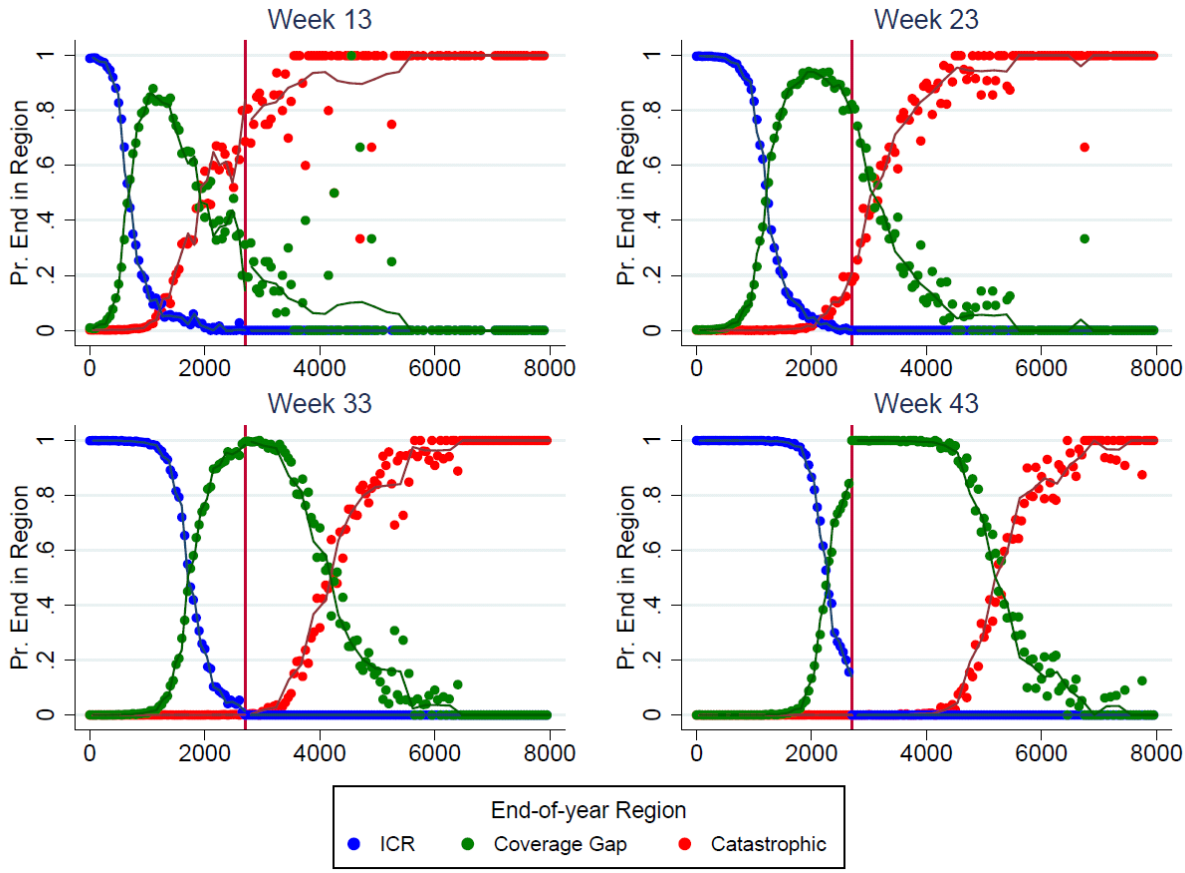
This section applies the heuristic approach to the empirical data proposed in Section 2 to recover the average objective expected end-of-year prices beneficiaries should expect at any given level of cumulative spending. One drawback of the heuristic method is that it requires a strong assumption that a beneficiary has access to the data to know her objective probabilities; however, this heuristic is still useful as a tool to both understand how optimal agents should behave. Further, if the stylized facts about beneficiaries sharply reducing their spending when approaching the coverage gap earlier in the year are generally true, then under this heuristic this behavior would only be explained by discontinuities in beneficiary's subjective end-of-year probabilities, i.e. if they fail to update their beliefs or receive health shocks that lead to surprises.

There are more regions and more noise in these estimates than the simulated example, but basic interpretations hold. First the probability distributions of ending in each of the contract regions based on the time of the year and beneficiary year-to-date spending are constructed. While these graphs are constructed at the cross-individual level, they should still provide insight for individuals to construct their internal beliefs.

For the baseline sample of beneficiaries with no deductible, Figure 4 displays the raw probability in weeks 13, 23, 33, and 43 of ending the year in each coverage region conditional on their cumulative total expenditures (through week 12, 22, 32, and 42 respectively). The heuristic expected marginal cost (HMC) is constructed using these probabilities as discussed in Section 2 and displayed in Figure 5. The both the probabilities and marginal costs are calculated from beneficiaries whose weekly cumulative total expenditure amounts fall in \$50 bins in the x-axis. The red line represents the initial coverage limit and the boundary between the initial coverage region of low coinsurance and the coverage gap. There are significantly more differences in the bin means for the probability ending the year in the catastrophic and initial coverage gap con-

ditional on the cumulative total spending. This is due in large part because the out-of-pocket threshold (i.e. the boundary between the coverage gap and catastrophic region) translates to different cumulative total expenditures for different insurance plans and drug consumption patterns. The noise in probabilities is exacerbated in the week 13 panel because there are few individuals who have accumulated high spending levels that early in the year.

Figure 4: Distribution of the Probability of Reaching each Coverage Region in 2009 Conditional on Cumulative Total Spending



Each point is the average probability of a beneficiary reaching coverage region R at the end of the year given week w and within a \$50 bin of the cumulative total spending Z_w. This makes up the distribution $F_R(Z_w, w)$. This figure is generated based on non-LIS beneficiaries who had PDP plans in 2009 that had the no deductible plans with the government-defined initial coverage limit at \$2,700 (shown by the red vertical line) and out-of-pocket threshold (not shown). The bin size = 50 and was chosen for illustrative purposes. Similar images for 2010-2012 are included in the Appendix. Because the out-of-pocket threshold limit for entering the catastrophic coverage region does not on aggregate map to a specific cumulative total amount, the average probability of ending in those phases as a function of the cumulative total amount has a wide dispersion of points.

While much of the prior literature has focused on the behavior directly at the initial coverage limit, the effect of the non-linear pricing structure could be evident far prior to the coverage gap earlier in the year. In week 13, if a beneficiary's spending is between \$500 and \$1,500, she is

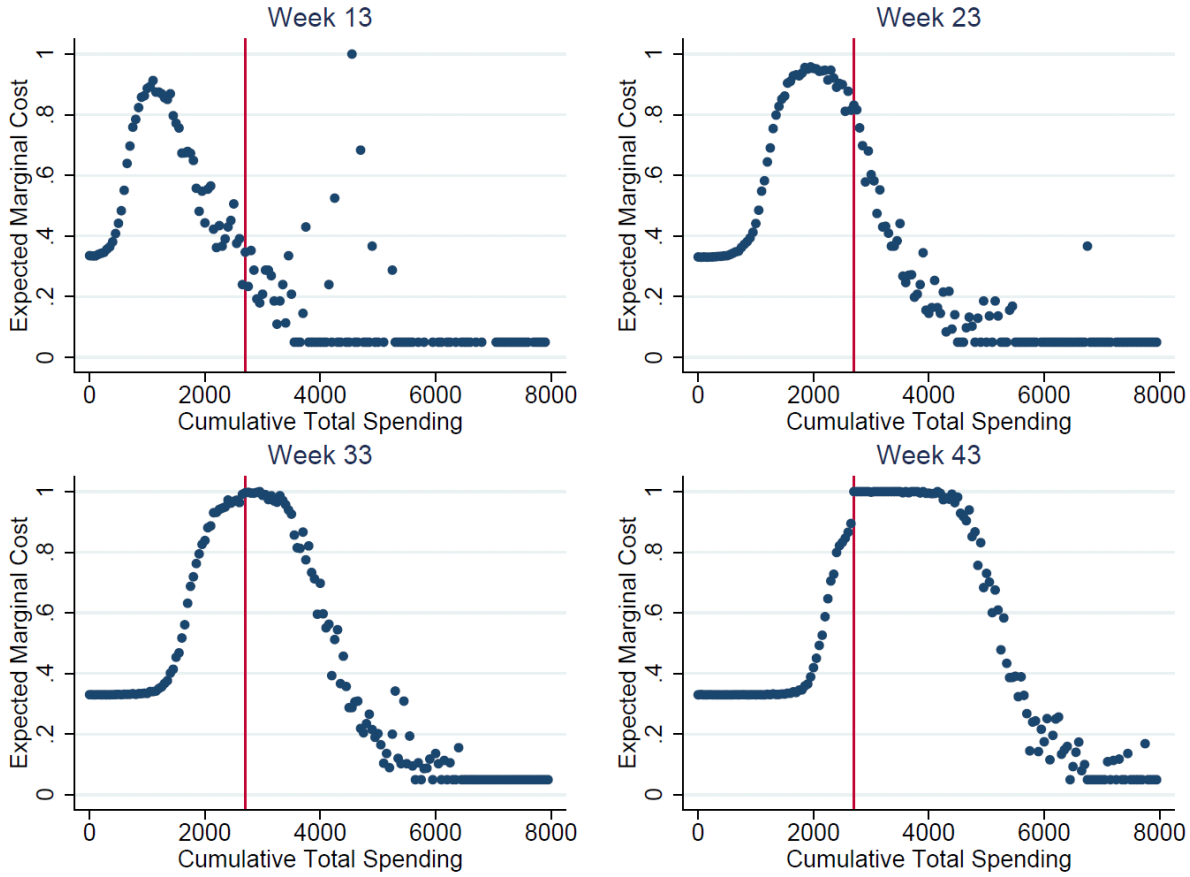
most likely to end the year in the coverage gap, and this translates to her highest HMC being in that region. If a beneficiary's spending is already over $1,500, she is most likely to end the year in the catastrophic zone, even though she is still far away from even the transition from the ICR to the coverage gap. If beneficiaries use the HMP as their perceived marginal cost, she should increase her spending as the HMP decreases, which begins well before the 2009 ICL of $2,700.

