INATTENTION AND SWITCHING COSTS AS SOURCES OF INERTIA IN MEDICARE
PART D

Florian Heiss
Daniel McFadden
Joachim Winter
Amelie Wuppermann
Bo Zhou

Inattention and Switching Costs as Sources of Inertia in Medicare Part D
Florian Heiss, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou
NBER Working Paper No. 22765
October 2016
JEL No. C25,D12,G22,I13

ABSTRACT

The trend towards giving consumers choice about their health plans has invited research on how good they actually are at making these decisions. The introduction of Medicare Part D is an important example. Initial plan choices in this market were generally far from optimal. In this paper, we focus on plan choice in the years after initial enrollment. Due to changes in plan supply, consumer health status, and prescription drug needs, consumers' optimal plans change over time. However, in Medicare Part D only about 10% of consumers switch plans every year, and on average, plan choices worsen for those who do not switch. We develop a two-stage panel data model of plan choice whose stages correspond to two separate reasons for inertia: inattention and switching costs. The model allows for unobserved heterogeneity that is correlated across the two decision stages. We estimate the model using administrative data on Medicare Part D claims from 2007 to 2010. We find that consumers are more likely to pay attention to plan choice if overspending in the last year is more salient and if their old plan gets worse, for instance due to premium increases. Moreover, conditional on attention there are significant switching costs. Separating the two stages of the switching decision is thus important when designing interventions that improve consumers' plan choice.

Florian Heiss
University of Duesseldorf
LS Statistics and Econometrics
Universitaetsstrasse 1, Geb. 24.31
40225 Düsseldorf
Germany
florian.heiss@hhu.de

Daniel McFadden
University of California, Berkeley
Department of Economics
508-1 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
mcfadden@econ.berkeley.edu

Joachim Winter
Department of Economics
LMU Munich
Ludwigstr. 33
D-80539 Munich
Germany
joachim.winter@lrz.uni-muenchen.de

Amelie Wuppermann
Department of Economics
LMU Munich
Ludwigstr. 33
D-80539 Munich
Germany
amelie.wuppermann@lrz.uni-muenchen.de

Bo Zhou
University of Southern California
635 Downey Way
Los Angeles, CA 90089-3331
zhoub@usc.edu

1 Introduction

The recent trend towards giving consumers more choice about their health plans has invited research on how good they are at making these decisions. Data come from private markets (such as employer-sponsored health coverage) as well as from health insurance programs that are offered or subsidized by governments. The introduction of Medicare Part D is an important example. A rapidly expanding literature analyzes enrollment and plan choices, using both survey and administrative data. Generally, initial plan choices in Medicare Part D have been found to reconcile with standard normative models of decision-making; see for example Abaluck and Gruber (2011) and Heiss *et al.* (2013).[1]

Initial plan choice is only one aspect of consumer choice in Medicare Part D. Once enrolled, consumers stay in Medicare Part D for many years. Over time, they experience changes in their health and prescriptions drug needs. On the supply side, the menu of plans offered also changes from year to year. Moreover, recent reforms implement changes in the copayment and coverage structure of Medicare Part D plans such as the gradual abolishment of the infamous coverage gap. In the open enrollment period at the end of each year, enrollees therefore face a switching decision for the next year that is as complex as the initial plan choice.[2] If consumers fail to make optimal switching choices, the welfare cost to them can be large. Inertia among health insurance enrollees also has implications for firm behavior, with potentially large effects on market outcomes and overall welfare; see Handel (2013), Ericson (2014), Handel and Kolstad (2015), Ho *et al.* (2015), Decarolis *et al.* (2015), and Polyakova (2016). There is, however, also the possibility that plan choices improve over time since enrollees learn, as stressed by Ketcham *et al.* (2012).

In this paper, our objectives are to characterize the sources of inertia, to study the events that trigger switching in the presence of inertia, and to simulate and compare the effects of interventions aimed at reducing inertia. We specify a structural consumer choice model with two crucial features. First, the model comprises separate attention and plan selection stages, to separate inattention and switching costs as reasons for inertia. As Abaluck and Gruber (2016a) note this is an important distinction. Second, it allows for unobserved heterogeneity that enters both stages. Our empirical results confirm that these features

---

[1] We review the literature on plan choices in Medicare Part D in Section 2.2.

[2] Each year during the Medicare Part D annual enrollment period that runs between October 15 and December 7, individuals on Medicare have the opportunity to enroll in Part D, or if they are enrolled, to switch plans. Switching plans does not involve any fees, and as at initial enrollment, plans have to accept all individuals. Important institutional features of the Medicare Part D market are described in Bach and McClellan (2005), Duggan *et al.* (2008) and Heiss *et al.* (2010), among others. We provide an overview in Section 2.1.

together provide a more realistic picture of plan switching and inertia in Medicare Part D than simpler models. Our simulations illustrate that these features also alter the evaluation of inertia-reducing interventions. Furthermore, our findings add to the mounting evidence on the interaction between inattention and the salience of product characteristics in complex environments such as health insurance markets.

Estimation and simulation of our two-stage model of plan switching are based on administrative data on Medicare Part D plan choice and prescription drug use for 2007 through 2010.[3] A preliminary, descriptive analysis confirms that annual plan switching rates in the Medicare Part D market are low, a finding that is in line with earlier studies of plan switching in this and many other markets, reviewed in Section 2.2 below. We characterize switching behavior descriptively in terms of overspending, defined as the difference between the consumer's total costs[4] induced by the plan the consumer has chosen and that of the least-cost alternative (for given prescription drug use).[5] We find that overspending is much lower among individuals who switch plans than for stayers. Overspending is furthermore lower for those who actively decide to switch than for those who are forced to make a plan choice (because they enter the Part D market for the first time or because their old plan is not available any more). This finding suggests that those individuals who make an active switching decision are selected, which is reflected in our structural model. Our descriptive analysis further suggests that overspending cannot be fully explained by unobserved plan characteristics that might lead consumers to optimally choose plans associated with higher overspending.[6]

The structural model of plan choice we present in this paper allows for different sources of persistence. Specifically, an individual may end up staying in the same plan as in the previous year for two reasons. First, last year's plan might be the optimal choice. The fact that plan features change relatively little over time together with preference heterogeneity will result in persistence of plan choices over time. Second, an individual might

---

[3] We restrict our analysis to individuals enrolled in a stand-alone prescription drug plan (i. e., not in managed care) and who are not eligible for a Low-Income Subsidy (LIS). Further details are given in Section 3.

[4] The total costs that the consumer has to bear include premiums, deductibles, and copayments that she has to pay for prescription drugs.

[5] Overspending has been defined and studied in several earlier studies of Medicare Part D plan choice, including Abaluck and Gruber (2011), Heiss *et al.* (2013), and Abaluck and Gruber (2016a). We discuss how the overspending measure we use here is constructed in Section 3 below.

[6] This observation is related to a more general discussion about the use of parameterized models of choice behavior to study the rationality of consumers in Medicare Part D and other markets, see Ketcham *et al.* (2016) and Abaluck and Gruber (2016b). We return to this issue in the concluding Section 7.

stay with the old plan even though it is not optimal any more – this is the definition of inertia. Much of the prior literature has used the concept of switching costs to rationalize this second cause of persistence. In discrete choice models, such switching costs are typically captured by including a dummy variable for the old plan among the plan characteristics. Its coefficient has been interpreted as a monetary estimate of utility loss arising from switching costs. See Farrell and Klemperer (2007) for a review of the literature on switching costs in industrial organization. Recent studies of switching behavior in Medicare Part D and other health care markets that use such models include Miller and Yeo (2012), Nosal (2012), Handel (2013), Ericson (2014), Handel and Kolstad (2015), Ho *et al.* (2015), Abaluck and Gruber (2016a), and Polyakova (2016). These studies also provide discussions of the sources of switching costs in health insurance markets, the most important one being transaction or "hassle" costs. We review theses studies in more detail in Section 2.2.

The approach of including a dummy variable for the old plan in a discrete choice model implies that each individual compares the available plans in each year and deliberately makes a choice – which seems unrealistic. Moreover, our simulations show that in such a model, switching costs would have to be unreasonably large to produce the typically low switching rates, a result several earlier studies found as well. Polyakova (2016) points out that including heterogeneity in these one-stage models tends to bring down estimated switching costs. The two-stage model we propose achieves this as well, but we add a behavioral dimension by including inattention as a second source of inertia. We argue that this set-up is more realistic as not all consumers are assumed to evaluate all alternatives in every year. Rather, they differ in their propensity of making a deliberate choice, which may be increased by attention triggers such as salient changes of plan features. Our empirical results show that the model with a separate attention stage fits the data better and provides more realistic estimates of switching costs than the standard model that ascribes inertia to switching costs alone.

The second innovative feature of our model is heterogeneity in consumers' ability and willingness to make diligent decisions. We model this heterogeneity as a latent variable, which we call acuity, consisting of both observed and unobserved components. Importantly, acuity enters both the attention and the plan-choice stage. Moreover, the model allows for heterogeneity in the implied switching costs. Taken together, these features provide a rich structural framework that can reproduce the descriptive evidence, including selectivity of switchers.

A few recent models of switching behavior in Medicare Part D and other markets allow for both inattention and switching costs. While they share certain features of our model, none comprises all of them.[7] Kiss (2014) studies liability insurance switching in Hungary. His specification of the attention stage is simpler than ours as he can rely on a natural experiment that creates variation in the attention probability. Further, his model does not allow for unobserved heterogeneity while the fact that such heterogeneity enters both stages is a crucial feature of ours. Hortaçsu *et al.* (2015) develop another model that comprises both attention and switching stages; they use it to analyze inertia in the Texas residential electricity market. Their model also does not allow for unobserved heterogeneity that enters both stages. Ho *et al.* (2015) is substantively related to the present paper as they analyze inattention in Medicare Part D and discuss attention triggers. Their model neither allows for switching costs in the second stage nor for unobserved heterogeneity that enters both stages. They focus on the implications of inertia for firm behavior while our objective is to study interventions that reduce it.

In order to characterize the individual and joint contributions of inattention, switching costs, and acuity to the observed persistence of plan choices, we estimate four versions of the model that sequentially add features. The fourth version is the full model, as outlined above. It provides the best fit of the data, and the nested simpler versions are statistically rejected.

The estimation results for the full model provide interesting insights into the nature of inattention in this market. Consumers are more likely to pay attention to plan choice if overspending in the last year is more salient and if their old plan gets worse, for instance due to premium increases. However, consumers are not more likely to pay attention when they experience the onset of a chronic health condition that involves costly prescription drugs, although such an onset could make it worthwhile to search for a plan with a more suitable formulary. These findings are in line with the recent literature on inattention and salience, an issue to which we return in the conclusions (Section 7).

Our model allows us to simulate interventions aimed at reducing inertia. Hoadley *et al.* (2013) discuss a variety of such interventions from a policy perspective. In our framework, some of them would affect attention while others would reduce switching costs, a distinction that seems to be made neither in the policy discussion nor in earlier structural

---

[7] Our study and the papers cited in this paragraph introduce inattention in two-stage models of consumer choice. Alternative but related motivations for introducing a first stage focus on decision costs. In a study of retirement investment, Luco (2014) distinguishes between decision costs and enrollment costs. The first-stage decision in Honka (2014) determines how many markets a consumers searches and thus endogenizes the size of her consideration set.

models of plan choice in this and other markets.[8] In the one-stage standard model, inertia is induced only by switching costs which suggests interventions that reduce the hassle costs of switching, for instance by facilitating paperwork. Removing switching costs in the simplest one-stage version of our model unsurprisingly yields a dramatic increase in switching rates and reduced overspending. However, just as switching costs are likely overestimated in this simple model, the effect of removing them is unrealistically large. When we simulate versions of our model that include an attention stage, removing switching costs has a much smaller effect. As we discuss in detail below, our simulations suggest that forcing people to actively consider switching plans has a much larger effect than reducing switching costs. This and additional simulation results highlight that compared to our full two-stage model with observed and unobserved heterogeneity, simpler models of plan switching may yield misleading findings.

Importantly, our results imply that simpler models that do not include a separate attention stage overestimate switching costs and underestimate inertia. Further, ignoring (in)attention leads to a selection problem as making an active decision is the outcome of the attention stage which involves unobservables – in our model, acuity. As active plan choices are observed only for a selected group of high-acuity deciders, simpler models would overestimate the gain from reductions in inattention and switching costs: they predict the behavior of the full population using the estimates of a selected population without adjusting for the acuity differences.

The remainder of this paper is structured as follows. We begin by describing the institutional arrangements of the Medicare Part D that are relevant for our analysis and the related literature on plan choice and plan switching in this and other markets in Section 2. Section 3 describes the data sources, the construction of our analytic dataset, and the definitions of the model variables. It also reports some descriptive empirical facts about plan switching in Medicare Part D which motivate the set-up of our two-stage discrete choice model. This model is then introduced in Section 4. The estimation results are presented and discussed in Section 5. In Section 6, we report the results of simulations of policy interventions aimed at reducing inertia in the Medicare Part D market. In Section 7, we summarize our findings and draw policy conclusions.

---

[8] We discuss this issue further in Section 6.

2 Background

2.1 Institutional framework of Medicare Part D

Medicare Part D, introduced in 2006, provides the Medicare-eligible population with universal access to standardized, heavily subsidized prescription drug coverage through government-approved plans sponsored by private insurance companies and health maintenance organizations (HMOs). In addition to providing access to affordable drug coverage to all Medicare beneficiaries (in particular to the chronically ill), a second policy goal was to create a "competitive, transparent marketplace offering a wide array of benefits" (Bach and McClellan (2005), p. 2733). The institutional design of Medicare Part D exemplifies the current trend toward "consumer-directed health care" (Goodman (2006)) as it relies on consumer behavior and competition among insurers to attain satisfactory market outcomes with limited government regulation. In the case of Medicare Part D, and arguably also in other similar programs, giving consumers more choice also means confronting them with difficult decisions. In the following, we describe the features of the program that are relevant for our present analysis.

Medicare beneficiaries can enroll in Part D when they become eligible for Medicare. If they enroll, they can choose between about 50 plans (the exact number varies across regions). Once enrolled, they can switch to a new plan annually during the open enrollment period at the end of each year. If individuals do not actively decide to switch to a different plan they are automatically re-enrolled in their old plan for the new year. The menu of available plans is the same for first-time enrollees and switchers.

Each firm that wants to enter this market has to offer at least a plan that is actuarially equivalent to a standard plan whose features are tightly specified. Features of the standard plan include formulary, the beneficiary's annual pharmacy bill for drugs in the plan formulary, the beneficiary's total out-of-pocket (OOP) payments for these covered drugs, the threshold for catastrophic coverage, and the monthly premium,. The standard Medicare Part D plan had the following benefit schedule in 2006:[9]

- The beneficiary has an annual deductible of $250.

- The beneficiary pays 25 percent of drug costs above $250 and up to $2,250 (the "initial coverage limit" or ICL). The expense born by the beneficiary is then $750 for a beneficiary whose pharmacy bill has reached $2,250.

---

[9] In subsequent years, these numbers were adjusted, see for example Heiss *et al.* (2013).

- The beneficiary pays 100 percent of drug costs between total OOP of $2,250 and $3,600; this is referred to as the coverage gap or doughnut hole. The threshold of $3,600 is attained at a drug bill of $5,100.

- The beneficiary pays 5 percent of drug costs above the drug cost threshold of $5,100, at which the total OOP threshold level is achieved; this is referred to as catastrophic coverage.

The Medicare Part D plans sponsored by private insurance firms may differ from the standard plan in their premiums and other plan features, provided that their benefits for any drug costs are, on average, at least as high as those of the standard plan. Enhancements may in particular include coverage for the $250 deductible and for the gap in the standard plan.

This design implies that Medicare Part D stand-alone plans are characterized by the following variables: premium, deductible (if any), ICL for those plans that are not classified as "standard benefit", and the formulary benefit design which specifies drug tiers and co-payments. Together with the specific prescription drugs an enrollee uses over the course of the year, these characteristics determine her OOP cost. The sum of these OOP cost and the premiums paid is the total cost to the consumer which for simplicity we refer to as "total costs" in the following.

Despite this complexity, there are two institutional features that provide beneficiaries with easy access to information on the plans that are available to them and their cost implications.

First, it is rather straightforward for an individual to figure out the total costs implied by her prescription drug use (or, if the calculation is made for the future, by her expected prescription drug use), for her current plan and any other plans that are available for the next year. The Medicare Plan Finder, provided by the Centers for Medicare & Medicaid Services (CMS) is an online decision tool supporting Part D plan choices.[10] Plan Finder allows consumers to determine the premiums and expected annual OOP costs of any given set of prescription drugs for the plans available at their location, and invites consumers to do this for their existing prescriptions. Plan Finder thus allows consumers to evaluate and compare total costs of all plans at low costs.[11]

---

[10] The CMS Plan Finder can be accessed at http://www.medicare.gov/find-a-plan/ (last visited: September 9, 2016).

[11] A further complication arises as the drug within a therapeutic class that implies the lowest copayment depends on the chosen plan. This introduces another dimension, drug substitution, into the choice

Second, before the beginning of the open enrollment period, each plan has to send a standardized "Plan Annual Notice of Change" (ANOC) to its enrollees, based on a template provided by CMS. This notice contains a table with cost-relevant plan features for the current and the following year (premium, deductible, copayments) as well as detailed information on all changes that will become effective in January of the following year.[12]

2.2 Existing literature on plan choice and switching in Medicare Part D

Since its introduction in 2006, a large number of studies pose the question of how good individuals are at making decisions in the Medicare Part D market. How consumers respond to prices and costs in this market is crucial for assessing the welfare effects of its introduction and subsequent reforms, in particular given that plan sponsors will adjust their pricing strategies accordingly.

Early studies of enrollment decisions and initial plan choices, such as Heiss *et al.* (2010) and Kling *et al.* (2012), analyzed survey data and documented choices that seemed unlikely to be optimal. At the same time, initial enrollment rates were high and overall the introduction of Medicare Part D was deemed to be successful, see Heiss *et al.* (2006), Goldman and Joyce (2008), and Duggan *et al.* (2008).

More recent research uses much more detailed administrative data on plan choices and prescription drug claims provided by CMS. A finding of several early studies in this literature is that for substantial fractions of those enrolled in Part D, initial plan choices imply overspending: For a given use of prescription drugs, plans with lower total costs than that of the chosen plan exist, see Abaluck and Gruber (2011), Zhou and Zhang (2012), and Heiss *et al.* (2013). For example, Heiss *et al.* document that in 2006–2009, less than a quarter of individuals were enrolled in plans that were, from an *ex ante* viewpoint, as good as the least-cost plan covering the same drugs. Their estimates indicate that consumers overspent about $300 per year, on average. Unobserved plan and taste heterogeneity may explain some of the overspending documented in this and other studies but most likely not all of it.

---

problem. Plan Finder suggests generic substitutes for branded drugs, but does not conveniently provide information that allows consumers to price out drug substitutes, or calculate risks associated with new conditions or complications. Relative to the first-order effect of choosing the least-cost plan for a given set of prescription drugs, drug substitution is, however, a second-order concern. See Heiss *et al.* (2013) for a detailed discussion.

[12] A recent example of an ANOC can be found at `https://www.yourmedicaresolutions.com/sites/default/files/2015%20%20Standard%20ANOC.pdf` (last visited: September 9, 2016).

There is less consensus about whether plan choices have improved over time. Given that beneficiaries' drug needs evolve over time and plan sponsors adjust key features such as premiums and the formulary benefit design every year, and as changing plans during the annual open enrollment period involves no monetary cost, one would expect that switching is frequent. However, switching rates are rather low – estimates of about 10 percent had already been found in 2006 and 2007 survey data by Heiss *et al.* (2010). Low switching rates have also been documented by Hoadley *et al.* (2013) in a descriptive analysis of plan switching using Medicare Part D claims data for 2006 through 2010. They document that switching rates are generally low; only 13 percent voluntarily switch plans during the annual enrollment period. Even though large premium increases were associated with higher switching rates, most enrollees with relatively large premium increases did not switch plans. Seven out of ten Medicare beneficiaries enrolled in stand-alone Part D prescription-drug plans during all four annual open enrollment periods from 2006 to 2010 did not voluntarily switch plans in any of these periods. These descriptive results are in line with those we report in Section 3 below and with those of several other studies that employ explicit choice models.

A natural approach to assess the effect of inertia at the individual level is to check how overspending evolves over time. Ketcham *et al.* (2012) find that enrollees were more likely to switch plans if their potential gain from doing so was larger and that overall, large reductions in overspending were realized from 2006 to 2007. In contrast, using the full universe of claims data that became available subsequently, Abaluck and Gruber (2016a) find that forgone savings from choosing suboptimal plans have increased during the first four years of Medicare Part D. They argue that there has been little consumer learning over time in the Part D market, and that increasing choice inconsistencies are driven by changes on the supply side that are not offset by consumers because of inertia.

As mentioned in the introduction, studies of plan inertia in this and other markets typically involve a discrete choice model that includes a dummy variable as a characteristic of the old plan. Miller and Yeo (2012) develop an algorithm that allows them to estimate switching costs using aggregate data on market shares. Applying this approach to data from Medicare Part D for 2006–2010, they find very high switching costs of around $1,700, which seems unrealistic. Estimates of switching costs based on individual level data tend to be smaller as they allow for some heterogeneity, see Polyakova (2016). Nevertheless, papers that use individual-level data find evidence of substantial switching costs and also of inattention (Ho *et al.* (2015)), and this is also the case for the present study. However, there is no consensus about how serious the problem of inertia is. Ketcham *et al.* (2015)

argue that while annual switching rates are low, over time a relevant fraction of those enrolled in Part D do switch: Of those enrolled in 2006, only about half remained in the same plan in 2010.

Another dimension of consumer behavior in health insurance markets that has been studied using administrative data from Medicare Part D concerns their reaction to nonlinear price schedules. Einav *et al.* (2015) show that beneficiaries react to the complex intertemporal incentives induced by non-linearities in Part D's copayment structure by substituting drugs within and across years. Their model assumes that all beneficiaries respond to these incentives similarly, but they differ with respect to their price elasticities.

3 Data construction and descriptive evidence on plan switching

In our analysis of plan switching and inertia in Medicare Part D, we use the claims records of a 20% representative sample of Medicare enrollees in the years 2007–2010. These claims data have also been used by other researchers working on Medicare Part D plan choice and switching, including Abaluck and Gruber (2011), Abaluck and Gruber (2016a), Heiss *et al.* (2013), Ho *et al.* (2015), Hoadley *et al.* (2013), and Polyakova (2016). As most of these earlier studies, we restrict our analysis to the claims for beneficiaries aged at least 65, in stand-alone prescription drug plans (PDPs) who do not receive low-income subsidies and are not dual-eligible for Medicare and Medicaid.

The claims data consist of several databases. Conceptually, the data can be thought of as being organized by beneficiary. The plan in which the beneficiary is enrolled in a given year is identified by an encrypted plan ID that allows plan characteristics (premium, deductibles, gap coverage) to be linked. The plan formularies which specify the copayments and are thus essential for determining the beneficiary's total costs associated with alternative plans were not available when the dataset we use was constructed. Information on plan formularies is thus inferred from copayment costs.[13]

Throughout the paper, we use the following convention for dating variables: Decisions with respect to plans covering prescription drug use in year t are taken during the open enrollment period at the end of year $t-1$. The plan in which a consumer is enrolled in year $t-1$ is referred to as the "old plan" or the "year $t-1$ plan", and the "new plan" is the one covering year t.

---

[13] The algorithm is described in Heiss *et al.* (2013); a similar approach has been used by other researchers including Abaluck and Gruber (2016a).

3.1 Sample

To study switching between Part D plans among seniors, we form working samples of Part D enrollees who have unrestricted choice among all the plans available in their Medicare region, and have sufficient data on their health, drug use, OOP costs, and premiums to estimate our plan-choice models.

The full 20% samples comprise between 9.3 and 10 million individuals per year, for the four years 2007–2010. We restrict our analytic sample to those individuals who are U.S. residents, aged 65 years or older, enrolled in stand-alone, non-employer (non-EGWP) Part D prescription drug plans; are neither entitled to the Low-Income Subsidy (LIS) nor dual eligible for Medicaid; are continuously enrolled in Part D in two adjacent years; and do not switch plans during the year (outside of the open enrollment period).[14]

Applying these criteria results in sample sizes of around 1.2 million individuals for the years 2008–2010 (Table 1). As we analyze plan choice for year t as made at the end of year $t-1$, conditional on year $t-1$ information, we estimate our choice models for these three years. The criteria for selecting individuals for our analytic sample are essentially the same as those used by Heiss *et al.* (2013) and other studies of plan choice in Medicare Part D. In the final step, we randomly select 100,000 individuals since the estimation of our model is computationally intensive. The number of observations available for each year is smaller (Table 1) as the resulting panel is unbalanced. The total number of choices across all individuals and years is 238,280. All descriptive statistics and model estimates reported in the remainder of the paper refer to this estimation sample, or to sub-samples thereof.

3.2 Plan choice, prescription drug use, and individual characteristics

For our subsequent analysis, we construct various variables that measure individual characteristics, health status, prescription drug use, and plan choice. These data are augmented with data on the features of the available plans (which may vary every year) such as premium and whether they provide gap coverage.

In addition, we construct cost variables based on the individual's prescription drug use during year $t-1$, i.e., the year at the end of which decisions for year t are made. The

---

[14] For those who are not in a stand-alone plan, we do not have data on individual prescription drug claims. Eligibility for the LIS, in turn, changes the nature of the choice problem an individual faces considerably, as discussed for instance by Decarolis (2015). These groups are commonly excluded in recent studies of plan choice and switching behavior using claims data, as for instance in Ho *et al.* (2015), Abaluck and Gruber (2016a), and Polyakova (2016).

list of all prescription drugs used combined with the formulary and benefit design of a Part D plan and the plan's premium gives the total costs of that plan, i.e., the sum of all copayments, the premiums, and any deductibles. We construct total costs for year t for the plan chosen in year $t-1$ and for any plan available in the next year.

We also construct variables that indicate whether according to her OOP costs, an individual (in her current year plan) ended up in the Part D catastrophic region, in the coverage gap, or below the coverage gap.

Finally, we construct indicators for the incidence of new health conditions, classified by whether they are costly in terms of prescription drugs or not. We classify conditions as costly if having the specific condition predicts high prescription drug costs in supplementary regressions.[15]

As the claims data contain very little demographic information on the individual (age, sex, and race), we further augment our data with regional information from the American Community Survey (ACS). We include the ZIP-code share of individuals aged 65 and above without a high school degree ("low education") as well the share of the population aged 65 or above with income below the federal poverty level ("low income") and above $5\times$ the federal poverty level ("high income").

Descriptive statistics for these variables are reported below in Section 5. In the remainder of the present section, we discuss the choice and cost variables descriptively.

3.3 Optional vs. forced choosers

In our subsequent analysis, we distinguish optional and forced choosers: Forced choosers are individuals who do not have a default plan and thus are forced to pay attention to plan choice. For optional choosers, on the contrary, the plan that they were enrolled in $t-1$ is part of their choice set in t. If they do not take action, they will by default stay in their old plan.

Optional choosers either switch to a new plan or stay in the old one. Thus, every year, each individual falls into one of three groups: Forced choosers, optinoal switchers, and

---

[15] Costly conditions include acute myocardial infarction, ischemic heart disease, stroke, breast and lung cancer, Alzheimer's disease and dementia, diabetes, depression and asthma. In comparison to these conditions, hip fracture, prostate cancer, endometrial cancer, colorectal cancer, chronic obstructive pulmonary disease, chronic kidney disease, osteoporosis, heart failure, rheumatoid arthritis/osteoarthritis, glaucoma, anemia, benign prostatic hyperplasia, hyperlipidemia, acquired hypothyroidism, hypertension, atrial fibrillation are relatively cheap and thus included as "cheap" conditions.

optional stayers. Table 2 displays the shares of these groups. The majority of individuals ends up staying in their old plan. On average across all three years, only a little more than 10% of the optional choosers leave their old plan. On average, and across all years, about 3.4% of individuals are forced to make a plan choice. The majority of them do not have a default plan because they are new to Medicare Part D (97%). Only 3% of forced choosers have been enrolled in a $t-1$ plan that is no longer available in year t. Despite some differences in sample definitions, these numbers are in line with those reported in an extensive descriptive analysis of administrative data on Medicare Part D plan switching between 2006 and 2010 by Hoadley *et al.* (2013).

3.4   Total costs and overspending

To motivate our analysis of inertia in Medicare Part D plan choices and the effect of policy interventions designed to reduce it, we now present descriptive evidence on the quality of plan choices of forced choosers, optional switchers and optional stayers. In the literature on plan-choice quality in Medicare Part D, and in health insurance more generally, a measure of overspending is frequently used for this purpose, see Abaluck and Gruber (2011), Heiss *et al.* (2013), and Abaluck and Gruber (2016a), among others. As stated above, conceptually, overspending is defined as the difference between the total costs (to the consumer) of the plan a consumer has chosen and the total costs of the least-cost alternative. This definition reflects the fact that relevant features of the plans such as premiums, deductibles, the plan formulary and copayments are summarized in the total costs.

In practice, measures of overspending can be constructed in different ways. A key distinction is that between an *ex ante* and an *ex post* perspective. Tables 3 and 4 display *ex post* and *ex ante* overspending for forced and optional choosers, separately for stayers and switchers.

Ex post overspending for year t is calculated based on the prescription drugs that an individual has taken in that year. It thus reflects the costs that beneficiaries actually realized during year t. As Table 3 shows, individuals who stay in their old plan have the highest overspending, on average, compared to both optional switchers and forced choosers. These numbers might be taken to imply that actively making decisions results in better outcomes. This is not what one would expect if individuals stayed in their old plans because of a preference for unobserved plan characteristics.

Individuals who are forced to do not choose quite as well as those who actively decide to switch to a different plan, which indicates substantial heterogeneity in the ability to choose plans. This implies that optional switchers are a selected group of individuals who appear to make better choices than forced choosers who, in turn, are mostly newly enrolled and therefore not selected at all. Our structural model of plan choice allows for heterogeneity of individuals' choice quality and can explain these features of observed overspending.

In Table 4, we turn to *ex ante* overspending for stayers and switchers, which is based on prescription drugs taken in year $t - 1$ (this measure cannot be computed for the newly enrolled individuals who make up 97% of the forced choosers as their prescriptions for $t - 1$ are not observed). This measure is meant to capture beneficiaries' information on their prescription drug needs at the time when they make their plan choice. This is also the information beneficiaries are asked to provide to the Medicare Plan Finder discussed above. The table shows *ex ante* overspending in year t for the plan that individuals chose for year t as well as for the plan that they chose in $t-1$. For the group of stayers, the two plans are obviously the same. For the switchers, overspending based on the $t-1$ plan is the amount they would have overspent had they stayed in their old plan. The difference in overspending between the old and new plans thus captures how much switchers have saved by switching plans – $180 per year, on average. Compared to stayers, switchers' overspending in the $t-1$ plan is slightly higher in all years except for 2010. This suggests that high *ex ante* overspending in the old plan may encourage individuals to switch to a different plan.

A particularly interesting group of individuals are those who have chosen the *ex ante* cost-minimizing plan. As discussed in Heiss *et al.* (2013), this is the choice implied by static optimization. It can easily be implemented, for instance by using a decision aid such as the Medicare Plan Finder discussed above. The lower panel of Table 4 displays the share of individuals with zero *ex ante* overspending, separately for the two groups of optional choosers. A much larger share of switchers than of stayers have zero *ex ante* overspending. These numbers also support the notion that switchers make better choices than stayers.

In the structural model of plan choice presented in the remainder of this paper, we take an *ex post* perspective. We cannot use the *ex ante* measures because they are not available for the group of forced choosers. *Ex post* measures of overspending might seem problematic as they condition on information on outcomes, realized during year t, that are not known to the individual at the end of year $t-1$. *Ex ante* measures, in contrast, do not condition

on year t outcomes but require assumption on how individuals form expectations at the end of year $t-1$ with respect to these variables. These assumption can also be problematic. We discussed the trade-off between using *ex ante* and *ex post* measures in detail in earlier work, see Heiss *et al.* (2013). In our analysis of static Medicare Part D choices, we found that the key substantive findings do not depend on whether *ex ante* or *ex post* measures are used.

4 A two-stage model of plan switching decisions

In this section, we introduce a comprehensive model of plan selection that is behaviorally more plausible and can explain more of the empirical findings presented in Section 3 than typical models used in the literature. It includes two features that have not been combined in previous papers. First, it comprises separate inattention and plan selection stages including switching costs. Without some sort of trigger, consumers might not pay attention to the plan-choice opportunity and fail to compare plans during the open enrollment period. If they compare plans and find a more attractive one, switching might still be hampered by hassle and other switching costs. Second, our model allows for both observed and unobserved heterogeneity in the ability and willingness to make diligent decisions. This "acuity" affects both stages, allowing for the effect that individuals who consider switching make more careful choices than the average. Leaving the attention stage aside, the choice model uses a mixed (i. e., random parameters) multinomial logit specification with heteroskedasticity, building on McFadden (1974) and McFadden and Train (2000).

We first present the structure of the choice problem in Section 4.1. The concept of acuity is introduced in Section 4.2. Then we discuss the two stages of the choice model in Sections 4.3 and 4.4. Identification of the two-stage model is discussed in Section 4.5. Since inattention is unobserved and acuity enters both stages simultaneously as a latent variable, all model parameters have to be estimated jointly. We use a maximum likelihood estimator with a numerical approximation of the analytically unavailable likelihood function; details are discussed in Section 4.6.

4.1 Structure of the choice problem

We model the plan choice for a set of consumers $i = 1, \ldots, N$ for years $t = 1, \ldots, T$.[16] Since the individuals start enrolling in Part D in different years, can die during the time our panel data covers, or can change the low-income subsidy status or other criteria for inclusion in our estimation sample discussed in Section 3.1, our panel data set is unbalanced and T differs by individual. For notational convenience, we do not explicitly index T by i in the remainder of this section. Individual i chooses among the available plans in year t. This choice set $\mathcal{Y}_{it}$ contains around 50 plans on average, but the number of plans and their features differ by t and Medicare region.

4.2 Acuity

Individuals differ in both their ability to, and their willingness to put effort into, making good plan-choice decisions. We introduce a latent variable, acuity, that captures these aspects. Acuity may comprise such diverse factors as intelligence, financial literacy, mental health, opportunity costs of time, or the availability of support from children, doctors, or other sources. It can affect choices in both stages: Individuals with high levels of acuity might be more likely to pay attention in the first stage of the model, i.e., they collect and compare information on the available plans. If they do pay attention, acuity might also affect the weight they put on the objective consequences of the plans such as their premiums and implied OOP costs and the error variance in the second stage (the multinomial choice model).

Acuity q_{it} of individual i at time t is a latent variable that depends on observed individual characteristics $\boldsymbol{w}_{it}$. We include variables such as age, gender, depression, and ZIP code level socio-economic characteristics. q_{it} also includes a time-constant unobserved heterogeneity term λ_i which we assume to be normally distributed:

$$q_{it} = \boldsymbol{w}_{it}\boldsymbol{\alpha} + \lambda_i \tag{1}$$

Since q_{it} is unobserved, it is only identified up to location and scale. For normalization, we do not include a constant in the model and define λ_i to have a zero mean and a variance of one.

---

[16] The time index t refers to the year for which the plan is chosen. The actual choice is made in the open enrollment period at the end of the previous year.