As time progresses, the probability of ending the year in any of the regions also shift to higher cumulative totals and tend closer to 0 and 1 and the beneficiary's highest heuristic expected marginal cost more closely resembles the plan spot prices. Over time the cumulative total spending level at which she experiences her highest HMP moves closer to the discontinuity between the initial coverage and the coverage gap regions. While the distribution of the probability of ending the year in the coverage gap was a maximum of 80% in week 13, there is an increasing group of individuals who have spent around $2,000 by week 20 and $2,700 in week 33 who are certain to end the year in the coverage gap. Thus the highest HMP, which should correlate with beneficiary's lowest levels of spending move to about $2000 in week 23 and to the ICL or $2,700 in week 33. In week 43, or approximately 2 months before the end of the year, there already exists a discontinuity in the probability of ending the year in the coverage gap at the initial coverage limit. These graphs are not generated based on the behavior of necessarily standard forward-looking agents, so the HMP may approach the spot price earlier than for standard rational agents.

In translating the beneficiary's perceived marginal cost to their spending patterns, if the demand function is smooth and quantity demanded is decreasing with the marginal price of prescription purchases, broad predictions can be made. The expectation is that within any time period the cumulative total expenditure with the lowest heuristic expected marginal prices in a time period should correspond with the cumulative total expenditure amount that has the highest level of spending. Similarly, within a time period, the cumulative total expenditure with the highest heuristic expected marginal cost should correspond with the lowest levels of spending. If the demand function is smooth, and the transitions in the HMC are smooth (as they are for the majority of the periods), then the spending should also be smooth.

The relationship between the HMC and spending is not necessarily expected to be one-to-one and would not be as dramatic as a simple "flip" of the HMC curve, but the expectations is that the location of spending changes should however correspond with the peaks and troughs of the HMC. In fact, because prescriptions purchases often have immediate and significant health

Figure 5: Heuristic Expected Marginal Price in 2009



Depicts the expected marginal price based off of the objective distributions of the probability of ending the year in each coverage region depicted in Figure 4. The heuristic expected marginal cost is $HMC(Z_w, w)) = F_{ICR}(Z_w|w) * MC(ICR) + F_{Gap}(Z_w|w) * MC(Gap) + F_{Cat}(Z_w|w) * MC(Cat)$, where MC(ICR) = .33, MC(Gap)=1, and MC(C=.05) the government standard plan amounts.

benefits, these drugs may be relatively inelastic goods and respond little if at all to the marginal price changes. Using the end-of-year purchases, Einav et al. (2016) measures the elasticities of different drug classes and find an overall elasticity of -0.037 so that a one percent increase in out-of-pocket cost leads to a 0.037 percent decrease in the probability of filling a claim. Thus, the expectation is that the HMC would result in small changes in the probability of filling a claim as well.

The next section will discuss the empirical approach to estimate a graphical representation of beneficiary spending patterns. It will also discuss how consistent behavior is with the simple heuristic model of spending.

5 Estimation Model and Results

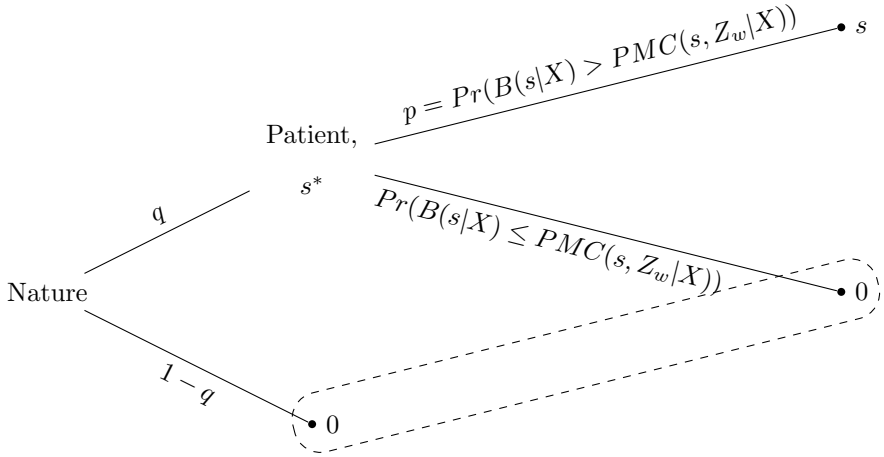
The following section details a simplified model of prescription purchasing behavior under Medicare Part D that ascribes prescription spending to both opportunities to spend and then beneficiary decisions to spend. The model emphasizes the effects of past cumulative spending on weekly prescription purchases. Ultimately the estimation shown in this paper lumps both the opportunities to spend and the decision to spend as one process, but the model is laid out to provide intuition on the beneficiary's underlying motivations. The intent of the simplified estimation is to first document the patterns of consumer behavior that emerge as patients progress through their insurance plans, and not to capture the full dynamics of the consumer prescription choice problem. The beneficiary's prescription spending decision is aggregated and examined on the weekly level to reduce the size of the problem.

The initial model below assumes beneficiaries have a choice to spend on prescriptions and a separate limited choice on the amount to spend. The only observables that are available in Medicare Part D data are prescription purchases and not the beneficiary's direct consumption of drugs. Because of the nature of chronic prescriptions, many are offered in 30 to 90 day refill amounts, which means that there is a periodicity to beneficiary's spending patterns that are not necessarily driven by a choice to spend or not. In order to capture some of this, the model assumes that the personal utility from medications should only apply if a patient has a medical event or shock that requires treatment. Thus, the decision to use Medicare Part D to purchase prescriptions should only occur when those events arrive and a doctor has written the patient a prescription. These events can be temporary health shocks that require treatment such as antibiotics for pneumonia, or continuations of existing conditions that require a prescription refill.

The probability of a medical event occurring q depends upon observables X such as demographics: age, risk scores, historic Medicare Part D usage, etc., and observable environment characteristics such as the time of the year. It may also depend upon unobservable patient characteristics and health shocks. However, essential to the model is the idea that the actual probability of a true medical event occurring should not change due to the arbitrary insurance coverage region R (or the distance to these regions) imposed by the patient's insurance plan coverage. $E(q|X) = E(q|X, R)$.

Once a patient receives a medical event, they have a choice to spend on prescriptions and a choice on the exact amount of out-of-pocket and total prescription spending. Within the second stage prescription filling decision, the patient has some flexibility in the total pre-

scription costs and thus their out-of-pocket costs by choosing branded or generic drugs, or by having the doctor prescribe alternative drugs in a class of drugs that treat their medical shock. I assume that the beneficiary spending, conditional on receiving an event, falls in some truncated distribution. Willingness to consume given a medical shock is s, and the dollar amount of prescriptions patient's actually purchase is censured at 0, and the probability $Pr(s > 0) = q * Pr(B(s|X) > PMC(s, Z_w|X))$. The total payment amount s depends on the patient's observable and unobservable characteristics X, but also depends on the region r of the beneficiary's insurance plan and the distance in spending to them. Changes in a patient's total spending between insurance regions is meant to capture changes in the patient's expectations of or response to to out-of-pocket marginal price changes.