4.3 Attention stage

Let a_{it} denote an unobserved indicator for whether individual i pays attention at time t ($a_{it} = 1$) or not ($a_{it} = 0$). Some individuals are forced to pay attention and choose a plan, for example in the first year of enrollment or if a plan is discontinued. Let a dummy variable, f_{it}, be equal to one if individual i is forced at time t. For optional choosers ($f_{it} = 0$), we model the attention probability as a function of observable attention triggers, collected in the vector $\boldsymbol{x}_{it}$, and unobserved individual acuity q_{it}, which in turn is modeled as a function of observed individual characteristics in equation (1).

Whether individual i pays attention in choice situation t is determined by a latent variable, a_{it}^*. We use a linear specification

$$a_{it}^* = \boldsymbol{x}_{it}\boldsymbol{\zeta} + \delta_a q_{it} + e_{it}. \tag{2}$$

Optional choosers pay attention if a_{it}^* is positive while forced choosers always pay attention. Assuming i.i.d. logistic errors e_{it}, this implies a conditional attention probability of

$$p_{it}^a(\lambda_i) \equiv \Pr(a_{it} = 1 | f_{it}, \boldsymbol{x}_{it}, \boldsymbol{w}_{it}; \lambda_i) = \begin{cases} \frac{1}{1+\exp(-\boldsymbol{x}_{it}\boldsymbol{\zeta} - \delta_a q_{it})} & \text{if } f_{it} = 0 \\ 1 & \text{if } f_{it} = 1 \end{cases} \tag{3}$$

4.4 Plan-choice stage

Given an individual pays attention, she compares plans according to the perceived utilities

$$u_{itj} = \boldsymbol{z}_{itj}\boldsymbol{\beta} + \gamma_i d_{itj} + v_{itj}, \tag{4}$$

where $\boldsymbol{z}_{itj}$ is a vector of plan characteristics, including premium and implied OOP costs given prescription drug use. The dummy variable d_{itj} indicates the plan she is enrolled in year $t-1$ so that $d_{ijt} = 0$ is associated with a switch. As in prior literature, the coefficient γ_i of this lagged choice dummy is interpreted as a switching cost. We allow for the possibility that these switching costs vary across individuals by using a random coefficient specification and assuming a normal distribution,

$$\gamma_i \sim \mathcal{N}\left(\mu_\gamma, \sigma_\gamma^2\right). \tag{5}$$

We allow individual acuity, q_{it}, to affect choices in the sense that individuals with higher values of q_{it} give relatively more weight to utility contributions of observed plan characteristics and the switching cost. Those with lower values of q_{it} are relatively more affected

17

by the noise involved in the error terms v_{itj}. We thus specify that the standard deviation of v_{itj} is related to acuity via

$$\sqrt{\text{var}(v_{itj}|q_{it})} \propto \exp(\delta_v q_{it}). \tag{6}$$

With a plan-choice parameterization resembling a heteroscedastic mixed multinomial logit model, the conditional probability that individual i chooses plan j from choice set $\mathcal{Y}_{it}$ at time t given she pays attention is

$$p_{itj}^y(\lambda_i, \gamma_i) \equiv \Pr(y_{it} = j | a_{it} = 1, \boldsymbol{z}_{itj}, \boldsymbol{w}_{it}, d_{ijt}; \gamma_i, \lambda_i) = \frac{\exp\left(\frac{\boldsymbol{z}_{itj}\boldsymbol{\beta} - \gamma_i d_{itj}}{\exp(\delta_v q_{it})}\right)}{\sum_{k \in \mathcal{Y}_{it}} \exp\left(\frac{\boldsymbol{z}_{itk}\boldsymbol{\beta} - \gamma_i d_{itk}}{\exp(\delta_v q_{it})}\right)}, \tag{7}$$

where as defined above acuity is $q_{it} = \boldsymbol{w}_{it}\boldsymbol{\alpha} + \lambda_i$.

Our model nests several special cases that have been studied in prior literature. Given our parametric structure, we can test the assumptions that produce special cases as follows.

As in other models of plan switching, $\gamma_i = 0$ implies that there are no switching costs. As we allow for heterogeneity of these costs, we test whether its population mean and variance are zero ($\mu_\gamma = 0$ and $\sigma_\gamma = 0$). Importantly, however, if decisions follow the full two-stage model, a restricted one-stage model that ignores attention would overestimate switching costs as all observed inertia would be attributed to this source of heterogeneity.

If $\delta_a = 0$ and $\delta_v = 0$, then acuity does not matter, i.e., there is not heterogeneity with respect to decision quality. If only one of these parameters is different from zero, then acuity only affects choices via the corresponding stage. Further, if at least one of these parameters is zero, then decision quality at the second stage is conditionally independent of attention. We expect that individuals with higher levels of acuity tend to have a higher attention probability ($\delta_a > 0$) and make better plan-choice decisions ($\delta_v < 0$). This is one of the novel testable behavioral implications of our model.

If this mechanism holds, optional choosers who end up paying attention have a relatively high acuity and therefore tend to pick "good" plans with low total costs. We have seen a first indication of this effect in Section 3.4 where we reported that optional switchers have a lower realized overspending than forced choosers. Ignoring this effect would lead to biases in the policy conclusions. Specifically, as optional choices are observed only for a selected group of high-acuity choosers, simpler models would overestimate the gain from reductions in inattention and switching costs as they cannot predict the behavior of the full population using the estimates of a selected population without adjusting for the acuity differences.

4.5 Identification

Data on repeated plan choices reveal whether a beneficiary switches plans for the next period. The identification problem arises because if she does not switch, we do not observe the reason why.

The simplest explanation for why a beneficiary decides to stay in the same plan in two consecutive years is that she carefully compared plans and came to the conclusion that her current $t-1$ plan is the most attractive one available for the next year, t. Given that on average around 50 plans are available and that both the list of available plans and their features (like the premium, deductible, list of covered drugs, copayments and others) change on a yearly basis, this alone is unlikely to explain the low switching rates observed in the data. We are able to account for a large part of the plan characteristics that should be relevant to an individual who chooses an insurance plan to minimize total costs. We observe exactly which drugs an individual takes and can reconstruct the financial consequences of buying these drugs under each available plan, as reflected in our total cost measure. We also account for specific plan features such as the premium and deductible as well as the risk associated with each plan. On top of that, we include alternative-specific constants for the largest plans. Allowing for preference heterogeneity provides a comprehensive model of deliberate plan choices. The remaining persistence that cannot be explained by plan features and preference heterogeneity is what we call inertia and attribute to inattention and switching costs.

For identification, we can exploit the fact that some individuals are forced to make an active choice because they enter the market for the first time or their $t-1$ plan is discontinued. For similar arguments, see Handel (2013), Handel and Kolstad (2015) and others. These forced choosers by definition do not have the opportunity to be inattentive. If they are otherwise comparable to the other consumers, their behavior alone identifies the plan choice parameters and the difference to the optional choosers is attributed to inertia. As forced and optional choosers have similar means of the observed variables (in particular age, education, and poverty rate; see Table 5 below), the assumption of similar preferences across the two groups seems reasonable. Also, they are very similar in terms of the choice sets they face; see Table 6 below.

The explanatory variables that drive the attention probability and the plan choices are different in a quite natural fashion, creating exclusion restrictions that add to the identification of the model. Take the premium as an example. If the premium of the $t-1$ plan *increases* in t – a change that is made salient in the ANOC letter – our model implies that

attention is triggered and the individual starts to compare alternatives. If on the other hand the *level* of the premium of the $t-1$ plan is high for year t relative to the alternatives – a fact that is only apparent after comparing plans – this might contribute to overcoming switching costs and changing the plan. The same holds for the other predictors. Attention is triggered by different shocks or changes in the past, whereas for the plan choice, comparison to the other plans provide the relevant predictors. Importantly, in this market attention triggers such as changes in the old ($t-1$) plan's premium and other features are highly salient as the ANOC letters prominently contain pairwise comparisons of the old and new values of plan features, see also Ho *et al.* (2015). The assumption that changes in state variable trigger attention is also in line with theoretical models of inattention such as Bordalo *et al.* (2013).

A final identification issue arises in dynamic panel data models with lagged dependent variables if unobserved initial conditions and unobserved heterogeneity are correlated, see Heckman (1981). In the present analysis, we focus on heterogeneity in acuity, λ_i, and switching costs, γ_i, neither of which are likely related to the specific plan chosen in the first year. Moreover, in Medicare Part D, we come close to observing the initial conditions; see also Polyakova (2016). As the first choices in our panel refer to 2008, the lagged dependent variable is the plan choice made for 2007 which was the first complete year of Medicare Part D. Also, the vast majority of the forced choosers in our sample are newly enrolled in Part D, so that their initial choices are observed and explicitly modeled.[17]

4.6 Maximum likelihood estimation

Given our parametric specification, we can estimate all model parameters simultaneously using maximum likelihood. Since we assume independence across individuals, the likelihood function is the product of the individual likelihood contributions $\mathcal{L}_i(\boldsymbol{\zeta}, \boldsymbol{\beta}, \delta_a, \delta_v, \mu_\gamma, \sigma_\gamma^2)$. The observed outcome for individual i is the sequence of plan choices made in years $t = 1, \ldots, T$ (while attention is unobserved). Let j_{it} denote the observed plan choice in year t. Thus, the likelihood contribution is given by

$$\mathcal{L}_i(\boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta_a, \delta_v, \mu_\gamma, \sigma_\gamma^2) = \Pr(y_{i1} = j_{i1}, \ldots, y_{iT} = j_{iT} | \cdot), \tag{8}$$

where for notational convenience we suppress the observed covariates (the histories of the $\boldsymbol{w}_{it}$'s, $\boldsymbol{x}_{it}$'s, and $\boldsymbol{z}_{it}$'s). Further, conditional on these covariates as well as individual i's

---

[17] In principle, we could use data from 2006 as well so that we would observe initial conditions for all individuals. However, as discussed in Heiss *et al.* (2010) and Heiss *et al.* (2013), the choice situation faced by consumers in the first year of Medicare Part D was special due to the extended enrollment period which ended in May, so we opted against using those data.

unobserved acuity part λ_i and her switching cost γ_i, her choices are independent over time. Put differently, any observed dependence is due to these two latent characteristics, conditional on observables. We can therefore write

$$\Pr(y_{i1} = j_{i1}, \ldots, y_{iT} = j_{iT} | \cdot, \lambda_i, \gamma_i) = \prod_{t=1}^{T} \Pr(y_{it} = j_{it} | \cdot, \lambda_i, \gamma_i). \qquad (9)$$

These probabilities are readily available given our assumptions on the attention and plan-choice stages. We first obtain the choice probabilities for period t:

$$\begin{aligned}
\Pr(y_{it} = j_{it} | \cdot, \lambda_i, \gamma_i) &= p_{it}^a(\lambda_i) \cdot \Pr(y_{it} = j | a_{it} = 1, \boldsymbol{w}_{it}, \boldsymbol{z}_{itj}, d_{ijt}, \lambda_i, \gamma_i) \\
&\quad + (1 - p_{it}^a(\lambda_i)) \cdot \Pr(y_{it} = j | a_{it} = 0, \boldsymbol{w}_{it}\boldsymbol{z}_{itj}, d_{ijt}, \lambda_i, \gamma_i). \quad (10)
\end{aligned}$$

This expression involves probabilities that condition on attention, which is not observable. By plugging in the quantities derived in (3) and (7) above, we see that the choice probabilities have different forms depending on whether the individual is a forced chooser or not and whether she stays in the same plan or switches to a different plan between years $t-1$ and t, which are both observable events. For forced choosers ($f_{it} = 1$), the attention probability is $p_{it}^a(\lambda_i) = 1$. If optional choosers ($f_{it} = 0$) do not pay attention, the probability to choose a different than the old plan (i.e., $d_{itj_{it}} = 0$), is zero, the probability to choose the old plan is one.

Putting everything together, we get

$$\Pr(y_{it} = j_{it} | \cdot, \lambda_i, \gamma_i) = \begin{cases} p_{itj_{it}}^y(\lambda_i, \gamma_i) & \text{if } f_{it} = 1 \\ p_{it}^a(\lambda_i) \cdot p_{itj_{it}}^y(\lambda_i, \gamma_i) & \text{if } f_{it} = 0 \wedge d_{itj_{it}} = 0 \, (11) \\ p_{it}^a(\lambda_i) \cdot p_{itj_{it}}^y(\lambda_i, \gamma_i) + (1 - p_{it}^a(\lambda_i)) & \text{if } f_{it} = 0 \wedge d_{itj_{it}} = 1 \end{cases}$$

To obtain the probability of individual i's entire choice sequence, we plug this expression into equation (9). To calculate the likelihood contribution, the final step is to integrate out the remaining two latent quantities, q_i and γ_i:

$$\Pr(y_{i1} = j_{i1}, \ldots, y_{iT} = j_{iT} | \cdot) = \int\!\!\int \prod_{t=1}^{T} \Pr(y_{it} = j_{it} | \cdot, \lambda, \gamma) \, g(\lambda) h(\gamma) \, \mathrm{d}\lambda \, \mathrm{d}\gamma, \qquad (12)$$

where $g(\cdot)$ and $h(\cdot)$ are the densities of λ and γ, respectively. Both latent variables are assumed to follow a normal distribution. Recall that we impose the normalization that λ is standard normally distributed, whereas mean and variance of the normal variate γ are left unrestricted. The double integral does not have an analytic solution. We use a Gaussian quadrature product rule to accurately approximate it.[18]

---

[18] With an even more flexible model including several random parameters, the dimension of this integral would increase and product rule Gaussian quadrature would become infeasible. However, in that case it

5 Results

In this section, we present the main empirical results. We begin by providing descriptive statistics of the observed covariates and choice outcomes. We then present the parameter estimates of our choice model, focusing on the most comprehensive specification, and interpret these results in terms of the implied marginal effects. We end the section by comparing these results with those obtained from estimating simpler, nested choice models.

5.1 Descriptive statistics

Table 5 contains descriptive statistics of the covariates that enter the acuity equation and the attention stage, separately for optional and forced choosers and stratified by switching status. The top panel shows variables that are potential triggers of attention. As most of the forced choosers newly enter Medicare Part D, the attention trigger variables that exclusively refer to information in $t-1$ are not available for the majority of forced choosers. The attention trigger variables are thus not shown for the forced choosers.The bottom panel of this table shows demographic controls that may affect individuals' acuity when making Part D choices.

There are three different sets of attention triggers: (1) experience in Part D in year $t-1$, (2) changes in features of the year $t-1$ plan, and (3) health shocks and prescription drug use in year $t-1$. The descriptive statistics suggest that Part D experience and changes in the old plan may be particularly relevant triggers: while among those who decide to switch plans 22% were in the coverage gap in year $t-1$, only 18% among stayers hit the gap. Similarly, switchers face an average premium increase of $130 per year while stayers only face an average increase of $70. 14% of switchers see changes in cost sharing, either from coinsurance to copayment or vice versa, compared to only 8% among stayers.

With respect to the variables that enter the acuity equation, switchers and stayers appear to be rather similar. The only marked difference is observed in the fraction of non-whites, which is 4% among switchers and 7% among stayers. As most forced choosers are newly enrolled in Medicare Part D, it is not surprising that forced choosers are younger, more likely to be male (which is correlated with age), and have fewer years of experience with Part D than optional choosers. With respect to all other demographic variables, optional and forced choosers are very similar.

---

could be approximated using either Monte-Carlo simulation or powerful multidimensional integration rules such as sparse grids, see Heiss and Winschel (2008).

Table 6 compares plan characteristics in the choice sets faced by the individuals in the estimation sample, again separately by switching status. Overall, the choice sets seem to be very similar across the three groups of individuals. On average, they can choose among roughly 52 plans, with very similar average premiums, deductibles, ICLs, and gap coverage. The last plan characteristic is the variance of the OOP costs on the plan level. This measure is meant to capture the variability of costs that consumers can expect with each plan. Including the plan-level variance of OOP as a regressor allows risk preferences to play a role.

Coming back to our discussion of identification in Section 4.5 above, these descriptive findings are reassuring for two reasons. First, if forced choosers were very different from optional choosers, it would be questionable whether the forced choosers can help to identify preferences for plan characteristics for the optional choosers. Second, the choice sets that optional and forced choosers face are similar on average, i.e., conditional on attention the two groups face similar choices, except that the choice sets of optional choosers include their old plans.

5.2 Specifications of the choice model

We specify and estimate four increasingly complex versions of the choice model introduced in Section 4. The features of these models are described in Table 7. All models allow for heterogeneity of switching costs by including a random coefficient on the dummy variable for the old plan (if it is still in individuals' choice sets). Model I is a standard mixed (random parameters) multinomial logit model with heterogeneous switching costs. As discussed in Section 2, the coefficient of a dummy variable for the old plan has been interpreted as a measure of switching costs in earlier literature. Importantly, this simple model attributes all inertia to switching costs. In Model II, we add an attention stage and thus allow for two different sources of inertia, inattention as well as switching costs. Model III adds the latent acuity variable in the attention stage, allowing for heterogeneity in how sensitive individuals are to attention triggers. Model IV further allows for an effect of acuity on plan choice, conditional on attention, which introduces heterogeneity of individuals with respect to acuity and heteroskedasticity at the plan-choice stage.

Model IV corresponds to the full two-stage model described in Section 4 while the other three models are nested sequentially. As all models are estimated by Maximum Likelihood, their predictive power can be readily assessed and compared. Table 7 reports the values of the Akaïke and Bayesian information criteria (AIC and BIC) as well as the likelihood-ratio test statistics for pairwise comparisons to Model IV. According to both the AIC

23

and the BIC, Model IV fits the data best, and the LR tests clearly reject the simpler models as well. Thus, the most comprehensive Model IV is our preferred model, and we focus the discussion of the estimation results and their substantive interpretation on this model. The discussion of the estimation results of the other three models concentrates on the question of whether using these simpler models would lead to different substantive conclusions with respect to Part D plan-choice behavior and the effects of interventions.

5.3 Estimated coefficients, marginal effects, and switching costs

Table 8 displays the estimates and their standard errors for the parameters of the attention stage and the acuity equation, and Table 9 those for the plan-choice stage, obtained from the preferred and most comprehensive Model IV. These tables also show the implied marginal effects of the covariates. For the attention stage and the acuity equation, we calculate the marginal effect on the probability to pay attention as well as on the probability to switch plans. For the plan-choice stage, we calculate the marginal effect on the probability to choose a plan conditional on attention as well as overall (i. e., not conditioning on attention).[19]

To facilitate the interpretation of the marginal effects reported below, it is useful to consider how the implied attention probabilities vary in the sample. Recall that attention is a latent binary outcome, and the probability of attention is heterogeneous, conditional on covariates. The attention probabilities implied by Model IV vary between 0.2% and 95.4%, with a median of 20.9%, and first and third quartiles of 17.1% and 25.3%, so there is considerable variation.[20]

Attention stage

The attention trigger variables fall into three groups: experience in Part D in $t-1$, changes in the features of the $t-1$ plan announced for year t, and health shocks as well as health-care use in $t-1$. All coefficients in the first two groups are statistically significant at the 0.001 level. Despite the large sample size, some of the health variables are not significant; they do not seem to raise enrollees' awareness of the opportunity to switch plans. In the following, we discuss the effect sizes of these covariates in terms of their marginal effects on the probability of attention. As switching requires attention, the marginal effects of the covariates on switching are always smaller than those on attention.

---

[19] Similar results for the other three models can be found in Tables A1 to A5 in the Appendix.

[20] We report and compare the attention probabilities implied by Models II–IV in Table 10 below.

Perhaps the most salient aspect of an individual's Part D experience is whether she hit the coverage gap and if so, whether she was enrolled in a plan that provides gap coverage. We thus include two dummy variables and their interaction. The marginal effect of having gap coverage in year $t-1$ for individuals who did not hit the gap in $t-1$ is -7.4 percentage points. For those who did hit the gap in $t-1$, it is 2.9 percentage smaller, for a total effect of -10.3 percentage points. Given a median attention probability of about 21%, this is a large effect: Individuals who hit the gap and do have gap coverage are much less likely to even consider switching for year t. Having hit the gap without gap coverage in year $t-1$ increases the attention probability by 5.8 percentage points, which also is a sizeable effect. Similarly, individuals who have hit the catastrophic coverage region in $t-1$, i.e., those with the largest drug bills, are more likely to be attentive as well (by 4 percentage points).

Recall that plans are required to send their enrollees ANOC letters before the open enrollment period each year. All changes of the features of the year $t-1$ plan should thus be salient. We find that individuals indeed are more likely to pay attention if the premium of their old plan increases, the deductible increases, the ICL increases, their old plan introduces changes to the formulary that would lead to an increase in OOP cost, or their old plan switches from copayments to coinsurance or *vice versa*. For instance, an increase of the year $t-1$ premium by 100 dollars results in a 6.1 percentage point increase of the attention probability. Interestingly, we find that premium changes are particularly salient, as the effect of a premium increase is much higher than the effects of an increase in the deductible or the annual OOP spending implied by a formulary change. The finding that individuals put more weight on premiums than on the cost implications of other plan features is in line with earlier research, including Abaluck and Gruber (2011) and Heiss *et al.* (2013). Furthermore, we find that the effect of the implied change in OOP cost due to formulary changes is much smaller (0.005 percentage point increase with a 100 dollar increase in OOP), than the effect of a change in the premium or the dedutible. This likely reflects that formulary changes are less salient than changes in premium and deductible.[21]

Attention probabilities are lower if a $t-1$ plan is consolidated and higher if there is a change in cost-sharing (from copay to coinsurance and *vice versa*). Such changes might make a plan more or less attractive, so we do not have a strong prior on the sign of these effects. Further, it might seem counter-intuitive that the probability of attention

---

[21] While the ANOC letters contain summary tables on changes in premiums, deductibles and copayments, the exact formulary changes, i.e., changes in the list of drugs and their tiers, are only provided in an Appendix to the letter.

increases with an increase of the ICL (as a higher limit indicates that the plan gets better) and with a lower share of tiers with increases in cost-sharing. However, as the regression already controls for change in total OOP spending, the sign of these effects are difficult to interpret as well.

One would expect that having been diagnosed with a new health condition or having had emergency room visits or hospital stays in the year $t-1$ raises attention to the Medicare Part D decision. However, this is not borne out by the data. The variables that measure health shocks and health-care use are mostly statistically insignificant. The only statistically significant but small effect we find is that of a dummy variable for whether an individual had five doctor visits or more in the previous year. As the model controls for the other health variables, this finding suggests that repeated interaction with a health professional raises attention. The data do not allow us to pursue such potential information channels further.

Acuity equation

The final covariate that enters the attention equation is acuity – a latent variable that captures how careful individuals are in making their choices. It is modeled as a function of observable individual characteristics and an unobserved heterogeneity term; in Model IV, it enters the plan-choice stage as well. It is statistically significant in both equations, and as the information criteria in Table 7 show, it also improves the overall predictive power. The variables that determine acuity also affect attention and plan choice indirectly. This indirect channel is accounted for in the marginal effects we report for the covariates in the acuity equation. As the effects of these covariates are modeled using a linear index in the attention equation, the indirect effect a predictor of acuity has on attention is 2.568 times its coefficient in the acuity equation.

Turning to the acuity equation itself, the coefficient estimates imply higher levels of acuity for males, whites, younger individuals, individuals who have less experience with Part D, individuals who live in ZIP code areas with higher average education among seniors, those living in middle-income ZIP codes, and those who have no depression diagnosis in their claims history.

The difference in acuity between males and females is small, and statistical significance is marginal given our sample size. The effect of being non-white, compared to all other groups, is large as this group's attention probability is lower by 8.4 percentage points, on average. Compared to individuals younger than 70, the attention probabilities of

individuals aged between 70 and 80, and of those older than 80, are 1.1 and 4.8 percentage points lower, respectively.

An important issue in the literature on Medicare Part D is whether choices improve over time as consumers learn to navigate this rather complicated market. The prior literature on this question delivered mixed results, see Ketcham *et al.* (2012) and Abaluck and Gruber (2016a). We find that for each additional year an individual has been enrolled in Part D, the attention probability is reduced by 1.6 percentage points, *ceteris paribus*. The fact that individuals become less attentive to plan switching over time, conditional on all other variables including attention triggers, suggests that individuals' decision-making does not improve over time. Rather, acuity declines, *ceteris paribus*.

It would be interesting to test whether acuity is predicted by measures of cognitive ability, decision making competence, or financial literacy, see Barcellos *et al.* (2014), among others. Measures of these constructs are not available in the Medicare claims data. However, the attention probability of individuals who have been diagnosed with depression is lower by 2.9 percentage points. This partial measure of mental health, which might also be related to decision making competence, predicts at least part of the heterogeneity in Medicare Part D decision-making.

Our proxies for socio-economic status, constructed at the ZIP code level, also predict acuity and therefore attention. An increase of the share of seniors without a high-school degree by 1 percentage point is associated with a decrease in the attention probability of 0.116 percentage points. As ZIP code level income measures, we include the shares of poor (income below the FPL) and rich (income above 500% of the FPL) seniors. Attention probabilities are largest for the reference group of individuals with incomes between 100% and 500% of the FPL, as the poor and rich shares increase by one percentage point, attention probabilities decline by 0.05 and 0.132 percentage points. As these ZIP-code level measures are noisy, the relatively small magnitudes of the marginal effects are not surprising. Nevertheless, they suggest that socio-economic status might play a role in determining the quality of Medicare Part D plan choices. We speculate that the relatively poor show lower levels of acuity because of lower decision-making competence. Among the relatively rich, acuity might be lower because they feel less inclined to worry about plan choice in Part D, which in turn might reflect their opportunity cost of time. This interpretation is consistent with the finding that richer individuals tend to select high-premium plans in Medicare Part D, independent of their risk; see Heiss *et al.* (2013).

Plan-choice stage

Turning to the results for the plan-choice stage, shown in Table 9, note first that in our two-stage model, plan choices are made only conditional on attention. Our model estimates imply an average attention probability of 23% while there is a 77% probability of being inattentive. Those individuals who are inattentive stay in their old plan and do not enter the choice stage. For the others, plan choices are modelled by a mixed (random parameters) multinomial logit specification with heteroskedasticity. In our preferred Model IV, the variance of the errors is a function of acuity. The results for the plan-choice stage show sensible signs for all coefficients, and to the extent that they are comparable, they are also in line with earlier results from static plan-choice models such as Heiss *et al.* (2013) and Abaluck and Gruber (2011). In particular, individuals are more likely to choose a plan if it has lower OOP, lower premium, lower deductible, higher ICL, and offers gap coverage. Also, the effect of a plan's premium is three times as large as that of its implied OOP cost, which confirms earlier findings that individuals are too sensitive to the premium as the most salient feature, see e.g. Abaluck and Gruber (2011) and Abaluck and Gruber (2016a).

In addition to these plan features, the plan-choice model also contains a dummy variable for the previous year's plan. It has a statistically significant and qualitatively important effect. We derive an estimate of switching costs from this coefficient below. Finally, we find that individuals are more likely to choose a plan whose implied OOP cost variance is lower, suggesting that individuals are risk averse and value the insurance component of Medicare Part D. To illustrate the size of the effect, our estimates imply that individuals have a willingness to pay of $163 for a *ceteris paribus* reduction of the OOP cost variance from the 75th to the 25th percentile.

As before, the further quantitative interpretation of the results focuses on marginal effects, which we compute both unconditional and conditional on attention.

Not conditioning on attention, the probability that a plan is chosen decreases by 3.3 percentage points when the annual premium increases by $1000, or by 0.33 percentage points with a more realistic increase in the annual premium by $100. Given that individuals choose, on average, between roughly 50 plans and thus a priori the probability of choosing a plan is roughly 2 percent, this corresponds to a sizable 16.5% increase.

When conditioning on attention, marginal effects get larger in absolute value, except for the marginal effect of the old plan dummy, which captures switching costs. This increase reflects the fact that the effects of the plan characteristics are watered down by inattention when not conditioning on attention. That is, when we consider all individuals, a plan's

annual premium has a lower impact on the probability that the plan is chosen than if we consider only individuals who pay attention. Those who are inattentive do not even consider the plan's premium when making their choice.

At this point, it is useful to recall the interpretation of the coefficient of the old plan dummy variable as a measure of switching costs. The raw model parameters are measured in 'utility units'. We can translate them into dollar units because a price variable (plan premium) is included among the regressors. The ratio of the coefficient of the old plan dummy variable and that of the premium is the implied compensation for staying in the old plan conditional on all other covariates. This yields our estimate of switching costs as $-\frac{3.238}{-9.565} \times \$1000 = \$338.53$.

These switching costs imply an increase of the probability to choose the $t-1$ plan by 11.5 percentage points conditional on attention. Not conditioning on attention, the overall probability of choosing the $t-1$ plan is 79.9 percentage points higher than the probability of choosing an otherwise identical plan. This large unconditional effect arises because all inattentive individuals stay in the $t-1$ plan by definition.

Finally, the estimate of δ_v is statistically significant. As the sign is negative, individuals with higher levels of acuity have a lower variance of the error term at the plan-choice stage, see the specification of plan utility in equation (4). Loosely speaking, they make better plan choices in the sense that the observed plan characteristics such as premium and OOP cost have a stronger impact on choice probabilities. Conversely, individuals with lower acuity have a larger variance of the error term and thus in the limit choose randomly among available plans.

5.4 Comparison of Models I–IV

In this section, we compare the four models in terms of their behavioral implications. In particular, it turns out that relative to our preferred Model IV, the simpler Models I–III lead to very different estimates of switching costs, inattention, and inertia. Thus, the four models would also have different implications for policy interventions, discussed in Section 6 below, aimed at reducing inertia.

We begin by comparing the attention probabilities implied by Models II–IV in Table 10. As we make the choice model more complex, the attention probabilities become smaller, at the mean and also across the entire distribution. In particular, allowing for unobserved heterogeneity in acuity in Model III and Model IV reduces the mean attention probability

by more than half, compared with Model II: Going from Model II to IV a larger share of inertia is attributed to inattention rather than switching costs.

Tables 11 and 12 show the marginal effects of the covariates on the probability to pay attention and the probability to switch plans, respectively, for all models that contain an attention stage, i.e., Models II–IV. Table 13 reports marginal effects on plan-choice probabilities for all four models.

In Table 11, most marginal effects get smaller when going from the simpler to the more complex models. Ignoring the role of unobserved heterogeneity in acuity and not taking the effect of acuity on plan choice into account tends to lead to an overestimation of the effects of attention triggers and demographics on attention. A particularly interesting difference arises for the marginal effect of experience with Part D: While in Model II, the probality to pay attention is not related to the number of years of experience with part D, there is a small but significantly positive relationship in Model III, and – as discussed above – a negative relationship in Model IV. Whether and how heterogeneity is allowed for in the model may thus be important for conclusions concerning the development of choice quality over time. The marginal effects reported in Table 12 are rather similar across models. If we are only interested in the effect of observed covariates on overall switching rates and do not intend to study policy interventions aimed at different aspects of inertia, the simpler models appear to be sufficient.

The four models lead to quite different estimates of switching costs, see Figure 1. As illustrated above for Model IV, the implied switching costs are calculated based on the estimation results for the plan-choice stage as the ratio of the coefficients of the $t-1$ plan dummy and the annual premium. They express how much individuals are, on average, willing to pay in order to stay in their old plan. Model I attributes all inertia to switching costs. It yields large average switching costs of more than \$1000. When we allow for inattention as a source of inertia, estimated switching costs are reduced substantially to roughly \$690. Given that average premiums are only roughly \$550 per year this is still a rather high number. Further reductions in estimated switching costs arise when we allow for unobserved heterogeneity in the attention stage (Model III) and in the plan-choice stage (Model IV). In the full model, average switching costs amount to roughly \$340 per year (the number reported above).

Table 13 displays the average marginal effects of the different plan characteristics on the probability that a plan is chosen, unconditional and conditional on attention (see the

top and bottom panels, respectively).[22] As Model I has no attention stage, the marginal effects of the regressors on the conditional and unconditional probability of choosing a plan are necessarily the same.

The average marginal effects of the covariates on the probability to choose a plan conditional on attention tend to increase in absolute value as we move from Model I to Model IV. This reflects that in the simpler models, the effects of plan characteristics on choice are watered down by inattention or neglected heterogeneity. The effects on the unconditional probabilities (which in Models II–IV integrate out the probability that individuals pay attention to plan choice) are very similar across models. The only exception to these patterns is the marginal effect of the year $t-1$ plan dummy (to be precise, the marginal effect of the mean of the random coefficient) whose patterns are more complex. Specifically, we quantify inertia as the additional probability of choosing the $t-1$ plan over an otherwise identical plan with the same features. Conditional on attention, this marginal effect reflects inertia due to switching costs alone. The unconditional marginal effect comprises both sources of inertia, switching costs in the second stage and inattention in the first stage. Figure 2 visualizes what the conditional and unconditional marginal effects imply for inertia, i. e., the additional probability of staying in the $t-1$ plan compared to another plan with identical plan features.

The height of the columns of Figure 2 corresponds to the unconditional marginal effects reported in the bottom panel of Table 13. They thus reflect the overall level of inertia estimated by each model. These are broken down by source, inattention and switching costs, as reflected in the conditional marginal effects. Model I does not include a separate attention stage, which results in a low estimate of inertia that is entirely attributed to switching costs. The other models produces larger estimates of inertia, about 80 percentage points in Models III and IV. These models differ in how much of this considerable inertia effect is attributed to inattention and switching costs, respectively. In the full model, the contribution of switching costs is relatively small while inattention is the major source of inertia.

These results correspond to the observation made above that the simpler models – most dramatically, Model I which does not allow for inattention – overestimate switching costs and underestimate total inertia.

---

[22] The underlying coefficients and standard errors for Model IV were discussed above (Tables 8 and 9); those for Models I–III are displayed in Tables A1 to A5 in the Appendix.

6 Simulations: Forcing attention and removing switching costs

Several policy interventions aimed at increasing switching rates have been suggested in the literature, both in Medicare Part D and in other markets where consumer inertia is an issue. Such proposals typically aim to reduce the transaction costs associated with switching plans (e. g., Hoadley *et al.* (2013)); increase attention using reminders, nudges, and other information shocks (e. g., Sinaiko *et al.* (2013)); or reduce the complexity of the choice environment, for instance by reducing the number of available options (e. g., Frank and Lamiraud (2009), Besedeš *et al.* (2015)).

We use the estimates of our four models to simulate the effects of two different policy interventions on the probability that individuals switch plans and on *ex post* overspending. In the first policy intervention, we force all individuals to pay attention. In the second policy intervention, we set switching costs to zero. These two scenarios are extreme and therefore allow us to bound the effects of more realistic interventions that increase attention or reduce switching costs. Our policy conclusions will be based on the most comprehensive and statistically superior Model IV. As discussed in section 5, Models I–III which are nested in Model IV are rejected both by likelihood ratio tests and perform much worse in explaining observed choices according to the AIC and BIC information criteria. However, it is not clear *a priori* whether this translates into strongly biased conclusions when they are used for policy simulations. It is therefore instructive to simulate the effects of the policy interventions for these three models as well.

Figures 3 and 4 display the simulation results. For Models I–IV, we report the average probability to switch and average overspending for three scenarios: the baseline case (i. e., using the estimates based on the observed choice data as reported in Section 5) and the two policy interventions. In our preferred Model IV, forcing all individuals to pay attention results in a dramatic increase in the switching probability, from 11.0% to 69.1%. An intervention that reduces switching costs to zero but lets attention unchanged has a smaller but still relevant effect on the average switching probability, which increases from 11.0% to 19.5%. These simulation results reflect the fact that persistence in observed plan choices is largely driven by inattention and not switching costs, as discussed above. The relative importance of attention and switching costs is also reflected in the simulated effects of the interventions on average *ex post* overspending, which results directly from the increased switching rates. From a baseline estimate of \$357, forcing attention reduces overspending by about 20% to \$284 while setting switching costs to zero results in a small reduction in average overspending to \$347.