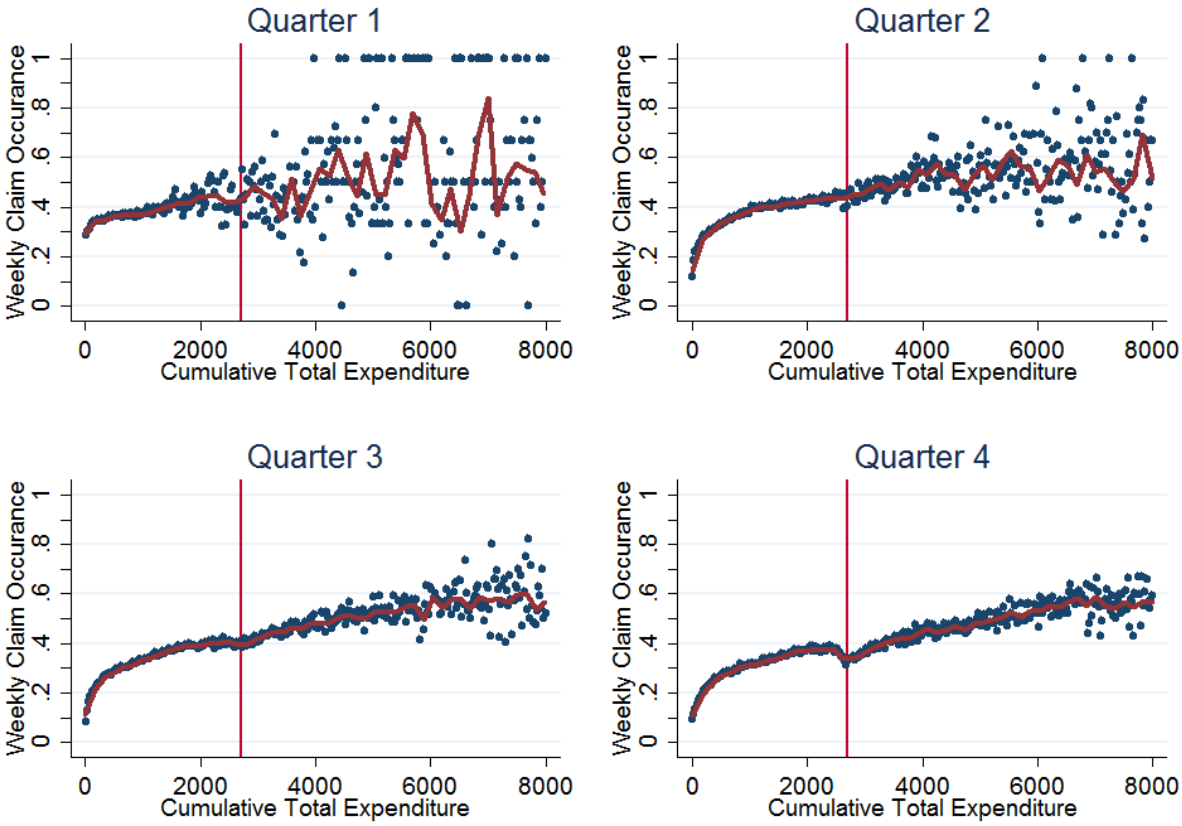


From what has been laid out so far, it is clear why a censured regression spending model such as a Tobit would be inappropriate. There are legitimate zeros when beneficiaries do not have prescriptions to refill or health events, where the benefit of any drug spending is minuscule and the cost would require a doctor's visit for a new prescription.

Rather than estimating the more complicated model of both the choice to spend and spending amount, this paper focuses on the empirical probability of spending occurring in a week versus the probability of observing zero spending $p' = q * p$ the probability of a claim occurring and the beneficiary decides to fill the claim. It is the probability of observing non-zero spending in a week. The probability $1 - p' = 1 - q + q * P(s \leq 0)$ is the probability of observing zero spending either because the beneficiary didn't have a claim, or she had a claim and she chose not to fill the claim.

Figure 6 uses the mean of the indicators of whether beneficiaries had claims in a week, conditions that on the cumulative total spending in every quarter of 2009 to illustrate the raw probabilities p' of beneficiary's spending. Einav et al. (2015) uses a similar graph of the monthly

Figure 6: Weekly Claim Occurrence Conditional on Cumulative Total Expenditure and Spending



Using a $50 dollar bin, the points are the average probability of spending on prescription purchases in a week conditional on the quarter of the year and the cumulative total expenditure. Einav et al. (2015) produced very similar graphs of the probability of a prescription purchase in a month rather than the week level. This image is only of 2009 claims.

probability of spending to help illustrate the empirical patterns that beneficiaries engage in at the coverage gap. However, in order to fully understand beneficiary behavior at and before the kink, researchers must take into consideration the significant amounts of heterogeneity in prescription needs that exist in the Medicare Part D population. Beneficiaries who are more likely to spend (and spend more), are also more likely to be observed in higher cumulative spending bins, while those who spend less are observed at the lower cumulative total expenditure levels.

While Einav et al. (2015) and other papers take a structural approach that imposes substantial assumptions on beneficiary behavior in order to control for beneficiary heterogeneity, this paper takes a reduced form fixed effects approach that differs from the prior literature. Along with a dynamic optimization model Dalton et al. (2015) also run simple linear regressions with fixed effects to examine the flat effect of being within $110 of and in the doughnut hole, on

individual's average weekly spending (and other measures of spend). This paper improves upon on their reduced form approach by applying it to four year panel of observed weekly spending patterns over a much larger number of beneficiaries to reduce the negative dynamic panel bias with fixed effects (see Section 5.2). Also, rather than using a single linear indicator of being near the coverage gap, this paper's analysis allows for a flexible form polynomial to characterize beneficiary's spending patterns. This is described in detail in the following Section 5.1.

Among the reduced form literature, other papers have handled the heterogeneity in other ways. Kowalski (2014) uses a quantile regression with an instrumental variable to analyze the change in a person's healthcare spending when their year-end marginal prices change due to accidental (and assumed exogenous) injuries to family members. Since Medicare Part consists of only individual plans, this paper does not use a similar IV approach. Abaluck et al. (2015) effectively net out individual fixed effects by levering their panel data and observing the difference in individual spending due to plan changes between different years. However, their analysis is purposefully focused on spending for individuals who are unlikely to cross coverage regions and is thus limited at spending kinks. Joyce et al. (2013) compares the difference in the spending patterns of patients who have standard Medicare Part D Plans with plan non-linearities with the patients who receive low-income subsidies (LIS) and thus do not have significant coverage gaps. However, using the spending patterns of the LIS as a baseline comparison group for non-LIS patients may ascribe inherent differences between the groups to the plan coverage structure.

5.1 Estimation

In order to identify how the probability of claims occurring responds to kinks in the marginal prices rather than any cross sectional differences between the heterogeneous beneficiaries, this analysis takes a fixed effects regression approach with a dynamic panel.

Equation 1 is a linear probability model and the main estimation approach for this paper. Suppose the Medicare Part D individual claims are grouped on a weekly level with spending s_{iyw} in year y, and week w across the four years of this dataset. Let the occurrence of spending $o_{iyw} = I(s_{iyw} > 0)$ be a binary variable.

$$o_{iyw} = \alpha_i + \boldsymbol{\gamma}\mathbf{X}_{iy} + f(Q_{iyw}, R_{iyw}, Z_{iyw}) + \tau_y + \varepsilon_{it} \tag{1}$$

where α_i is an individual fixed effect constant across years and nests any gender, race, and age in 2009 information about the beneficiary. $\mathbf{X}_{iy}$ the set of year-varying individual demographics that include beneficiary RxHCC 2011 demographic and risk scores for that year that are based

off of their known health conditions from the prior year. The real life purpose of the risk scores are to reimburse prescription spending, and as such they are an important measure to capture any between year changes in beneficiary's probabilities of spending. τ_y is the coefficient for year y that is the same for all individuals. $Q_w \in \{1, 2, 3, 4\}$, which are simplified to be called quarter variables, indicate whether a week is in the first, second, third, or fourth set of 13 consecutive weeks in a year.

The most important variables of interest are $Z_{iyw} = \sum_{u=1}^{w-1} s_{iyu}$, the measure that represents the cumulative total expenditures within a year t up until week w, and $R_{iyw} \in \{1, 2, 3\}$ an indicator for the region of the insurance contract beneficiary i is in in week w of year y. The values of R—1,2, and 3 correspond to the initial coverage, coverage gap, and catastrophic regions respectively.

The most important variables of interest enter the estimation through f in Equation 2, a piecewise function with a polynomial form describing the effect on the weekly claims probability of the cumulative total expenditures measure interacted with the time of the year and spending region.

$$f(Q_{iyw}, R_{iyw}, Z_{iyw}) = \sum_{q=1}^{4} \sum_{r=1}^{3} \left(I(Q_{iyw} = q) * I(R_{iyw} = r) \left(\eta_{qr} + \sum_{j=1}^{3} \beta_{qrj} Z_{iyw}^{j} \right) \right) \quad (2)$$

This function is used to measure the extent to which beneficiary's respond to the non-linear pricing contract in Medicare Part D, even far away from the marginal pricing discontinuities. Specifically, f is the interaction of the quarter and region terms with a flexible cubic polynomial of Z_{iyw}.

For the polynomial Z terms, coefficients are estimated that differs by the quarter of the year and the coverage region R the beneficiary is in. This functional form allows the slope (and form) of the relationship between cumulative total spending and the beneficiary weekly claim probability to vary separately in each region-quarter grid.