As Model I does not include an attention stage, forcing individuals to pay attention affects neither the probability to switch nor average overspending. Setting switching costs to zero has much larger effects on the switching probability and on average overspending in Model I than in Model IV. This is not surprising as in Model I, switching costs are the only mechanism that could generate inertia. The simulated effects are very large: The switching probability increases from the baseline estimate of 10.7% to 80.4% while overspending is reduced from $354 to $262. Going from Model I to Model IV, the impact of forcing attention becomes gradually larger while the impact of reducing switching costs to zero is reduced. This is in line with the results in Figure 1 that switching costs decrease when introducing an attention stage and unobserved heterogeneity into the model.

Forcing everybody to pay attention results in different switching rates and average overspending when comparing Models III and IV as in the latter, individuals with low acuity both have a lower attention probability and put less weight on OOP costs in the choice stage. This induces selection; in particular, an intervention that forces everybody to pay attention results in lower average acuity in the plan-choice stage and therefore poorer choices, on average. Our simulations show that this effect is quantitatively important. While 46% are induced to switch by forced attention in Model III, 69.1% are in Model IV. Nevertheless, average overspending is *lower* in Model III ($273) than in Model IV ($284).

The simulation results obtained from Model IV provide additional insights on the heterogeneity of the effects of policy interventions. Figure 5 shows the empirical cumulative distribution functions (CDFs) of overspending in the baseline scenario and in the two intervention scenarios. Forcing attention results in a much more compressed distribution of overspending. This illustrates the fact that attention is the most important source of heterogeneity as it prevents a large share of individuals from making any change.

Figure 6 displays the empirical CDFs of the reduction in overspending that is induced by the two policy interventions relative to the baseline case. Not surprisingly given the relatively minor impact of switching costs in Model IV, the change in overspending is close to zero for most individuals when switching costs are removed. When individuals are forced to pay attention overspending decreases on average, but it increases for about a third of the individuals in the sample.

An important implication of our simulation results is that while some individuals profit when they are forced to pay attention, others are harmed – at least with respect to *ex post* overspending. For policymakers, the balance between winners and losers of such reforms as the ones we simulate in this section is of particular interest. The magnitudes of these

effects depend on whether heterogeneity in acuity is allowed for, as we do in our preferred Model IV.[23]

7   Summary and conclusions

In Medicare Part D and other health insurance markets, only few consumers switch plans across years and thereby forego significant savings. In this paper, we separated channels that contribute to this high level of inertia: inattention to plan choice and perceived costs of switching plans. To characterize these two sources of inertia, we developed a two-stage panel data model of plan choices. The first stage models whether individuals pay attention to their plan choice or not. Given that individuals pay attention, the second stage models the plan choice in a discrete choice framework. The model includes switching costs and allows for unobserved individual-specific heterogeneity that leads to correlation of the two stages.

Our results suggest that in Medicare Part D, attention to the possibility of switching plans is triggered by events and experiences that are salient during the open-enrollment period. In particular, for those individuals who do not have gap coverage in their current plan, having reached the gap is associated with a higher propensity to switch. Premium and formulary changes also trigger attention, but the effect of premium changes is much larger than that of changes in OOP costs induced by changes to the formulary. This finding is interesting since formulary changes are clearly relevant for switching decisions but not as salient as premium changes. An implication for firm behavior is that consumers' reactions to increases in their plan's costs will be larger when they are implemented via the premium rather than other characteristics such as copayments and obscure details of the formulary benefit design, as for instance the tier assignment of specific drugs. Furthermore, new plans in the market may not be considered to the extent that their market entry is not salient to consumers.

Another interesting finding is that acuity declines over time, which seems inconsistent with the notion that people learn. This result corroborates recent findings Abaluck and Gruber (2016a) who also argue that choice inconsistencies in Medicare Part D increased over time. Also, individuals who pay attention in the baseline scenario have higher acuity and therefore make better decisions, at least in terms of total costs associated with their plan

---

[23] In further simulations, we computed the effects of the two interventions using Models I–III. These simulations illustrate how modeling assumptions – in particular, the way how heterogeneity enters a structural choice model – affect the policy conclusions. Detailed results can be found in the Appendix.

choice and current prescription drug use, than those who do not pay attention. Finally, our estimates indicate that even conditional on attention, there are still considerable implicit switching costs. Understanding the sources of these switching costs is an important topic for future research. These open issues not withstanding, our two-stage model opens up the possibility to simulate the effects of policies that aim at increasing plan switching by making the economic determinants of optimal health insurance plan choices more salient.

The simulations we performed using the estimates of our full two-stage choice model provide several interesting insights that differ from those one would obtain from simpler ones. We found that removing inattention, in our case by forcing every beneficiary to make an active choice, has much larger effects on switching rates than removing switching costs. This important finding echoes Kiss (2014) who found that in the Hungarian car insurance market, a public campaign aimed at increasing switching rates mainly acted through manipulating attention, rather than switching costs. Recent literature has stressed that reducing the number of available plans is another intervention that would conceivably improve plan choices and increase switching rates. We leave to future research the issue of whether a structural choice model like ours could reliably capture the effects of changing the number of available options.[24]

Our simulations underscore the importance of heterogeneity in individual responses to policy interventions. Specifically, when we force everybody to pay attention in our simulations, average acuity in the choice stage decreases, resulting in higher switching rates but, on average, also in worse choices in terms of total costs: At least some individuals end up in a worse plan when forced to make a choice. Such effects make the evaluation of the welfare effects of interventions much harder. We also found that most of those beneficiaries who switch do not choose the least-cost plan, a result that is in line with descriptive evidence by Hoadley *et al.* (2013).

The welfare consequences of consumer inertia in health insurance markets are complex and depend on the specific setting. Our findings add to the existing evidence on the excess costs associated with suboptimal plan choices and on the importance of inertia observed in health insurance markets. It seems obvious that consumers should be encouraged to review their health plans more frequently. However, Handel (2013) argues that policies that nudge consumers to better decisions by reducing inertia also exacerbate adverse selection, which potentially leads to a reduction in overall welfare. A further issue is

---

[24] The key issue here is that in data on Medicare Part D, the number of available plans has varied rather little over time. In experimental studies such as Besedeš *et al.* (2015), the number of options can be manipulated over a much larger range.

that removing frictions might lead to market unravelling, but as Handel and Kolstad (2015) argue, risk-adjustment could mitigate the negative effects of reducing information frictions. Handel *et al.* (2015) explore this issue further. We leave an extension of our present analysis that would allow for the computation of welfare effects to future research. We note, however, that in addition to these issues, such an extension would necessarily involve additional functional from assumptions, as pointed out in a related context by Ketcham *et al.* (2015) and Abaluck and Gruber (2016b).

Another area of research where our two-stage model might be fruitfully employed is the industrial organization of health insurance markets. Polyakova (2016) develops a model of plan choice that allows both for private information about health risk (giving rise to adverse selection) and switching costs. She argues that in a setting such as Medicare Part D, switching costs help sustain an equilibrium with adverse selection characterized by large differences in risk between more and less generous plans. Our two-stage model could be used to study how different interventions that reduce inertia affect market equilibria and prices over time. However, as Abaluck and Gruber (2016a), Decarolis (2015), and Decarolis *et al.* (2015) argue, analyzing the choice set evolution is especially complicated due to the rules governing assignment of low-income subsidy enrollees to plans.

We close by pointing out the relevance of our empirical findings for the emerging literature on the effects of inattention on consumer choices, and their implications for firm behavior, see Grubb (2012) for an overview. A prominent example is the theory of context-dependent choice developed by Bordalo *et al.* (2013). In that model, "a consumer's attention is drawn to salient attributes of goods, such as quality or price. An attribute is salient for a good when it stands out among the good's attributes, relative to that attribute's average level in the choice set (or more broadly, the choice context). Consumers attach disproportionately high weight to salient attributes and their choices are tilted toward goods with higher quality/price ratios." Our empirical findings are very much in line with these predictions, and they add to the still relatively scarce evidence on the relevance of inattention and attention shocks for high-stakes financial decisions.[25]

---

[25] This evidence includes Stango and Zinman (2014) on bank overdraft fees, Crossley *et al.* (2016) on retirement savings, and Helmers *et al.* (2016) on online shopping.

References

Abaluck, J. and J. Gruber (2011): Choice inconsistencies among the elderly: Evidence from plan choice in the Medicare Part D program. *American Economic Review*, 101, 1180–1210.

Abaluck, J. and J. Gruber (2016a): Evolving choice inconsistencies in choice of prescription drug insurance. *American Economic Review*, 106(8), 2145–2184.

Abaluck, J. and J. Gruber (2016b): The robustness of tests for consumer choice inconsistencies: Reply to Ketcham, Kuminoff and Powers. *American Economic Review*, forthcoming.

Bach, P. B. and M. B. McClellan (2005): A prescription of a modern Medicare program. *New England Journal of Medicine*, 353, 2733–2735.

Barcellos, S. H., A. Wuppermann, K. G. Carman, S. Bauhoff, D. McFadden, A. Kapteyn, J. Winter, and D. Goldman (2014): Preparedness of Americans for the Affordable Care Act. *Proceedings of the National Academy of Sciences of the United States of America*, 111(15), 5497–5502.

Besedeš, T., C. Deck, S. Sarangi, and M. Shor (2015): Reducing choice overload without reducing choices. *Review of Economics and Statistics*, 97(4), 793–802.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013): Salience and consumer choice. *Journal of Political Economy*, 121(5), 803–843.

Crossley, T. F., J. de Bresser, L. Delaney, and J. Winter (2016): Can survey participation alter household saving behavior? *Economic Journal*, forthcoming.

Decarolis, F. (2015): Medicare Part D: Are insurers gaming the low income subsidy design? *American Economic Review*, 105(4), 1547–1580.

Decarolis, F., M. Polyakova, and S. P. Ryan (2015): The welfare effects of supply-side regulations in Medicare Part D. NBER Working Paper No. 21298.

Duggan, M., P. Healy, and F. Scott Morton (2008): Providing prescription drug coverage to the elderly: America's experiment with Medicare Part D. *Journal of Economic Perspectives*, 22(4), 69–92.

Einav, L., A. Finkelstein, and P. Schrimpf (2015): The response of drug expenditure to nonlinear contract design: Evidence from Medicare Part D. *Quarterly Journal of Economics*, 130(2), 841–899.

Ericson, K. M. M. (2014): Consumer inertia and firm pricing in the Medicare Part D prescription drug insurance exchange. *American Economic Journal: Economic Policy*, 6(1), 38–64.

Farrell, J. and P. Klemperer (2007): Coordination and lock-in: Competition with switching costs and network effects. In M. Armstrong and R. Porter (Eds.), *Handbook of Industrial Organization*, Volume 3, chapter 31, 1967–2002. Amsterdam: North-Holland.

Frank, R. and K. Lamiraud (2009): Choice, price competition and complexity in markets for health insurance. *Journal of Economic Behavior and Organization*, 71(2), 550–562.

Goldman, D. P. and G. F. Joyce (2008): Medicare Part D: A successful start with room for improvement. *Journal of the American Medical Association*, 299(16), 1954–1955.

Goodman, J. (2006): Consumer directed health care. Policy Brief 2006-PB-20, Networks Financial Institute, Indiana State University.

Grubb, M. D. (2012): Dynamic nonlinear pricing: Biased expectations, inattention, and bill shock. *International Journal of Industrial Economics*, 30, 287–290.

Handel, B. (2013): Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7), 2643–2682.

Handel, B. and J. T. Kolstad (2015): Health insurance for "humans": Information frictions, plan choice, and consumer welfare. *American Economic Review*, 105(8), 2449–2500.

Handel, B. R., J. T. Kolstad, and J. Spinnewijn (2015): Information frictions and adverse selection: Policy interventions in health insurance markets. NBER Working Paper No. 21759.

Heckman, J. J. (1981): Heterogeneity and state dependence. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data and Econometric Applications*, 91–139. Cambridge, MA: MIT Press.

Heiss, F., A. Leive, D. McFadden, and J. Winter (2013): Plan selection in Medicare Part D: Evidence from administrative data. *Journal of Health Economics*, 32, 1325–1344.

Heiss, F., D. McFadden, and J. Winter (2006): Who failed to enroll in Medicare Part D, and why? Early results. *Health Affairs*, 25, w344–w354.

Heiss, F., D. McFadden, and J. Winter (2010): Mind the gap! Consumer perceptions and choices of Medicare Part D prescription drug plans. In D. A. Wise (Ed.), *Research Findings in the Economics of Aging*, 413–481. Chicago and London: University of Chicago Press.

Heiss, F. and V. Winschel (2008): Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144, 62–80.

Helmers, C., P. Krishnan, and M. Patnam (2016): Attention and saliency on the internet: Evidence from an online recommendation system. CEPR Discussion Paper No. 10939.

Ho, K., J. Hogan, and F. Scott Morton (2015): The impact of consumer inattention on insurer pricing in the Medicare Part D program. NBER Working Paper No. 21028.

Hoadley, J., E. Hargrave, L. Summer, J. Cubanski, and T. Neuman (2013): To switch or not to switch: Are Medicare beneficiaries switching drug plans to save money. Issue Brief, The Henry J. Kaiser Family Foundation, Menlo Park, CA.

Honka, E. (2014): Quantifying search and switching costs in the US auto insurance industry. *RAND Journal of Economics*, 45(4), 847–884.

Hortaçsu, A., S. A. Madanizadeh, and S. L. Puller (2015): Power to choose? An analysis of consumer inertia in the residential electricity market. NBER Working Paper No. 20988.

Ketcham, J. D., N. V. Kuminoff, and C. A. Powers (2016): Choice inconsistencies among the elderly: Evidence from plan choice in the Medicare Part D program: Comment. *American Economic Review*, forthcoming.

Ketcham, J. D., C. Lucarelli, E. J. Miravete, and M. C. Roebuck (2012): Sinking, swimming, or learning to swim in Medicare Part D. *American Economic Review*, 102, 2639–2673.

Ketcham, J. D., C. Lucarelli, and C. A. Powers (2015): Paying attention or paying too much in Medicare Part D. *American Economic Review*, 105, 204–233.

Kiss, A. (2014): Salience and switching. Unpublished manuscript.

Kling, J. R., S. Mullainathan, E. Shafir, L. Vermeulen, and M. V. Wrobel (2012): Comparison friction: Experimental evidence from Medicare drug plans. *Quarterly Journal of Economics*, 127(1), 199–235.

Luco, F. (2014): Switching costs and competition in retirement investment. Unpublished manuscript, Northwestern University.

McFadden, D. (1974): Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, 105–142. New York: Academic Press.

McFadden, D. and K. Train (2000): Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447–470.

Miller, D. P. and J. Yeo (2012): Estimating dynamic discrete choice models of product differentiation: An application to Medicare Part D with switching costs. Unpublished manuscript.

Nosal, K. (2012): Estimating switching costs for Medicare Advantage plans. Unpublished manuscript, University of Mannheim.

Polyakova, M. (2016): Regulation of insurance with adverse selection and switching costs: Evidence from Medicare Part D. *American Economic Journal: Applied Economics*, 8(3), 165–195.

Sinaiko, A. D., C. C. Afendulis, and R. G. Frank (2013): Enrollment in Medicare Advantage plans in Miami-Dade County: Evidence of status quo bias? *Inquiry*, 50(3), 202–215.

Stango, V. and J. Zinman (2014): Limited and varying consumer attention: Evidence from shocks to the salience of bank overdraft fees. *Review of Finance*, 27(4), 990–1030.

Zhou, C. and Y. Zhang (2012): The vast majority of Medicare Part D beneficiaries still don't choose the cheapest plans that meet their medication needs. *Health Affairs*, 31(10), 2259–2265.

Tables and figures

Table 1: Sample selection

	Entire 20%	Selection criteria	Estimation sample
2007	9,299,848		
2008	9,530,609	1,176,895	76,352
2009	9,781,213	1,215,532	80,546
2010	10,016,372	1,136,255	81,382

Notes: Selection criteria: US residents, aged 65+, enrolled in stand-alone non-employer PDP plan, do not receive low-income subsidies (non-LIS), not dual-eligible, enrolled in Part D continuously in 12 months in the reference year (t) and prior year ($t-1$) unless forced chooser (i. e., individual has no default plan). Estimation sample randomly selects 100,000 individuals who meet all selection criteria in at least one year. Because of later enrollment, death, or not meeting all selection criteria for the whole time period, the panel data set is unbalanced and the number of observations is smaller than 100,000 in each year.

Table 2: Optional and forced choosers

| | Forced choosers | Optional choosers | | |
| | | Stayers | Switchers | |
	(%)	(%)	(%)	N
2008	3.65	86.09	10.27	76,352
2009	3.65	86.29	10.06	80,546
2010	2.78	87.79	9.44	81,382
All years	3.35	86.73	9.91	238,280

Notes: Optional choosers are individuals who do have a default plan, forced choosers are those without default (i. e., because they are new to Part D (roughly 97% of forced choosers) or because their old plan terminates (roughly 3%)).

Table 3: Average *ex post* overspending

Year	Forced choosers	Optional choosers		
		Stayers	**Switchers**	**Total**
2008	280.5	325.0	220.4	312.6
2009	272.5	394.6	203.4	370.9
2010	255.3	396.1	202.6	374.0
All years	270.4	373.0	208.8	353.3

Notes: Based on estimation sample. *Ex post* overspending = total costs in chosen plan – total costs of cheapest available plan in year t, based on drugs used in year t.

Table 4: *Ex ante* overspending in old and new plan

	Stayers	Switchers	
	Old plan (t-1) = new plan (t)	Old plan (t-1)	New plan (t)
Mean			
2008	341.3	409.4	239.1
2009	405.6	416.5	220.1
2010	423.9	400.7	224.2
All years	391.5	409.0	227.7
Median			
2008	256.4	313.8	166.6
2009	334.0	346.9	144.5
2010	342.4	307.6	172.3
All years	313.9	322.8	163.7
Share with zero ex ante *overspending*			
2008	3.37	0.96	13.15
2009	1.84	0.91	17.64
2010	2.45	1.17	18.32
All years	2.54	1.01	16.37

Notes: Based on estimation sample. *Ex ante* overspending = total costs in (old or new) plan − total costs in cheapest plan, based on drugs used in $t-1$. As *ex ante* overspending is not available for most of the forced choosers, they are not included.

Table 5: Descriptive statistics – estimation sample

	Optional choosers				Forced choosers	
	Switchers		**Stayers**			
Attention Triggers:						
Part D experience in t-1:						
Gap coverage [D]	0.14	(0.35)	0.13	(0.34)		
Hit the gap [D]	0.22	(0.41)	0.18	(0.38)		
Gap coverage & hit the gap [D]	0.06	(0.23)	0.04	(0.21)		
Hit catastrophic region [D]	0.03	(0.17)	0.02	(0.15)		
Change in t-1 plan:						
Plan consolidated [D]	0.10	(0.31)	0.19	(0.39)		
Change of premium [\$1000]	0.13	(0.11)	0.07	(0.08)		
Change of deductible [\$1000]	0.01	(0.05)	0.00	(0.05)		
Change of ICL [\$1000]	0.15	(0.05)	0.15	(0.05)		
OOP effect of formulary change [\$1000]	0.05	(0.17)	0.04	(0.15)		
Tiers with increases in copayments [share]	0.38	(0.35)	0.42	(0.30)		
Switch between copay and coinsurance [D]	0.14	(0.35)	0.08	(0.27)		
Health shocks and health care use in t-1:						
Onset of costly condition [D]	0.10	(0.30)	0.10	(0.30)		
Onset of cheap condition [D]	0.25	(0.43)	0.25	(0.43)		
Five or more doctor visits [D]	0.77	(0.42)	0.76	(0.43)		
At least one ER visit [D]	0.24	(0.43)	0.25	(0.43)		
At least one hospital stay [D]	0.17	(0.37)	0.17	(0.38)		
Determinants of Acuity						
Male [D]	0.36	(0.48)	0.37	(0.48)	0.45	(0.50)
Non-white [D]	0.04	(0.20)	0.07	(0.25)	0.08	(0.28)
Age 70–79 years [D]	0.49	(0.50)	0.48	(0.50)	0.44	(0.50)
Age 80 years or older [D]	0.30	(0.46)	0.32	(0.47)	0.27	(0.44)
PDP experience [years]	2.83	(0.85)	2.87	(0.86)	0.36	(0.94)
Low education [ZIP-code share]	0.21	(0.11)	0.21	(0.12)	0.21	(0.12)
Low income [ZIP-code share]	0.09	(0.07)	0.09	(0.07)	0.08	(0.07)
High income [ZIP-code share]	0.24	(0.16)	0.25	(0.16)	0.26	(0.16)
Ever had depression [D]	0.24	(0.43)	0.24	(0.43)	0.23	(0.42)
N	23,623		206,672		7,985	

Notes: Means with standard deviations in parentheses. Dummy variables are marked with [D]. Pooled across years 2007/8-2009/10. Information on experience in old plan for forced choosers omitted as only available for small fraction of forced choosers (those whose old plan was terminated). Change in cost sharing is defined as the share of tiers on the formulary for which an increase in cost-sharing occurs. Dollar amounts are measured in 2010 dollars using the CPI.

Table 6: Descriptive statistics – choice sets in estimation sample

| | Optional choosers | | | | Forced choosers | |
	Switchers		Stayers			
# plans	51.79	(4.38)	51.11	(3.72)	51.65	(3.82)
OOP cost [$1000]	1.01	(1.19)	0.93	(1.08)	0.90	(1.07)
Plan characteristics						
Annual premium [$1000]	0.55	(0.04)	0.55	(0.04)	0.54	(0.04)
Deductible amount [$1000]	0.13	(0.02)	0.13	(0.02)	0.12	(0.02)
ICL amount [$1000]	2.75	(0.10)	2.75	(0.10)	2.74	(0.10)
No gap coverage [D]	0.76	(0.04)	0.76	(0.04)	0.75	(0.04)
Old plan [D]	0.02	(0.00)	0.02	(0.00)	0.00	(0.00)
Variance of OOP cost [($1000)^2]	10.27	(5.88)	9.95	(5.67)	9.55	(5.50)
N	23,623		206,672		7,985	

Notes: Means across averages in individuals' choice sets with standard deviations in parentheses. Dummy variables are marked with [D]. Pooled across years 2007/8-2009/10. OOP cost= out of pocket cost, ICL=initial coverage limit. Old plan is a dummy variable for the plan the individual was enrolled in in $t-1$. Dollar amounts are measured in 2010 dollars using the CPI.

Table 7: Estimated models

	Model			
	I	II	III	IV
Model components:				
Switching cost (random coefficient for old plan)	✓	✓	✓	✓
Attention stage		✓	✓	✓
Unobserved heterogeneity in acuity			✓	✓
Acuity enters choice stage				✓
Number of parameters	13	41	42	43
Log likelihood	-166,907.2	-163,429.8	-162,716.8	-162,004.0
LR test against model IV (df)*	9806.4 (30)	2851.6 (2)	1425.6 (1)	—
AIC	333,840.4	326,941.6	325,517.6	324,094.0
BIC	334,026.5	327,528.6	326,118.9	324,709.6

Notes: All models are estimated on the unbalanced estimation sample with $238,280$ observations on $100,000$ individuals. On average, they faced around 51.2 alternatives in each choice situation which translates into 12,199,769 alternatives in total.

*: All LR tests clearly reject the respective simpler model against Model IV.

Table 8: Results Model IV – attention and acuity

	Coefficient	SE		Marginal effects on attention	switching
Attention stage					
Constant	-1.656	(0.113)	***		
Gap coverage [D]	-0.965	(0.047)	***	-0.074	-0.042
Hit the gap [D]	0.681	(0.037)	***	0.058	0.036
Gap coverage & hit the gap [D]	-0.370	(0.067)	***	-0.029	-0.017
Hit catastrophic region [D]	0.460	(0.075)	***	0.040	0.024
Plan consolidated [D]	-0.153	(0.044)	***	-0.012	-0.007
Change of premium [$1000]	7.453	(0.116)	***	0.612	0.366
Change of deductible [$1000]	1.517	(0.225)	***	0.125	0.074
Change of ICL [$1000]	1.641	(0.336)	***	0.135	0.081
OOP effect of formulary change [$1000]	0.558	(0.068)	***	0.046	0.027
Tiers with increases in copayments [share]	-0.390	(0.043)	***	-0.032	-0.019
Switch between copay and coinsurance [D]	0.663	(0.040)	***	0.058	0.036
Onset of costly condition [D]	-0.014	(0.041)		-0.001	-0.001
Onset of cheap condition [D]	0.008	(0.028)		0.001	0.000
Five or more doctor visits [D]	0.086	(0.033)	**	0.007	0.004
At least one ER visit [D]	-0.011	(0.033)		-0.001	-0.001
At least one hospital stay [D]	-0.033	(0.038)		-0.003	-0.002
Acuity coefficient (δ_a)	2.568	(0.041)	***		
Determinants of acuity					
Male [D]	-0.035	(0.015)	*	-0.007	-0.003
Non-white [D]	-0.438	(0.031)	***	-0.084	-0.037
Age 70–79 years [D]	-0.050	(0.017)	**	-0.011	-0.005
Age 80 years or older [D]	-0.235	(0.021)	***	-0.048	-0.021
PDP experience [years]	-0.075	(0.008)	***	-0.016	-0.007
Low education [ZIP-code share]	-0.550	(0.076)	***	-0.116	-0.051
Low income [ZIP-code share]	-0.237	(0.109)	*	-0.050	-0.022
High income [ZIP-code share]	-0.628	(0.056)	***	-0.132	-0.058
Ever had depression [D]	-0.140	(0.017)	***	-0.029	-0.013
Unobserved Heterogeneity	1	(by normalization)			

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

Notes: Average marginal effects on the probability to pay attention and the probability to switch plans. Stars indicate whether coefficients are significantly different from zero. The attention stage model also includes year dummies.

Table 9: Results Model IV – plan choice

	Coefficient	SE		Marginal effects on plan choice	
				unconditional	conditional
Choice stage					
OOP cost [$1000]	-3.130	(0.066)	***	-0.011	-0.036
Annual premium [$1000]	-9.565	(0.189)	***	-0.033	-0.111
Deductible amount [$1000]	-5.216	(0.111)	***	-0.018	-0.061
ICL amount [$1000]	0.622	(0.157)	***	0.002	0.007
No gap coverage [D]	-0.569	(0.038)	***	-0.002	-0.008
Old plan coefficient mean (μ_γ)	3.238	(0.078)	***	0.799	0.115
Old plan coefficient variance (σ_γ^2)	1.385	(0.046)	***		
Variance of OOP cost [($1000)^2]	-0.219	(0.012)	***	-0.001	-0.003
Acuity coefficient (δ_v) in error var.	-0.497	(0.010)	***		

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

Notes: Average marginal effects on the probability that plan is chosen, unconditional and conditional on attention. Stars indicate whether coefficients are significantly different from zero. The plan choice model also includes five dummy variables for the largest plans.

Table 10: Probability of attention among optional choosers

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
II	0.004	0.374	0.464	0.455	0.556	0.994
III	0.001	0.176	0.216	0.224	0.263	0.979
IV	0.002	0.171	0.209	0.218	0.253	0.954

Notes: Models described in Table 7. Minimum, 1st Quartile, Median, Mean, 3rd Quartile and Maximum of attention probability implied by the models with attention stage.

Table 11: Marginal effects on probability to pay attention

	Model		
	II	**III**	**IV**
Attention Triggers			
Gap coverage [D]	-0.330 ***	-0.120 ***	-0.074 ***
Hit the gap [D]	0.117 ***	0.048 ***	0.058 ***
Gap coverage & hit the gap [D]	-0.041 **	-0.016 *	-0.029 ***
Hit catastrophic region [D]	0.037 *	0.018 *	0.040 ***
Plan consolidated [D]	0.078 ***	0.014 **	-0.012 ***
Change of premium [$1000]	1.028 ***	0.685 ***	0.612 ***
Change of deductible [$1000]	0.538 ***	0.212 ***	0.125 ***
Change of ICL [$1000]	0.301 ***	0.095 **	0.135 ***
OOP effect of formulary change [$1000]	0.110 ***	0.063 ***	0.046 ***
Tiers with increases in copayments [share]	-0.012	-0.022 ***	-0.032 ***
Switch between copay and coinsurance [D]	0.171 ***	0.080 ***	0.058 ***
Onset of costly condition [D]	-0.008	-0.005	-0.001
Onset of cheap condition [D]	-0.001	-0.000	0.001
Five or more doctor visits [D]	0.018 **	0.005	0.007 **
At least one ER visit [D]	-0.014 *	-0.006	-0.001
At least one hospital stay [D]	-0.010	-0.007	-0.003
Determinants of Acuity			
Male [D]	-0.021 ***	-0.012 ***	-0.007 *
Non-white [D]	-0.158 ***	-0.085 ***	-0.084 ***
Age 70–79 years [D]	-0.040 ***	-0.017 ***	-0.011 **
Age 80 years or older [D]	-0.087 ***	-0.045 ***	-0.048 ***
PDP experience [years]	0.003	0.006 *	-0.016 ***
Low education [ZIP-code share]	-0.137 ***	-0.099 ***	-0.116 ***
Low income [ZIP-code share]	-0.171 ***	-0.078 ***	-0.050 *
High income [ZIP-code share]	-0.279 ***	-0.141 ***	-0.132 ***
Ever had depression [D]	-0.017 **	-0.012 ***	-0.029 ***

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

Notes: Average marginal effects. Stars indicate whether underlying coefficients are significantly different from zero. No marginal effects for Model I shown as it has no attention stage or acuity.

Table 12: Marginal effects on probability to switch

	Model		
	II	**III**	**IV**
Attention Triggers			
Gap coverage [D]	-0.082 ***	-0.060 ***	-0.042 ***
Hit the gap [D]	0.026 ***	0.023 ***	0.036 ***
Gap coverage & hit the gap [D]	-0.009 **	-0.008 *	-0.017 ***
Hit catastrophic region [D]	0.008 *	0.008 *	0.024 ***
Plan consolidated [D]	0.018 ***	0.007 **	-0.007 ***
Change of premium [$1000]	0.233 ***	0.326 ***	0.366 ***
Change of deductible [$1000]	0.122 ***	0.101 ***	0.074 ***
Change of ICL [$1000]	0.068 ***	0.045 **	0.081 ***
OOP effect of formulary change [$1000]	0.025 ***	0.030 ***	0.027 ***
Tiers with increases in copayments [share]	-0.003	-0.011 ***	-0.019 ***
Switch between copay and coinsurance [D]	0.039 ***	0.038 ***	0.036 ***
Onset of costly condition [D]	-0.002	-0.002	-0.001
Onset of cheap condition [D]	-0.000	-0.000	0.000
Five or more doctor visits [D]	0.004 **	0.002	0.004 **
At least one ER visit [D]	-0.003 *	-0.003	-0.001
At least one hospital stay [D]	-0.002	-0.003	-0.002
Determinants of Acuity			
Male [D]	-0.005 ***	-0.006 ***	-0.003 *
Non-white [D]	-0.036 ***	-0.040 ***	-0.037 ***
Age 70–79 years [D]	-0.009 ***	-0.008 ***	-0.005 **
Age 80 years or older [D]	-0.020 ***	-0.022 ***	-0.021 ***
PDP experience [years]	0.001	0.003 *	-0.007 ***
Low education [ZIP-code share]	-0.031 ***	-0.047 ***	-0.051 ***
Low income [ZIP-code share]	-0.039 ***	-0.037 ***	-0.022 *
High income [ZIP-code share]	-0.063 ***	-0.067 ***	-0.058 ***
Ever had depression [D]	-0.004 **	-0.006 ***	-0.013 ***

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

Notes: Average marginal effects. Stars indicate whether underlying coefficients are significantly different from zero. No marginal effects for Model I shown as it has no attention stage or acuity.

Table 13: Marginal effects on probability to choose plan

	Model			
	I	**II**	**III**	**IV**
Conditional on Attention				
OOP cost [$1000]	-0.010 ***	-0.020 ***	-0.037 ***	-0.036 ***
Annual premium [$1000]	-0.029 ***	-0.061 ***	-0.115 ***	-0.111 ***
Deductible amount [$1000]	-0.018 ***	-0.039 ***	-0.069 ***	-0.061 ***
ICL amount [$1000]	0.003 ***	0.007 ***	0.011 ***	0.007 ***
No gap coverage [D]	-0.004 ***	-0.004 ***	-0.010 ***	-0.008 ***
Old plan coefficient mean (μ_γ)	0.334 ***	0.282 ***	0.192 ***	0.115 ***
Variance of OOP cost [($1000)^2]	-0.001 ***	-0.002 ***	-0.003 ***	-0.003 ***
Unconditional				
OOP cost [$1000]	-0.010 ***	-0.009 ***	-0.010 ***	-0.011 ***
Annual premium [$1000]	-0.029 ***	-0.029 ***	-0.030 ***	-0.033 ***
Deductible amount [$1000]	-0.018 ***	-0.019 ***	-0.018 ***	-0.018 ***
ICL amount [$1000]	0.003 ***	0.003 ***	0.003 ***	0.002 ***
No gap coverage [D]	-0.004 ***	-0.002 ***	-0.003 ***	-0.002 ***
Old plan coefficient mean (μ_γ)	0.334 ***	0.660 ***	0.801 ***	0.799 ***
Variance of OOP cost [($1000)^2]	-0.001 ***	-0.001 ***	-0.001 ***	-0.001 ***

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

Notes: Average marginal effects. Stars indicate whether underlying coefficients are significantly different from zero.

**Figure 1:** Estimates of switching costs
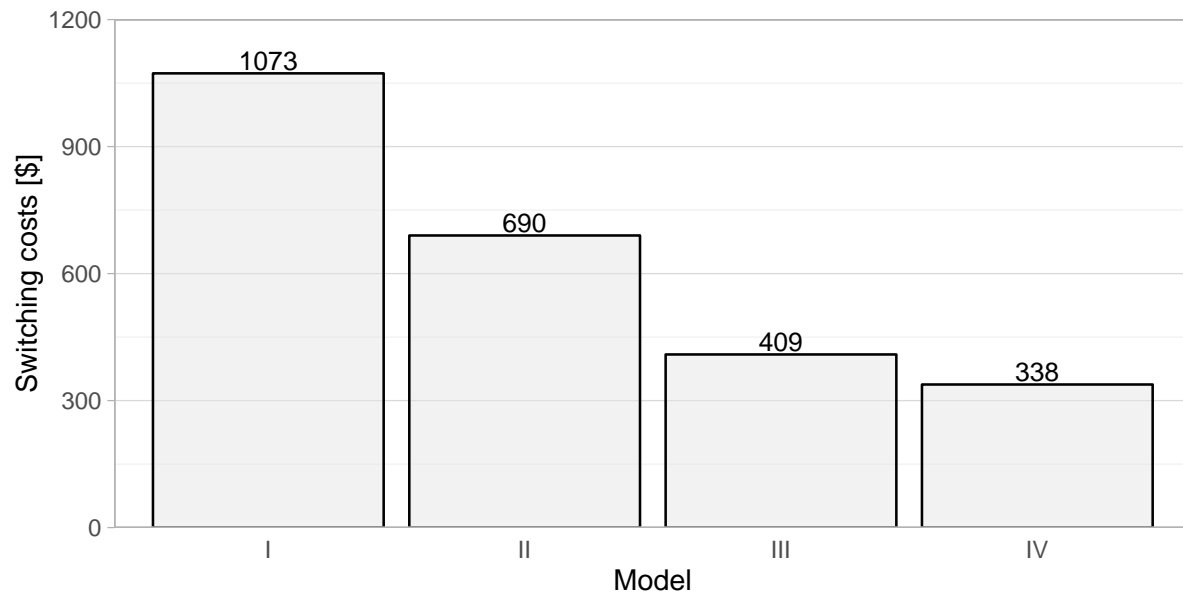
Notes: Based on coefficient estimates after estimation of Models I–IV. Ratio of the mean of the random coefficient for the old plan dummy and the coefficient of the premium. As premium measured in USD 1000, ratio multiplied by 1000.