This estimation model assumes that the relationship in f is the same across all four years, with the year effects only altering the level of spending across all beneficiaries through τ_y in Equation 1. It is possible that the levels and slopes of the beneficiary claims response within the coverage gap differ across years, because the beneficiary plans differ across the years. Their marginal costs in 2011 and 2012 coverage gap are significantly closer to the ICR coinsurance levels, and thus this should result in fewer changes in spending. The expectation then by the

estimation presented in Equation 1 and 2 across all four years is that the predicted f will be "too flat" to describe 2009-2010 and not flat enough to describe 2011-12. An extension model could allow for the f function to have separate slopes at the year-quarter-region level, however it was not included here.

Alternative models were considered such as ones that included an individual-year fixed effect for more flexibility. However, such models were rejected in order to mitigate the potential bias with dynamic panels and fixed effects. See the discussion below.

5.2 Bias in a Dynamic Panel with Fixed Effects

While the estimates of fixed effects models applied to dynamic panel are known to be biased if T the number of time periods is small and the cross sectional size of the panel (N) is large (Nickell, 1981), this paper takes a few precautions to reduce the influence of the Nickell bias on the analysis. First, the analysis spans beneficiary behavior over four years or $T = 208$ weeks, a longer time span than normally cited in the literature. As Nickell (1981)'s demonstrates in their dynamic panel with a simple lag, as $N \to \infty$, the inconsistency of the estimated lagged parameter is of the order 1/T. So while the number of beneficiaries in the "No Deductible" sample $N = 89,354$ is large, the potential bias with a larger T is greatly reduced.

This papers analysis is not a direct of the classic Nickell bias, since the Equation 1 is not only a function of its own lag. However, the cumulative total expenditure Z_{iyw} and region R_{iyw} are both functions of the lagged claim observation o_{iyw} and make this analysis a dynamic panel. For a given person and within year iy, $\sum_{u=1}^{w-1} s_u = o_{w-1} * s_{w-1} + o_{w-2} * s_{w-2} + ... + o_1 * s_1$. The region effects are $R_{iyw} = \mathbf{I}((\text{Region lower limit})_{iy} \leq Z_{iyw} < (\text{Region upper limit})_{iy})$, indicates which contract region the beneficiary i faces in year y and week w. In the baseline "No Deductible" sample, there are the initial coverage region (ICR), the coverage cap (Gap), and catastrophic (C) regions. While the spending regions R_{iyw}, conditional on year y, maps directly to the cumulative total spending Z_{iyw} at the initial coverage limit (the upper bound of the ICR) to enter the coverage gap, there is significant between-person heterogeneity on the spending level at which beneficiary's reach their OOPT and enter the catastrophic region. Both Z_{iyw} and R_{iyw} are functions of the lagged dependent variable.

The classic Nickell bias as it applies to the coefficient on the lag of the dependent variable is negative, and the coefficient on the cumulative total spending and region dummies would be similarly negatively biased. To verify, the direction of the bias as it applies to the specific types of lags that are produced here are simulated with noise and presented in the Appendix

Section C. Any bias in the estimates are more likely to be observed at low cumulative spending levels and leads to a more negative slope. The bias alone does not lead to discontinuities in the affect of the cumulative total spending on the probability of spending.

Other estimation approaches that would have circumvented the bias problem of the fixed effects approach such as that presented by Anderson and Hsiao (1982) or Holtz-Eakin et al. (1988) (popularized and more commonly known as Arellano and Bond (1991)) have their own problems. These methodologies use a first-differences approach to net out the fixed-effect α_i and then use further lags of the lagged variable (in this case the Z_{iyw} and R_{iyw}) as instruments in a 2SLS and GMM style estimation respectively. While not biased, Anderson and Hsiao (1982) does not use all available data and can result in imprecise estimates. Further, because of the large size of the data, the number of interactions of Z_{iyw} and R_{iyw} terms, and the length of the panel, the Arellano-Bond methods become computationally intractable.

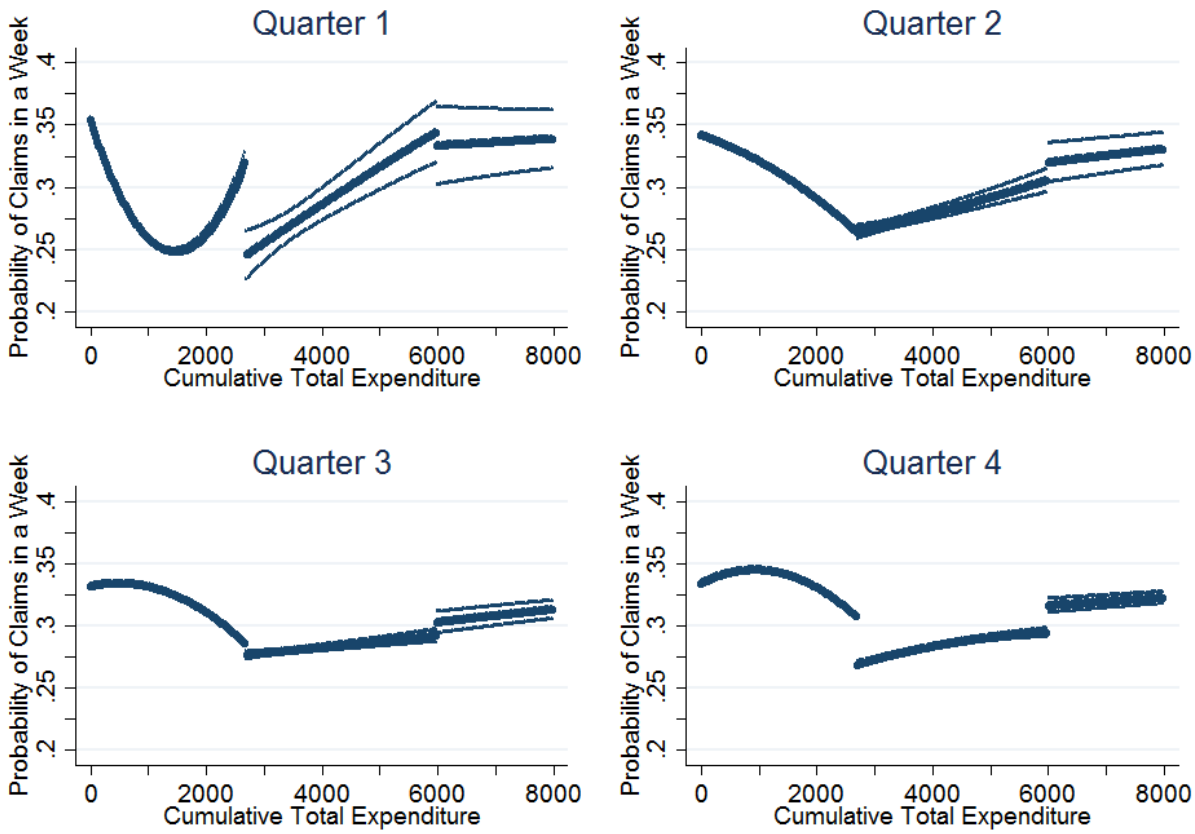
5.3 Baseline Estimation Results

The results of the linear probability fixed-effects estimation of Equation 1 and 2 are displayed in Figure 7 and Table 7. Errors are clustered at the individual level. Rather than displaying the coefficient estimates of the piecewise f function, the figure displays the predicted values from those estimates that describe the effect that the specific coverage regions have on the probability of filling a prescription claim in a week. The remaining coefficient estimates for the beneficiary risk scores and year (γ and τ_y) are included in the table.

Overall the estimates from the linear probability model are highly statistically significant, and the 95% confidence interval bounds on the figure are incredibly tight. These estimates may be tight due to the large number of individual and claims observations and the rich set of individual fixed effects. However, in the robustness checks that use logit and Poisson regression models (Section 5.4), while the estimates are similar, the 95% CI are tremendously wider, so this paper is cautious about the statistical significance of these results. Further, perhaps in a reflection of the inelastic nature of prescription purchases, the effect on the probability of a claim in a week may not necessarily be economically significant.

The directions of the coefficients of individual risk and demographics scores are highly significant, but the scale of the effects may be economically small. On average for the predicted sample, the point estimates of the probability of a claim occurring in a week all fall in the range of around 25-35%. These point estimates translate into economic meaning of making claims every 4 weeks versus making claims every 2.9 weeks for a difference of only a little over a week

Figure 7: Linear: Probability of Claims Occurring in a Week, Conditional Year-Cumulative Total Spending



The predicted values of the f function from a fixed effects panel regression of Equation 1 on the beneficiaries in the "No Deductible" sample. Each panel represents a quarter of the year and displays the estimated probability of having at least one prescription claim in a week, conditional on the cumulative total spending in a year. Each quarter consists of 13 weeks except for quarter 4, where the last "week" of the year consists of the remainder 8 or 9 days of the year. The probabilities are predicted assuming that the fixed effect is zero and the beneficiary has the sample average risk and demographic scores from 2009 .4753 and .4196 respectively. Images display a 95% confidence interval around the predicted values. Each line segment represents a prediction made assuming the year is 2009 with 2009 spending limits and across time average individual risk and demographics. The discontinuity for the OOPT into the catastrophic region was chosen at $6,153.8 the total expenditure equivalent OOPT from the government-defined 2009 plan (Table 1).

between claims. These differences would be higher for individuals with lower risk scores and the differences would be smaller for high-risk types.