Figure 2: Contribution of inattention and switching costs to inertia
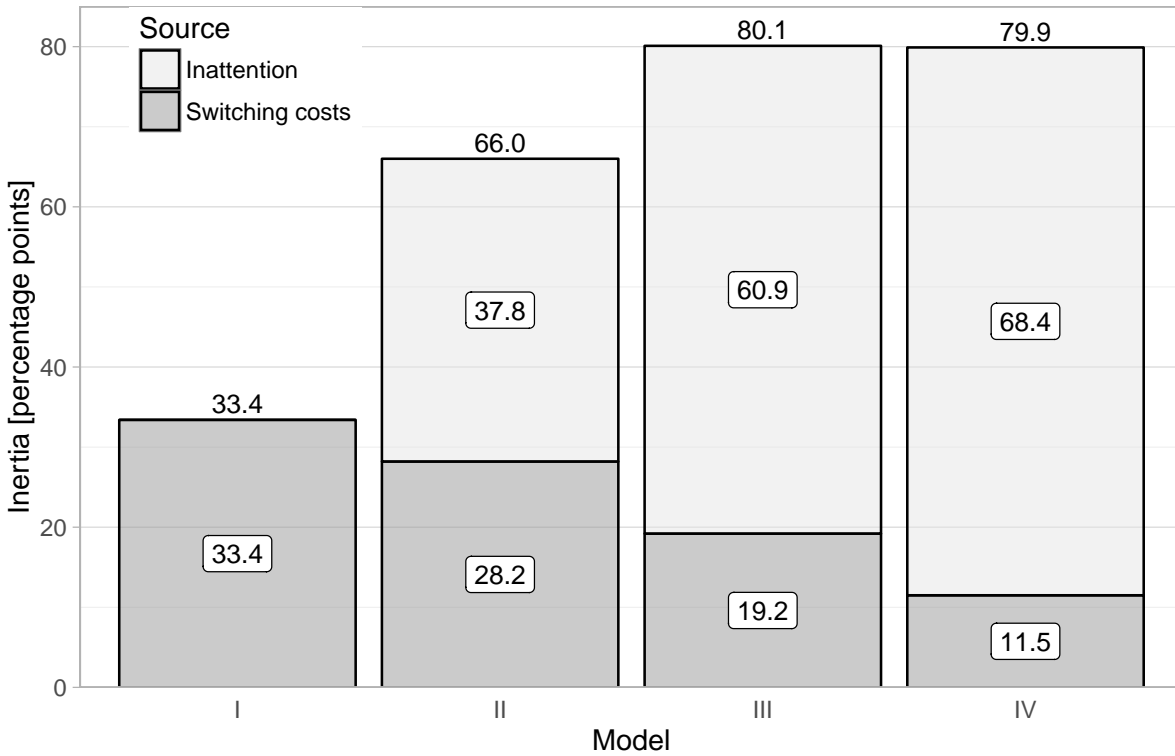
Notes: Overall inertia measured by the unconditional marginal effect of the mean old plan coefficient (bottom panel of Table 13). The respective marginal effect conditional on attention (top panel of Table 13) indicates how much of overall inertia can be attributed to switching costs.

Figure 3: Comparison of simulation results: Switching probability
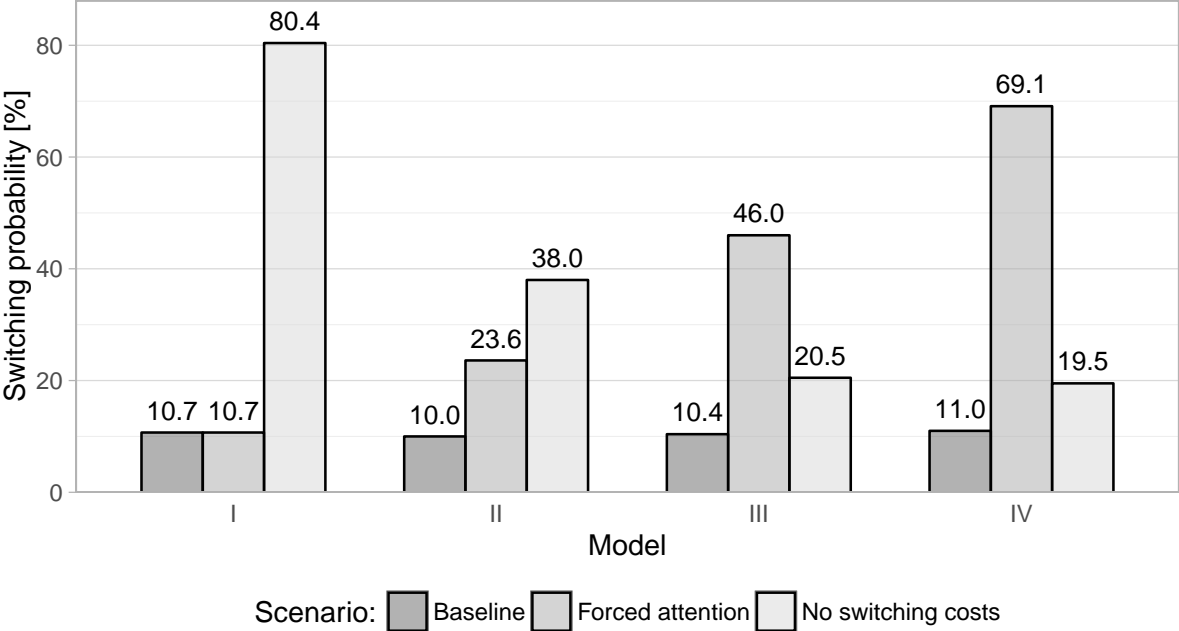
Notes: Policy simulation: Forced attention setting the probability of attention to 1 for everyone. No switching costs reduces switching costs to zero.

Figure 4: Comparison of simulation results: Overspending
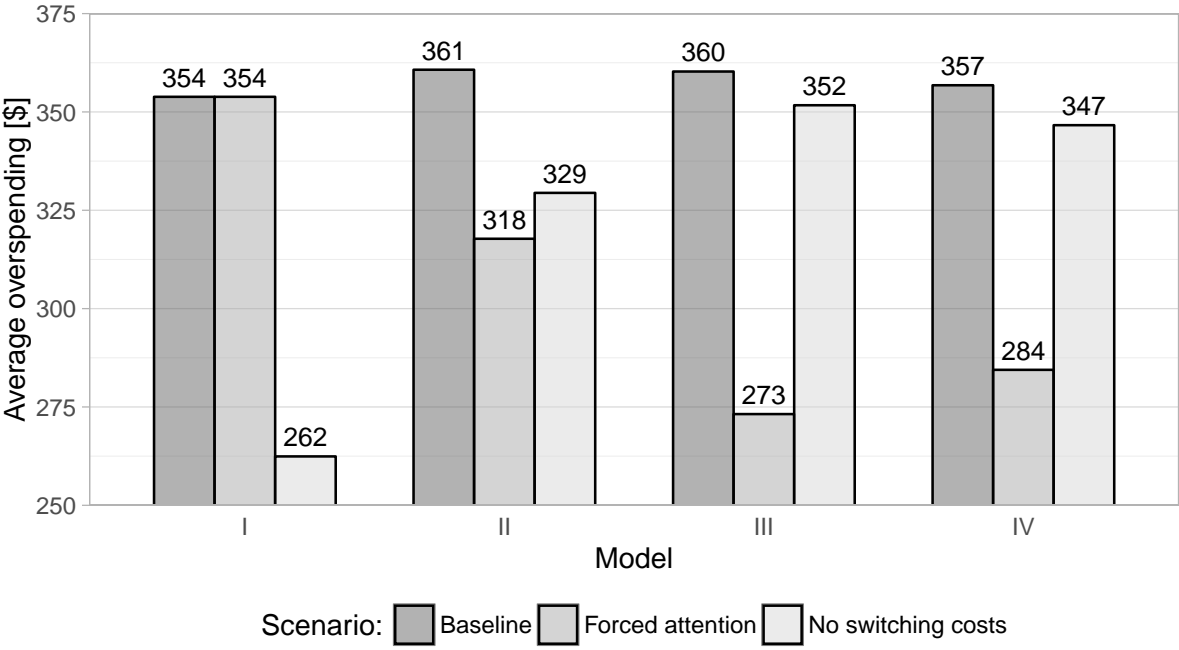
Notes: Policy simulation: Forced attention setting the probability of attention to 1 for everyone. No switching costs reduces switching costs to zero.

Figure 5: Predicted overspending in baseline and after policy interventions

Notes: Empirical CDFs based on Model IV. Policy simulation: Forced attention setting the probability of attention to 1 for everyone. No switching costs reduces switching costs to zero.

Figure 6: Predicted reduction in overspending induced by policy interventions

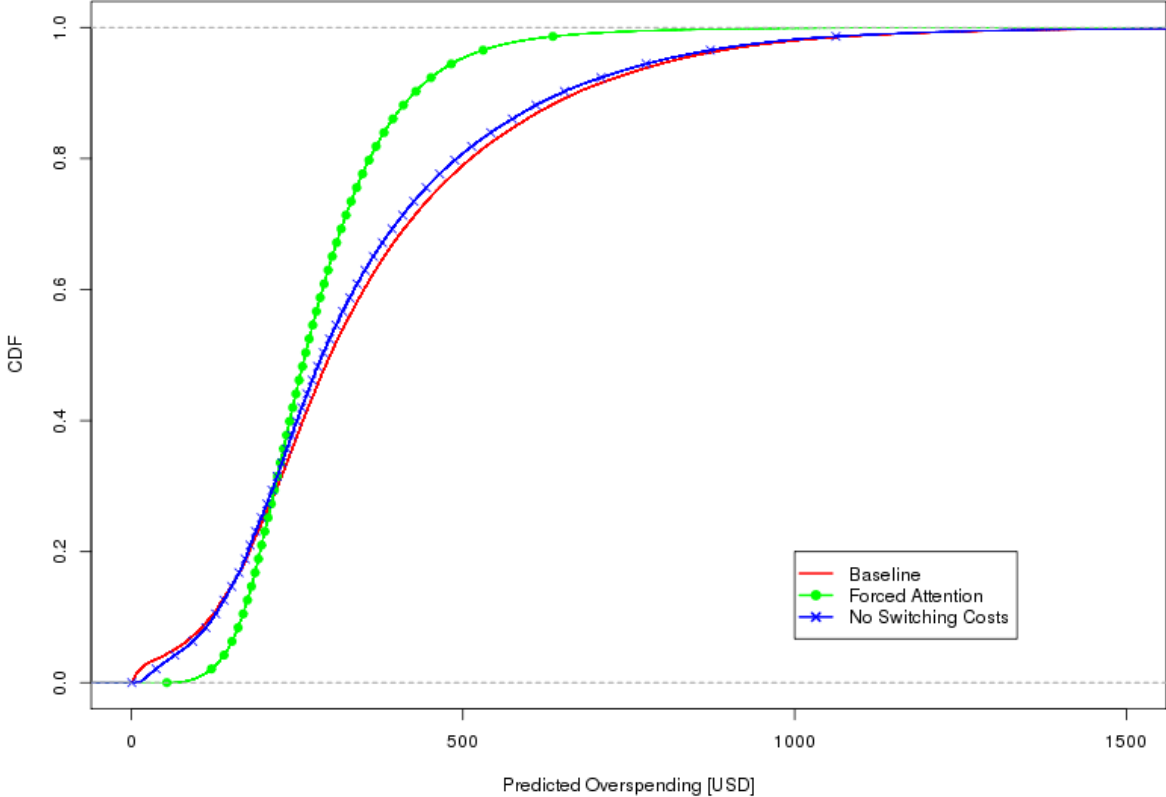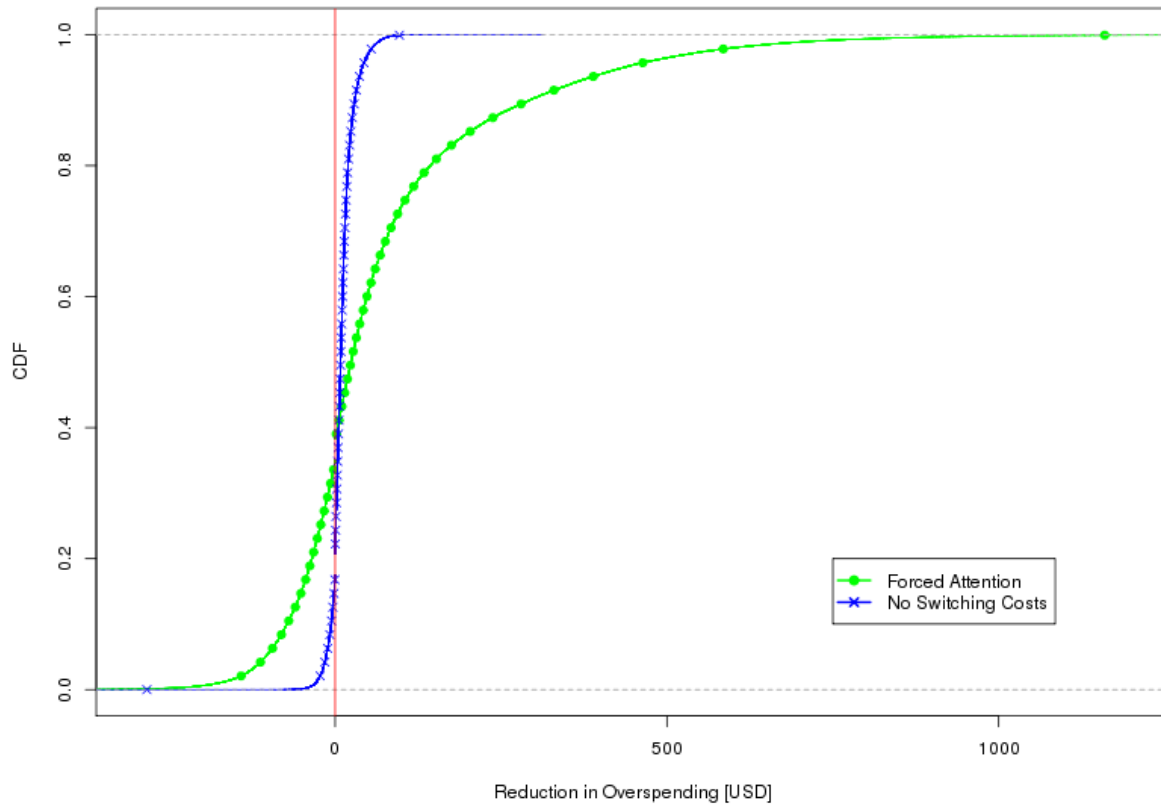*Notes:* Empirical CDFs based on Model IV. Overspending at baseline – overspending after policy simulations. Policy simulation: Forced attention setting the probability of attention to 1 for everyone. No switching costs reduces switching costs to zero.

Online Appendix

A.1 Estimation results for Models I–III

Tables A1–A5 report detailed estimation results for Models I–III. The results are presented analogously to the results for the preferred Model IV in Tables 8 and 9. Tables A1, A3, and A5 present coefficients, standard errors and marginal effects for the plan-choice stage of Models I, II, and III, respectively. Tables A2 and A4 report coefficient estimates, standard errors, and marginal effects for the attention stage, and the acuity equation of Models II and III (Model I does not contain an attention stage nor an acuity equation). In the following, we first describe the results for the plan-choice stage for the different models and then focus on the results for the attention stage and acuity equation. We mainly discuss the coefficient estimates. The marginal effects are discussed in Section 5.4 in the main text.

Plan-choice stage

As Table A1 shows, all coefficients in the plan-choice stage of Model I are highly statistically significant. They furthermore have the expected signs: Individuals are less likely to choose plans with higher OOP, higher premium, higher deductible, and those plans that do not offer gap coverage. An increase in a plan's ICL increases the probability that this plan is chosen. Similar to the results in the literature, the premium coefficient is more than three times the OOP coefficient, suggesting that individuals largely overvalue premiums in their decisions, see also Abaluck and Gruber (2016a). Furthermore, the estimates of the mean for the old plan dummy – or, to be precise, the mean of the random coefficient – is positive and high, translating into an estimate of high switching costs – as displayed in Figure 1. We further find even in our simplest model that an increase in the plan-level variance of OOP significantly decreases the probability that the plan is chosen, which indicates that risk aversion plays a role in Part D plan choice.

Adding an attention stage in Model II results in only relatively minor changes to the coefficient estimates in the plan-choice stage, as the comparison of the values in Tables A1 and A3 shows. The only relevant changes appear for the mean (and the variance) of the old plan dummy. The mean is reduced from 9.2 in Model I to 6.5 in Model II (the variance from 3.6 to 2.8). Taken together with a relatively stable coefficient for the plan premium, the reduction in the mean implies a reduction in the estimated switching costs, see also Figure 1. As part of inertia is explained by inattention in Model II, this change is expected.

When going from Model II to Model III, the coefficients of all but the old plan dummy hardly change as the comparison across Tables A3 to A5 shows. Only the mean (and the variance) of the old plan dummy are reduced, from 6.5 to 3.9 (2.8 to 1.2 for the variance). Allowing for heterogeneity thus further decreases the estimated switching costs.

Attention stage

Tables A2 and A4 report the coefficient estimates, standard errors and marginal effects for Models II and III, respectively, for the three sets of attention triggers: experience with Part D, changes in plan features across years, and health events/health care use in $t-1$.

In both models, hitting the gap increases attention for individuals who do not have gap coverage; so does hitting the catastrophic coverage region. Furthermore, changes in plan features trigger attention in both models. As in Model IV, changes in premiums are particularly important for attention in both Models II and III. In addition, changes in deductible, ICL, and the benefit design significantly trigger attention in both models. Again, similar to the results presented in Table 8 for Model IV, health events and health care use in $t-1$ seem to play a smaller role for attention. While in Model II, individuals with five or more doctor visits have a higher probability of attention, individuals with at least one visit to the Emergency Room pay attention with slightly lower probability. In Model III none of the coefficients for health events or health care use are significantly different from zero.

Acuity

Results for the determinants of acuity are presented in the bottom panels of Tables A2 and A4 for Models II and III, respectively. As in Model IV, all coefficient estimates except for the years of experience with Part D are negative and highly statistically significant in Models II and III. According to these models, men have lower acuity than women, non-whites have lower acuity than whites, and older individuals have lower acuity than younger ones. Individuals who reside in a ZIP code with a higher share of low educated seniors have lower acuity, and so do individuals in low and very high income ZIP codes. Furthermore, individuals who ever had depression have lower acuity.

Interestingly, the coefficient estimate for experience with Part D is positive in Models II and III, but not significantly different from zero in Model II and only marginally so in Model III. Whether heterogeneity is allowed for in the attention and the plan-choice stage thus seems important for conclusions on how acuity evolves over time in Part D. More specifically, neglecting that unobserved heterogeneity in the willingness and ability to engage in Part D plan choice can play a role for consumers' valuation of plan

characteristics would lead to the conclusion that consumers' acuity, attention to plan choice and probability to switch increase over time. In contrast, when heterogeneity is accounted for, a longer experience with Part D is associated with a decrease in acuity.

A.2 Policy simulations comparing Models I–IV

Figure A1 compares the reduction in predicted overspending induced by forced attention for Models II–IV (there is no attention stage in Model I). Compared to the effect we obtain in our preferred Model IV, Model II and III yield interesting differences. In particular, the distribution of the reduction of overspending is more concentrated in Model II and the fraction of individuals for whom overspending increases is much smaller in both models. These differences are due to the fact that Model II and III are more restrictive w.r.t heterogeneity in acuity. This highlights again the importance of allowing for heterogeneity in structural models of individual choice. This is particularly relevant if the policymaker is interested in the trade-off between 'winners' and 'losers' of an intervention that forces individuals to attend to a choice. (This insight most likely holds irrespective of any additional assumptions that would be required to make welfare calculations.)

Figure A2 compares the reduction in predicted overspending by eliminating switching costs for Models I-IV. Again, the distribution of the effects of this intervention depends on the model specification. Interestingly, the effects are most dispersed in Model I, which contains no attention stage at all and no heterogeneity in acuity. In that model removing switching costs induces switching rates to increase the most. As Figure 3 shows, eliminating switching costs results in 80% switching rate in Model I compared to only 38% in Model II, 20.5% in Model III, and 19.5% in Model IV. Due to the lack of an attention stage and acuity in Model I, the random choice error is exacerbated so that overspending increases even more for those individuals who end up in a worse plan. In contrast, in Models III and IV the attention stage is more important than switching costs (as discussed above) so that removing switching costs has a very small effect.

Appendix Tables and Figures

Table A1: Results Model I - plan choice

	Coefficient	SE		Marginal effects	
				Choose Plan (unconditional)	**Choose Plan** (cond. on attention)
Choice Stage					
OOP cost [$1000]	-2.783	(0.026)	***	-0.010	-0.010
Annual premium [$1000]	-8.532	(0.052)	***	-0.029	-0.029
Deductible amount [$1000]	-5.340	(0.051)	***	-0.018	-0.018
ICL amount [$1000]	0.783	(0.151)	***	0.003	0.003
No gap coverage [D]	-0.824	(0.031)	***	-0.004	-0.004
Old plan coefficient mean (μ_γ)	9.156	(0.035)	***	0.334	0.334
Old plan coefficient variance (σ_γ^2)	3.574	(0.025)	***		
Variance of OOP cost [($1000)^2]	-0.279	(0.011)	***	-0.001	-0.001

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

Notes: Average marginal effects on the probability that plan is chosen, unconditional and conditional on attention. Stars indicate whether coefficients are significantly different from zero. The model also includes five dummy variables for the largest plans.

Table A2: Results Model II - attention and acuity

	Coefficient	SE		Marginal effects Attention	Switch
Attention Stage					
Constant	0.173	(0.102)			
Gap coverage [D]	-1.776	(0.040)	***	-0.330	-0.082
Hit the gap [D]	0.542	(0.044)	***	0.117	0.026
Gap coverage & hit the gap [D]	-0.193	(0.060)	**	-0.041	-0.009
Hit catastrophic region [D]	0.172	(0.072)	*	0.037	0.008
Plan consolidated [D]	0.362	(0.044)	***	0.078	0.018
Change of premium [$1000]	4.776	(0.108)	***	1.028	0.233
Change of deductible [$1000]	2.498	(0.273)	***	0.538	0.122
Change of ICL [$1000]	1.397	(0.353)	***	0.301	0.068
OOP effect of formulary change [$1000]	0.512	(0.070)	***	0.110	0.025
Tiers with increases in copayments [share]	-0.057	(0.042)		-0.012	-0.003
Switch between copay and coinsurance [D]	0.800	(0.044)	***	0.171	0.039
Onset of costly condition [D]	-0.038	(0.039)		-0.008	-0.002
Onset of cheap condition [D]	-0.007	(0.028)		-0.001	0.000
Five or more doctor visits [D]	0.085	(0.031)	**	0.018	0.004
At least one ER visit [D]	-0.063	(0.032)	*	-0.014	-0.003
At least one hospital stay [D]	-0.047	(0.036)		-0.010	-0.002
Determinants of Acuity:					
Male [D]	-0.097	(0.026)	***	-0.021	-0.005
Non-white [D]	-0.766	(0.048)	***	-0.158	-0.036
Age 70–79 years [D]	-0.187	(0.034)	***	-0.040	-0.009
Age 80 years or older [D]	-0.408	(0.039)	***	-0.087	-0.020
PDP experience [years]	0.012	(0.027)		0.003	0.001
Low education [ZIP-code share]	-0.639	(0.138)	***	-0.137	-0.031
Low income [ZIP-code share]	-0.796	(0.189)	***	-0.171	-0.039
High income [ZIP-code share]	-1.296	(0.102)	***	-0.279	-0.063
Ever had depression [D]	-0.081	(0.029)	**	-0.017	-0.004

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

Notes: Average marginal effects on the probability to pay attention and the probability to switch plans. Stars indicate whether coefficients are significantly different from zero. The model also includes year dummies.

Table A3: Results Model II – plan choice

| | Coefficient | SE | | Marginal effects | |
				Choose Plan (unconditional)	Choose Plan (cond. on attention)
Choice Stage					
OOP cost [$1000]	-3.021	(0.029)	***	-0.009	-0.020
Annual premium [$1000]	-9.382	(0.058)	***	-0.029	-0.061
Deductible amount [$1000]	-5.955	(0.059)	***	-0.019	-0.039
ICL amount [$1000]	1.017	(0.165)	***	0.003	0.007
No gap coverage [D]	-0.572	(0.035)	***	-0.002	-0.004
Old plan coefficient mean (μ_γ)	6.469	(0.041)	***	0.660	0.282
Old plan coefficient variance (σ_γ^2)	2.799	(0.031)	***		
Variance of OOP cost [($1000)^2]	-0.277	(0.012)	***	-0.001	-0.002

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

Notes: Average marginal effects on the probability that plan is chosen, unconditional and conditional on attention. Stars indicate whether coefficients are significantly different from zero. The model also includes five dummy variables for the largest plans.

Table A4: Results Model III - attention and acuity

	Coefficient	SE		Marginal effects Attention	Switch
Attention Stage					
Constant	-1.794	(0.115)	***		
Gap coverage [D]	-1.399	(0.054)	***	-0.120	-0.060
Hit the gap [D]	0.468	(0.042)	***	0.048	0.023
Gap coverage & hit the gap [D]	-0.166	(0.069)	*	-0.016	-0.008
Hit catastrophic region [D]	0.176	(0.086)	*	0.018	0.008
Plan consolidated [D]	0.144	(0.047)	**	0.014	0.007
Change of premium [$1000]	6.927	(0.125)	***	0.685	0.326
Change of deductible [$1000]	2.141	(0.241)	***	0.212	0.101
Change of ICL [$1000]	0.965	(0.352)	**	0.095	0.045
OOP effect of formulary change [$1000]	0.642	(0.071)	***	0.063	0.030
Tiers with increases in copayments [share]	-0.225	(0.045)	***	-0.022	-0.011
Switch between copay and coinsurance [D]	0.744	(0.045)	***	0.080	0.038
Onset of costly condition [D]	-0.052	(0.043)		-0.005	-0.002
Onset of cheap condition [D]	-0.001	(0.030)		0.000	0.000
Five or more doctor visits [D]	0.052	(0.034)		0.005	0.002
At least one ER visit [D]	-0.065	(0.035)		-0.006	-0.003
At least one hospital stay [D]	-0.073	(0.039)		-0.007	-0.003
Acuity coefficient (δ_a)	2.075	(0.040)	***		
Determinants of Acuity:					
Male [D]	-0.058	(0.015)	***	-0.012	-0.006
Non-white [D]	-0.462	(0.031)	***	-0.085	-0.040
Age 70–79 years [D]	-0.084	(0.018)	***	-0.017	-0.008
Age 80 years or older [D]	-0.226	(0.022)	***	-0.045	-0.022
PDP experience [years]	0.030	(0.014)	*	0.006	0.003
Low education [ZIP-code share]	-0.482	(0.080)	***	-0.099	-0.047
Low income [ZIP-code share]	-0.381	(0.113)	***	-0.078	-0.037
High income [ZIP-code share]	-0.687	(0.059)	***	-0.141	-0.067
Ever had depression [D]	-0.060	(0.017)	***	-0.012	-0.006
Unobserved Heterogeneity	1	(by normalization)			

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

Notes: Average marginal effects on the probability to pay attention and the probability to switch plans. Stars indicate whether coefficients are significantly different from zero. The model also includes year dummies.

Table A5: Results Model III – plan choice

	Coefficient	SE		Marginal effects	
				Choose Plan (unconditional)	**Choose Plan** (cond. on attention)
Choice Stage					
OOP cost [\$1000]	-3.096	(0.028)	***	-0.010	-0.037
Annual premium [\$1000]	-9.556	(0.059)	***	-0.030	-0.115
Deductible amount [\$1000]	-5.710	(0.059)	***	-0.018	-0.069
ICL amount [\$1000]	0.882	(0.165)	***	0.003	0.011
No gap coverage [D]	-0.670	(0.035)	***	-0.003	-0.010
Old plan coefficient mean (μ_γ)	3.904	(0.057)	***	0.801	0.192
Old plan coefficient variance (σ^2_γ)	1.320	(0.056)	***	-0.001	-0.003
Variance of OOP cost [(\$1000)^2]	-0.268	(0.012)	***		

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

Notes: Average marginal effects on the probability that plan is chosen, unconditional and conditional on attention. Stars indicate whether coefficients are significantly different from zero. The model also includes five dummy variables for the largest plans.

Figure A1: Reduction in overspending induced by forced attention – model comparison
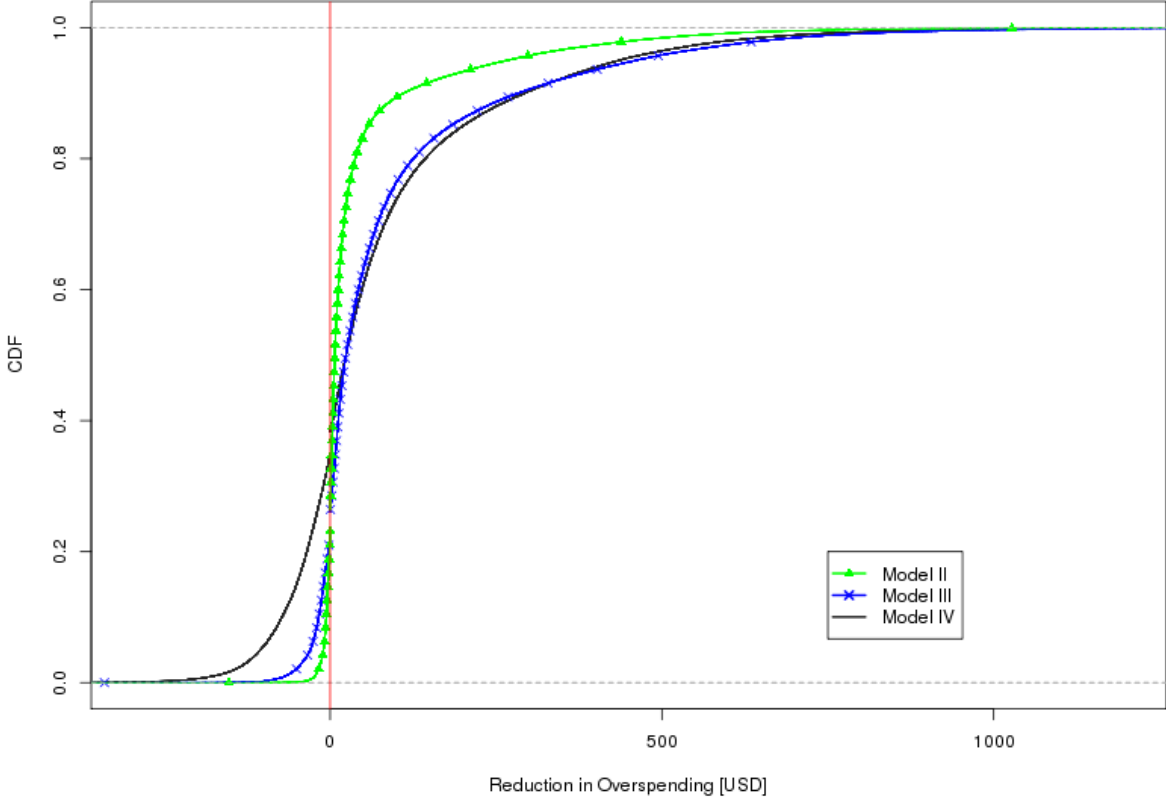
Notes: Empirical CDFs of predicted reduction in overspending (overspending at baseline – overspending after policy intervention). Results shown for Models II – IV (no attention stage in Model I). Policy simulation: Forced attention setting the probability of attention to 1 for everyone.

Figure A2: Reduction in overspending induced by no switching costs – model comparison
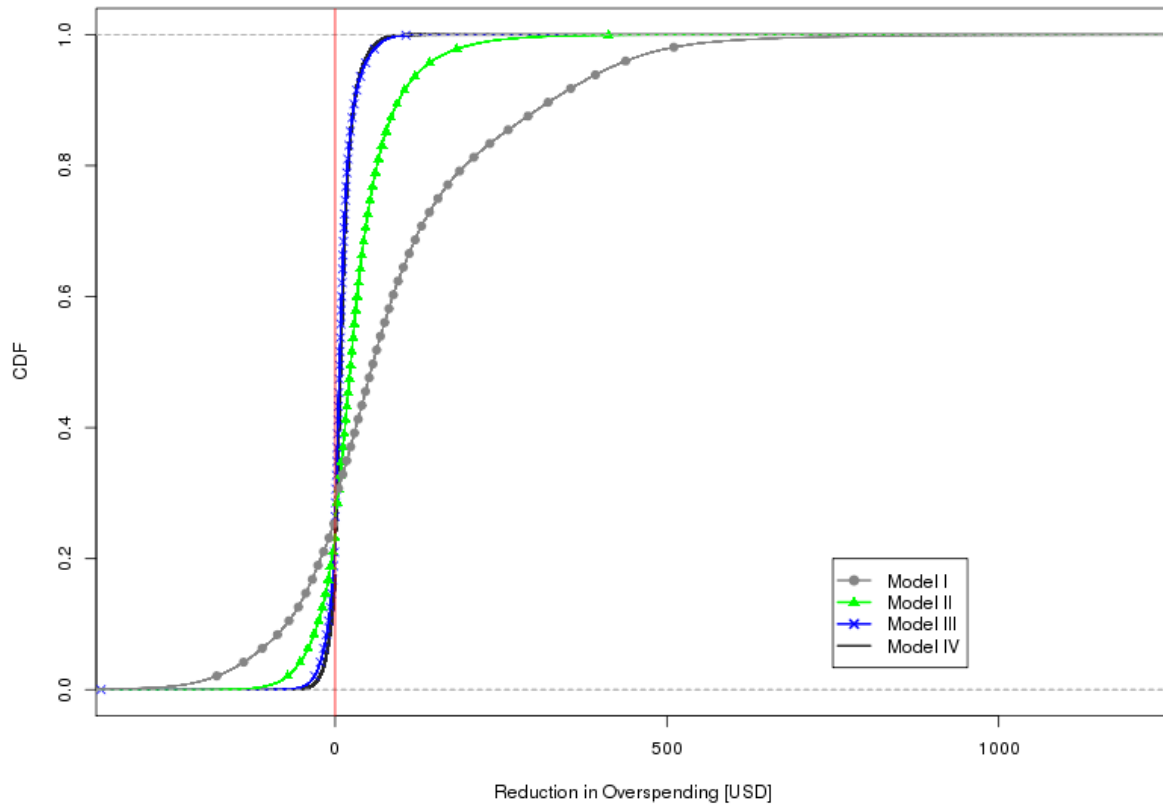
Notes: Empirical CDFs of predicted reduction in overspending (overspending at baseline – overspending after policy intervention). Results shown for Models I – IV . Policy simulation: No switching costs reduces switching costs to zero.