Referencing Table 3, the standard deviation in risk scores in this sample is approximately 0.3, so for a standard deviation increase in the risk score, the probability of a claim being observed in a week is expected to increase by approximately 1 percentage points. The standard deviation of the demographic weight is even smaller at 0.01, so a standard deviation increase in the demographics

Table 7: Effect of Cumulative Total Expenditures on the Probability (%) of a Claim in a Week

Coefficient	Estimate	Std. Error
RxHCC Risk Weight	3.60	0.11
RxHCC Demographic Weight	-25.42	6.13
2010	0.32	0.036
2011	0.66	0.044
2012	1.16	0.05
N	18,585,632	

The estimated coefficients on the risk scores $\mathbf{X}_{iy}$ and year-time dummies τ_y from the fixed effects panel linear probability regression of Equation 1 on the beneficiaries in the "No Deductible" sample. All estimates in the table are significant at the 1% level.

score results in an expected 0.25 percentage point drop in the probability of observing at least one claim in the week. While the demographics score is a function of an individual's fixed effects characteristics, it is a nonlinear function and thus still included in this analysis as beneficiaries change over time. At first glance the negative coefficient on the demographics score appears counter intuitive, because these scores are used by the administration to reimburse prescription usage, which should be positively correlated with the probability of observing claims. However, these scores are defined across individuals in conjunction with the risk scores to adjust payments, and the demographics score actually decreases as age increases. In the analysis that controls for fixed effects, the demographics score having a negative coefficient essentially means that there is a positive effect of age on the probability of having claims in a week. The positive and significant coefficients on the year fixed effects also capture the effect of aging on the sample population and any year trends that lead to increases in medication purchasing frequency.

Figure 7 depicts in four panels the estimated probability of a claim occurring in a week in each of four quarters (13 week periods) in the year. Within each panel, the f is displayed as a piecewise function with the first segment representing the initial coverage region, the second the coverage gap, and the third the catastrophic region. The panels do not include individual fixed effects in the weekly claim probability over a four year period. Further these estimates are the predicted values from the regression assuming the year is 2009 where beneficiaries have the average demographics and risk scores and face the 2009 ICL and OOPT limits. Because the OOPT limit is in terms of the cumulative patient spending rather than total spending, the predicted value uses \$6,153.8, the cumulative total expenditure equivalent OOPT from the government-defined 2009 plan (Table 1). While the predicted values would be different across years and different levels of cumulative total expenditure equivalent OOPT cumulative total expenditure equivalent, these differences would result in only minor qualitative differences in

the beneficiary's responses.

In order to analyze the predicted f, it is important to analyze both the beneficiary's response to the coverage limits (within a panel) and the beneficiary's response through time (across the four panels). Within a panel, an especially important limit that will be discussed is the initial coverage limit, where beneficiaries transition from the initial coverage region to the coverage gap (i.e. the "doughnut hole"). In comparing across panels, there is a striking difference between the response in quarter one relative to the other panels. The existence of these differences between the panels are consistent with the existing literature in Einav et al. (2015) and Aron-Dine et al. (2015), which showed the response to changes in the coinsurance rates differ depending on the amount of time left in the plan.

Broadly, the beneficiary response in the first quarter to the initial coverage limit exhibits forward-looking behavior that has been thus far not documented in the literature. At the same time, the drop in the probability of prescription purchases at the ICL that has already been established through raw data plots in the literature persists (Einav et al., 2015; Dalton et al., 2015).

First, in quarter one, there is a sharp decrease in the probability of a claim when crossing from the ICR to the coverage gap. The point estimate difference is approximately an 8% drop in the probability of spending which might be economically significant if it were sustained over a long period of time. While the coefficients are tightly estimated here, the CI may be driven by the functional form estimate. Given the amount of noise in the raw probabilities in Quarter 1 from Figure 6 and 4 and the relatively large CI bounds in the coverage gap segment, more work may be needed to be confident of the significance of the estimate at the boundary. The estimate is suggestive that at least some people are reaching the gap early in the year and pausing. While this paper uses a different estimation approach, the qualitative drop in the probability of spending at the kink is consistent with what was found in Dalton et al. (2015). Further, like in that paper, the drop in spending is not consistent with either the standard or the heuristic approach to determine spending in the first quarter of the year. Based on the heuristic approach, individuals who have accumulated up to \$2,700 in cumulative total expenditures should be highly likely to end the year in the catastrophic region. It is possible that this empirical evidence is the result of the heterogeneity in combining types—combining individuals who have foresight and respond to lower expected marginal costs due to the high probability of ending in the catastrophic region with individuals who respond to the actual spot price.

The slope of the estimated f in quarter one in the ICR region supports the hypothesis that there are individuals who may be aware that they should expect a lower future price and thus have a higher claims frequency. In this region the f is concave. This means that at low cumulative total spending levels, there is a downward slope which suggests that even at the beginning of the year, some beneficiaries who have spent less than $1,750 realize they have a potential or high likelihood of reaching the coverage gap and lower their probability of spending. Others who have spent a more substantial amount,[11] may realize that they are likely to end the year in the catastrophic region and increase the frequency of claims.

In comparing the behavior in quarter one to the other periods, the response at the ICL differs in the first period compared to the other. While there is a large discontinuity at the ICL in the first quarter, it is nonexistent in the second and third and nowhere near statistically significant in the fourth. Based on the simulations from Section 2, theory predicts that any drop in the probability of spending is greatest in the last time period, because there is less ambiguity as to the beneficiary's end-of-year marginal costs. Supporting the hypothesis laid out in Section 4 a discontinuity if any occurs in the later part of the year, this drop is certainly larger in Quarter 4 than any discontinuities in Quarter 2 or 3, and it is significantly estimated. The difference in the probability of spending at the ICL kink is approximately 4%, which on a weekly spending level may not be economically significant.

In addition, the response to approaching the spending regions differs within the ICR region as the estimated f transitions from concave in quarter one to increasingly convex from quarter two through four. In quarter four, while the 95% confidence interval around the estimate is large, the slope within the ICR region appears to be convex. At the end of the year, when beneficiaries would have more information on their end-of-year region, the expectation should be that beneficiaries on the lower end of the ICR may have constant or increase spending (expecting to end within the ICR), while those on the higher end may decrease (avoiding the coverage gap). The "humped" shape of the estimated f in quarter four is consistent with that hypothesis. The comparison across time periods in the ICR region is similar to the analysis in Aron-Dine et al. (2015) as they compare the initial spending behavior of beneficiaries of company insurance plans who join the companies (and thus the insurance plans) at different times of the year (with the same end date).

The slopes and spending discontinuities of the f function from the coverage gap and catas-

[11] While the image is constructed over an average of all individuals in the first quarter across all four years, the cumulative total amount at the minimum of f within the ICR corresponds with the rise in the probability of ending the year in the catastrophic region in Figure 4 Quarter 12.

trophic regions do not differ significantly between the four periods. There is an upwards slope of the line segment in the coverage gap in all four quarters, and it is steepest (though not significantly) in quarter one. In quarter one, this upwards slope is potentially a continuation of the upwards response seen in the ICR and may be indicative of beneficiaries who are more and more certain to end the year in the catastrophic region. Across the time period quarters, the slope of the claims probability function flattens. This flattening is also consistent with the hypothesis, because as time increases and given a fixed cumulative total expenditure amount lower than the OOPT limit, the probability of entering the catastrophic region decreases. Beneficiaries who are in the coverage gap in the later quarters are less likely than those in earlier regions to update their perceived marginal price to the low coinsurance rate in the catastrophic region. Further, as beneficiaries enter the catastrophic region, the slope levels out. Again, this response is consistent with expectations, because within the catastrophic region the beneficiary's perceived marginal price is certainly equal to the spot price and is no longer changing.

In analyzing the f function estimates, it is important to remember that while the fixed effects may "net-out" the heterogeneity in the probability of spending between individuals, the predicted f function effectively averages the potentially heterogeneous response to the marginal cost kinks. The heterogeneity in response is probably most evident in the first quarter of the year. While the majority of observations in the first quarter are for individuals who have cumulative total expenditure levels within the initial coverage region, there are a substantial number of beneficiaries who have reached even the catastrophic region.

There are many sources of heterogeneity that may be affecting these estimates. Further study could help clarify whether or not the predictions here may be true in empirical data. First, experience with the Medicare Part D pricing schedule should have a significant impact on individuals expectations of their end-of-year marginal price. Individuals with more experience (and no surprises) would be expected to have more constant spending patterns that result in a flatter overall f response curve.

There are a few reasons why there is a kink at the ICL. Individuals with less experience with Medicare Part D may be less able to appropriately estimate their end-of-year region and thus their perceived marginal price is off. They may be more prone to being "surprised" when reaching the coverage gap and contribute to the drop in the probability of spending at the ICL in the first period. Additionally, beneficiaries who reach actual budget constraints may also reduce their spending at the ICL. However, I would expect that there are individuals who are "surprised" and reach budget constraints to exist in all time periods and not just the first

quarter. The response of these individuals may be more pronounced in period one because there are fewer overall individuals who reach the ICL within the first 3 months of the year. And in the later periods, when these surprises occur, there are more flat spenders as a proportion of the population. A type analysis may be helpful to quantify the extent of individuals who may be more constant spenders versus those who react to the coverage gap.

In order to improve the estimates, adding more years of data could help to reduce the dynamic panel bias; however that strategy may result in even further overestimates on the year fixed effects or risk scores. Certainly over a long time period, the net trend is for beneficiaries to get sicker. Since that trend would not be captured in the individual fixed effect, it may be attributed to year effects (which also capture general trends in prescription update) or the risk scores that typically see an increase whenever individuals are diagnosed with a new health condition.

Another drawback of using the probability of spending rather than the actual amount spent is that there is a significant amount of claims information that is unused in this analysis. Further, there are often multiple claims of different drugs within a week, and if the beneficiary's decision was to drop one specific prescription within that week, this analysis would fail to make a distinction between that scenario and one where the frequency of claims remained constant. In order to alleviate some of these concerns, a Poisson regression model on the number of prescription claims per week is analyzed in Section 5.4 to similar results. This paper does not analyze the effect that the insurance contract has on the total spending amount in part because any such mean analysis could conflate the separate processes that drive the probability of spending within a week and the probability of switching to generics.

5.4 Robustness: Logit and Poisson

While the linear probability model is often adequate, this paper also considered the logit regression model with fixed effects that resulted in extremely similar point estimates. This is in part because logit probability is relatively linear from .2 to .8 and the linear model is an adequate approximation of this.

5.4.1 Logit

Recall that the occurrence of a claim within a week $o_{iyw} = I(s_{iyw} > 0)$ a binary variable. Assume instead of a linear probability model that the probability of a claim occurring in a week follows a Bernoulli distribution. The resulting equation is similar to the linear probability case

Equation 1.

$$o_{iyw} \sim Bernoulli(p_{iyw})$$

$$logit(p_{iyw}) = \alpha_i + \gamma \mathbf{X}_{iy} + f(Q_{iyw}, R_{iyw}, Z_{iyw}) + \tau_y \tag{3}$$

Overall the estimation of the logit[12] model produces strikingly similar point estimates to the linear probability model; however standard error estimates are substantially larger in the logit regression case. As in the linear case, the standard errors are also clustered on the individual level. Figure 8 presents the predicted f function, and the 95% confidence intervals are approximately 10% across the board. The latent function estimates for the risk score, demographic weights, and year fixed effects are presented in the top panel of Table 8.

Table 8: Effect on the Latent Function for the Probability of Weekly Spending

Coefficient	Estimate	Standard Deviation
RxHCC Risk Weight	0.175	0.005
RxHCC Demographic Weight	-1.434	0.315
2010	0.0174	0.0019
2011	0.0349	0.0023
2012	0.0604	0.0026
Marginal Effects		
RxHCC Risk Weight		0.0397***
		(19.71)
RxHCC Demographic Weight		-0.325***
		(-5.55)
N		18,585,632

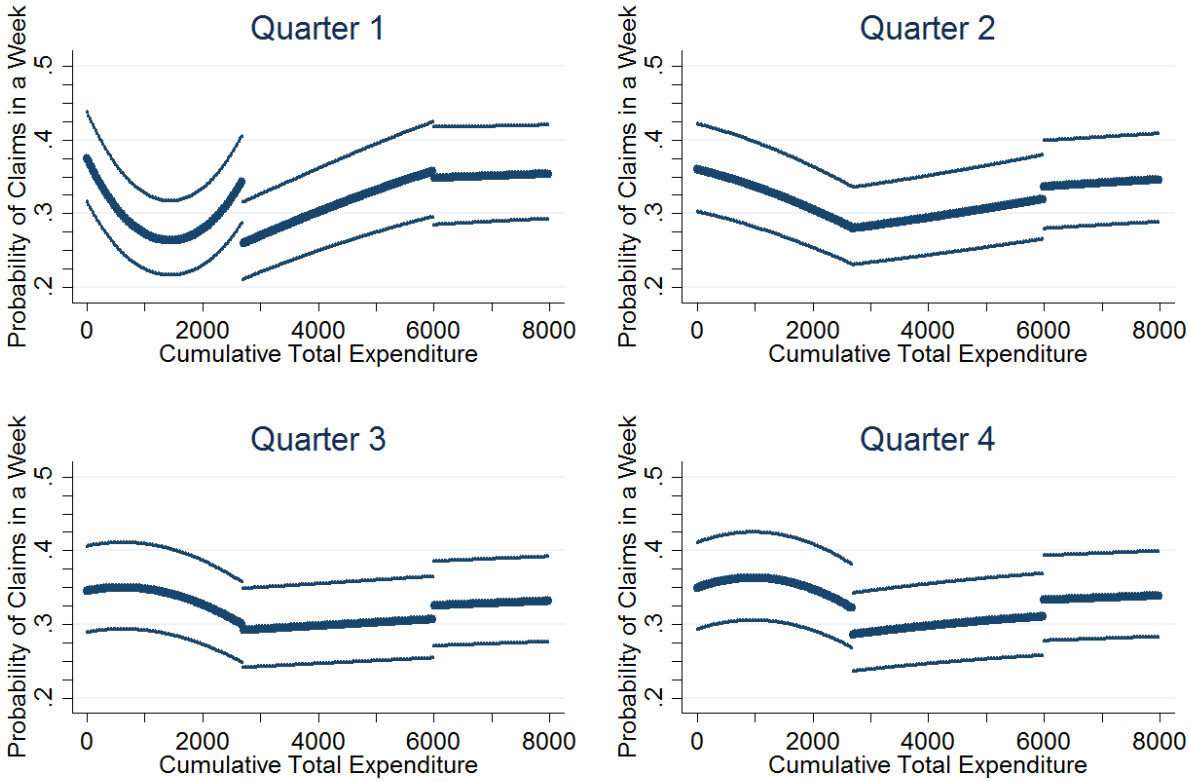
t-statistic in parenthesis. * p<0.05, **p<0.01, *** p <0.001

The estimated coefficients on the risk scores $\mathbf{X}_{iy}$ and year-time dummies τ_y from the fixed effects panel regression of Equation 3 on the beneficiaries in the "No Deductible" sample. The estimates for the latent function are all significant at the 1% level. The Stata margins command run after a `clogit` command was used to create the bottom panel. The standard deviations and p-values were automatically generated. It is curious that the standard errors in parenthesis are very large but the margins returned a significant p-value and something that will be followed up on.

The marginal effect of the risk score and demographic weights are presented in the bottom panel.[13] The estimate for the risk score is statistically significantly higher than from the linear case (comparing 3.6% to 3.97%) but does not result in an economically meaningful probability of spending. For a standard deviation change of the risk score (approx. 0.3), the probability of spending increases by approximately 1.2% according to the logit regression instead of 1% via

[12] This is actually implemented as a conditional logit regression so as to avoid the incidental parameters problem.
[13] The margins here are presented as a probability and not as a percentage as in Table 7

Figure 8: Logit: Probability of Claims Occurring in a Week, Conditional Year-Cumulative Total Spending



See additional information in the note from Figure 7.

the linear. The coefficient on the demographic specific weight is also more negative in the logit regression but is similarly unlikely to be economically significant since a one standard deviation change in this weight (approx. 0.01) results in a 0.325% decrease in spending as opposed to a 0.25% decrease.

The estimation's predicted values of the f function from the logit regression[14] are presented in Figure 8 and have strikingly larger bounds. It is an open question in this paper's analysis as to why the standard errors are substantially larger in the logistic regression case from the linear model. The initial bootstrap errors (50 draws) of the linear probability model did not result in significantly different error estimates from the original estimation presented in Section 5.1 and did not resemble the logit regression results. Further work will include more draws for the bootstrap and other methods of estimating these equations using the Mundlak (1978) approach.

The 95% confidence interval leads to a slightly different interpretation of the logit regression

[14] Using Stata's `clogit` command.

results than the linear probability results. The changes in the slopes of the probability of filling claims in a week and the discontinuities at the coverage gap are no longer significantly different from each other.

5.4.2 Falsification Test

In order to more fully investigate whether or not the discontinuity in quarter 1 at the initial coverage limit really exists is to do falsification tests at other points. One such falsification test using a linear probability model with fixed effects indicate that this discontinuity may not necessarily actually occur and does indicate that the small standard errors are driven off of estimates at other regions of the curve. The following linear probability falsification test was conducted where the initial coverage region is broken down into 2 separate regions $ICR1$, $ICR2$ where

$$ICR1 = \mathbb{I}(Z_{iyw} < ICL_y - 1000) \text{ and } ICR2 = \mathbb{I}(ICL_y - 1000 \leq Z_{iyw} < ICL_y)$$

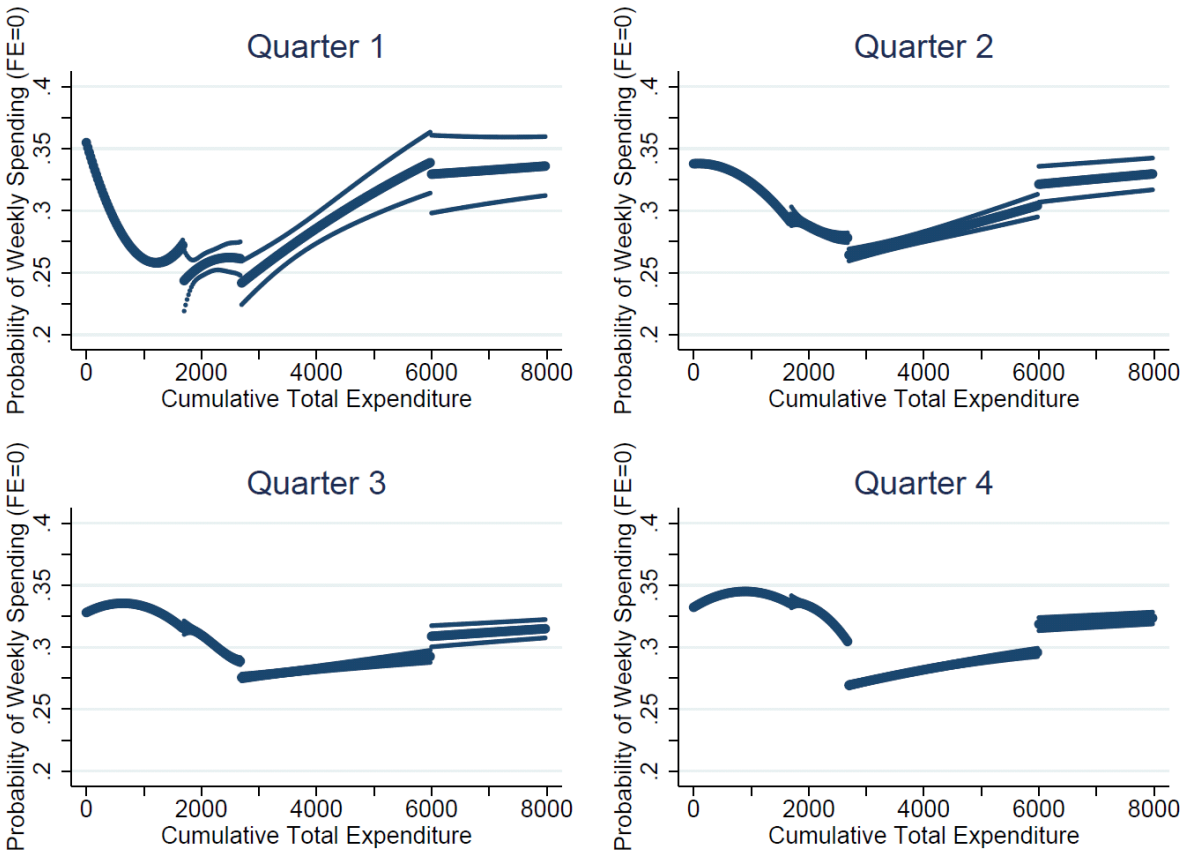
Then the new coverage region definitions that enter the f function are $R'_{iyw} \in \{1, 2, 3, 4\}$ regions that correspond to an ICR1, ICR2, coverage gap, and catastrophic regions. Equation 1 is again estimated with standard errors clustered at the individual level and the estimated f function is displayed in Figure 9.

While the falsification test did not result in any obvious discontinuities in quarter 2-4, it did result estimate changes in the area around the initial coverage region in quarter 1. The suspicion that the error bars in the quarter 1 panel around the ICL were misrepresented by the functional form are confirmed, and it seems likely that the estimates were being driven by the spending behavior at lower cumulative spending levels and the functional form. The drop in spending at the quarter 1 ICL still exists at the point estimate level, but is substantially smaller because the increase in spending prior to entering the coverage gap is much lower.

The falsification test did change some of the estimates for the spending patterns in quarter 2-4 as well; however they remain quantitatively very similar to the estimates from the original specification. The main difference is that the transition towards lower spending the end of the initial coverage region in quarters 2-3 appears more gradual than the Figure 9 suggests.

Overall, because the point estimates are incredibly similar between this specification and the original linear probability regression, there is still little economically significant differences throughout these images on the probability of claims occurring during a week.

Figure 9: Linear Falsification: Probability of Claims Occurring in a Week, Conditional Year-Cumulative Total Spending



See additional information in the note from Figure 7.

5.4.3 Poisson

One drawback of both the logit and linear probability regressions presented here is that the dependent variable is the occurrence of any claim in a week rather than the actual number of claims or the total dollar amount spent on claims. This regression studies the beneficiary's response to the contract schedule in filling any claim in a week. However, given that these beneficiaries often have multiple conditions that require prescription drugs, beneficiaries may have multiple prescription claims that they have the choice to fill in a week. So observing the occurrence of any claim in a week would understate the beneficiary's decisions to fill a claim or not.

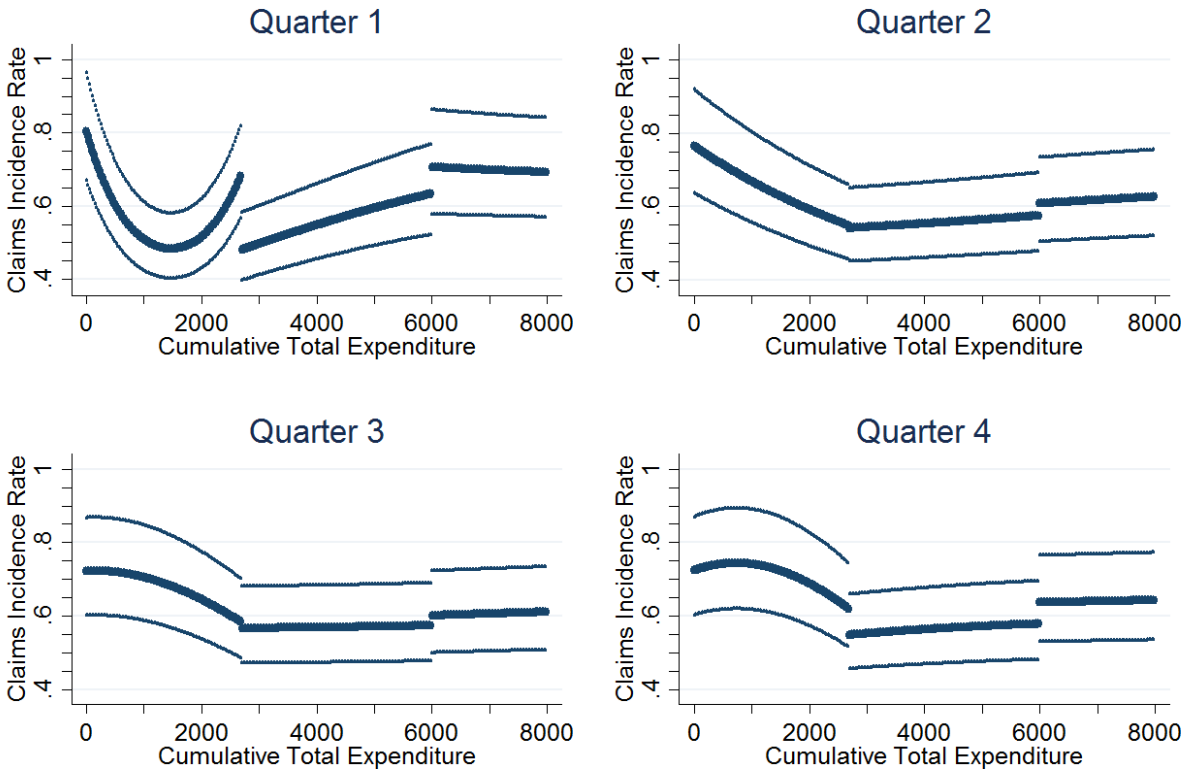
An alternative specification uses the number of claims observed in a week n_{iyw} as the dependent variable assuming that the model fits has a Poisson link function. The fixed effects Poisson

regression faces the same critique due to dynamic panel bias as the prior two regressions, and will be taken into consideration in the analysis.

Let ir_{iyw} be the incidence rate such that $n_{iyw} \sim Poisson(ir_{iyw})$. Then the fixed effects Poisson regression in Equation 4, where the f function retains the same functional form as before.

$$log(R_{iyw}) = \alpha_i + \gamma \mathbf{X}_{iy} + f(Q_{iyw}, R_{iyw}, Z_{iyw}) + \tau_y \tag{4}$$

Figure 10: Poisson: Incidence of Claims Occurring in a Week, Conditional Year-Cumulative Total Spending



See additional information in the note from Figure 7.

The general shape of the estimates is very similar to those estimated from the linear and logit regressions. The level of claims incidence are much higher, which reflect the fact that many beneficiaries file more than one claim a week. The confidence intervals are wider than those from the logit regression averaging about a 0.15 band around the estimates. The point estimates range from approximately 0.5 to 0.8, translating to a claim occurring every 2 weeks (14 days) to every 8.75 days.

Similar to the probabilities from the linear regression, the incidence rate varies across the different time periods and cumulative total spending amounts. Individuals in quarter 1 who having spent approximately $1,500 cumulatively, still seem to ramp up their spending prior to entering the coverage gap. As predicted by the heuristic expected marginal cost, the lowest claims occurrence shifts from $1,500 cumulative spending in quarter 1 to the coverage gap. It also appears as if on average individuals in the coverage gap in quarter 1 increase their claims rates as cumulative total spending increases, possibly anticipating the lower marginal costs in the catastrophic region.

Similar to the linear probability estimation, the discontinuity in incidence claim rates at the beginning of the coverage gap occur in quarter 1 and in quarter 4, where the quarter 1 discontinuity was not predicted. Again, knowing that there are few individuals and observations in the beginning of the year coupled with the large confidence intervals it is difficult to make a substantial claim as to the beneficiary behavior here. The point estimate difference between the spending at the end of the initial coverage region and the beginning of the coverage gap appears to be about 0.2 which is quite large. However, based on the low observations in this region and the results from Section 5.4.2, this difference may also be due to the functional form assumptions.

The incidence of claims is highest in every one of the periods at the beginning of the year and decreases and steadies at higher spending levels. This trend may be due to the fact that there are substantial numbers of typically low spending beneficiaries who receive acute prescription shocks rather than chronic period prescriptions that they constantly need to refill. A concern with that interpretation is that that is precisely the pattern that dynamic panel bias would generate. Further analysis using either a longer sample or some other method described in Section 5.2 may be necessary.

6 Conclusion

This paper builds on the existing literature on beneficiary's dynamic response as they approach the many discontinuities in the Medicare Part D pricing structure. Part of the motivation for this work is that throughout all health insurance, the government and insurers have significant control over the cost-sharing features that are responsible for non-linear pricing, and with the rise in health care costs, these institutions are more likely to use them as cost control measures. Unfortunately, the effect of these cost control measures on beneficiary behavior and health is not fully explained. While the literature has identified sharp drops in spending particularly at the

Medicare Part D coverage gap, trying to explain this behavior using time-discounting models have resulted in discounting estimates far lower than the broader economics literature.

The first main contribution of this paper is a through its discussion of an expected price model that uses the objective probability distributions of beneficiaries' end-of-year prices given their spending probabilities. The key takeaway from this heuristic is that if beneficiaries know the objective probabilities of ending the year in each region for the population, the expected marginal price that a beneficiary responds to can be constructed and be used to make purchasing decisions. Because the end-of-year region probabilities have relatively smooth transitions (in all but approximately the last 10 weeks of the year), beneficiaries' marginal price and then spending should also be smooth (except for the last weeks). Even if beneficiaries have inaccurate beliefs on their objective end-of-year probabilities or are present-biased, as long as they update those beliefs in each time period, a heuristic marginal price would not generate sharp spending changes unless there were sharp changes in probabilities.

Another significant contribution of this paper is the graphical representation of beneficiaries' claims rates conditioning on the cumulative sums of their total spending. Using separate linear probability, logit, and Poisson regressions with individual fixed effects to control for individual heterogeneity in their base levels of spending, this paper illustrates beneficiary claims rates in a way that allows direct visual comparison with their heuristic marginal spending.

The estimation finds that there are significant changes to beneficiary's claims rates, some of which are consistent with the predictions of the heuristic marginal price, but that those changes may not be economically significant. The lowest amount of claim rates in each quarter of the year broadly matches the cumulative total expenditures values that produced the lowest expected marginal prices. The changes in claims rates for a predicted beneficiary with average risk and demographic scores indicate that the economic magnitude of the claims occurrences are on the scale of filing claims every 3 weeks versus filing claims every four weeks. Further work to understand the welfare consequences of these reductions would be to analyze whether the claims changes were by discontinuing drugs entirely or just small delays in going to the pharmacy for refills.

This papers findings differ from prior literature's empirical results and may be due in part of the sample selection. Because this sample involved individuals who retained similar plan structures through all four years, they are mechanically more likely to have experience with Medicare Part D, their plan structure, and their prescription needs than individuals in the papers mentioned in the literature. Further, the individuals who do not switch between plans

with different limits may be different onto themselves as people who have high inertia and do not switch plans or just happen to have have expected spending amounts that align well with their chosen plans. Further work could explore whether experience with Medicare Part D promotes individuals to exhibit more optimal behavior, because that would imply that informational and educational programs could promote that behavior. A more detailed subsample analysis may be warranted.

This paper also included estimation complications that require more research to confirm the validity of the coefficient estimates and predicted probabilities particularly in the quarter one coverage gap. The role of heterogeneous types of responses to the coverage gap should also be considered in future work. While this paper controlled for heterogeneous levels of claim rates for individual beneficiaries, it is likely that due to random health shocks or experience with the Medicare Part D pricing schedules, individuals may separately increase or decrease their claims in a predictable way that adds noise to this papers estimates.

Another challenge to this paper's research question was the large variety of plans and coinsurance rates offered from 2009-2012 and the policy changes introduced by the Affordable Care act to fill in the donut hole. The lowered coinsurance rates within the donut hole in 2011 and 2012 could be partially responsible for the more stable spending estimates that are found throughout this paper. An obvious related research project would be to study the actual impact of the Affordable Care Act's policy of filling in the coverage gap and subsequent health outcomes to determine whether it truly impacted beneficiary spending rates. The plan variety that was a challenge for the analysis in this paper, would be a boon for follow-up research projects.

References

(2016). The Medicare Part D Prescription Drug Benefit. Technical report, The Henry J. Kaiser Family Foundation.

Abaluck, J. and Gruber, J. (2016). Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance. *American Economic Review*, 106(8):2145–2184.

Abaluck, J., Gruber, J., and Swanson, A. T. (2015). Prescription Drug Use Under Medicare Part D: A Linear Model of Nonlinear Budget Sets. SSRN Scholarly Paper ID 2572135, Social Science Research Network, Rochester, NY.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82.

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

Aron-Dine, A., Einav, L., Finkelstein, A., and Cullen, M. (2015). Moral Hazard in Health Insurance: Do Dynamic Incentives Matter? *Review of Economics and Statistics*, 97(4):725–741.

Baicker, K., Mullainathan, S., and Schwartzstein, J. (2015). Behavioral Hazard in Health Insurance*. *The Quarterly Journal of Economics*, page qjv029.

Baker, D. (2006). The Origins of the Doughnut Hole: Excess Profits on Prescription Drugs. Technical Report 2006-19, Center for Economic and Policy Research (CEPR).

Dalton, C., Gowrisankaran, G., and Town, R. (2015). Myopia and Complex Dynamic Incentives: Evidence from Medicare Part D. *Unpublished*.

Einav, L., Finkelstein, A., and Polyakova, M. (2016). Private Provision of Social Insurance: Drug-Specific Price Elasticities and Cost Sharing in Medicare Part D.

Einav, L., Finkelstein, A., and Schrimpf, P. (2015). The Response of Drug Expenditure to Nonlinear Contract Design: Evidence from Medicare Part D. *The Quarterly Journal of Economics*, 130(2):841–899.

Ho, K., Hogan, J., and Morton, F. S. (2015). The Impact of Consumer Inattention on Insurer Pricing in the Medicare Part D Program. *Working Paper*.

Hoadley, J., Summer, L., Hargrave, E., and Cubanski, J. (2011). Understanding The Effects of The Medicare Part D Coverage Gap in 2008 and 2009. Technical Report 8221, The Henry J. Kaiser Family Foundation.

Holtz-Eakin, D., Newey, W., and Rosen, H. S. (1988). Estimating Vector Autoregressions with Panel Data. *Econometrica*, 56(6):1371–1395.

Joyce, G. F., Zissimopoulos, J., and Goldman, D. P. (2013). Digesting the doughnut hole. *Journal of Health Economics*, 32(6):1345–1355.

Kowalski, A. E. (2014). Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care. *Journal of Business and Economic Statistics*.

MedPAC (2016). Prescription Drugs. In *Health Care Spending and the Medicare Program*, pages 170–171.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, 46(1):69–85.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, 49(6):1417–1426.

Zhang, Y., Donohue, J. M., Newhouse, J. P., and Lave, J. R. (2009). The Effects Of The Coverage Gap On Drug Spending: A Closer Look At Medicare Part D. *Health Affairs*, 28(2):w317–w325.