Online Appendix

"The Response of Drug Expenditure to Non-Linear Contract Design: Evidence from Medicare Part D" by Einav, Finkelstein, and Schrimpf

A. Spending around the deductible

The same standard economic theory that generates bunching at the (convex) kink as individuals enter the gap, should also generate "missing mass" at the concave kink created by the sharp price decreases when individuals hit the deductible amount or hit the catastrophic coverage limit (see Figure I). It is difficult to analyze the distribution of spending around the catastrophic limit.[1] Appendix Figure A3, however, shows no evidence of such missing mass around the deductible level for individuals in plans with the standard deductibles. We exclude from the analysis the roughly 10% of people in plans with a (non-zero) deductible that is not the standard deductible level. As with the location of the kink, the level of the deductible is set differently each year in the standard benefit. It is $265 in 2007, $275 in 2008, and $295 in 2009.

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

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

B. Estimation details

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

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

To calculate $m_s(\varphi)$ we simulate data given a vector of parameters. To do so, we first calculate the value function for each latent type and plan combination as described below. For each observation we then simulate S claim histories. Given a person's chosen plan, age, and other characteristics we simulate a sequence of claims. We first draw the person's type m_{is} from a multinomial distribution with probabilities $\exp(z_i\beta_m) / \left(\sum_{k=1}^{M} \exp(z_i\beta_k)\right)$. Then, starting from the first week of the year ($t = 51$) and going until the final week of the year ($t = 0$), we simulate a claim history.[2]

Cumulative spending begins with $x_{is,51} = 0$. The initial health state, λ_{ist}, is drawn from its type specific stationary distribution. Each week there is an event with probability λ_{ist}. When there is an event, the log potential claim is $\log \theta_{ist} \sim N(\mu_{m_{is}}, \sigma^2_{m_{is}})$. The utility cost of not filling the claim is ω_{ist}, which is equal to θ_{ist} with probability $1 - p_{m_{is}}$ and uniform on $(0, \theta_{ist})$ with probability $p_{m_{is}}$. The claim is filled if

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

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

We repeat this simulation until $t = 0$. We then use the simulated data to calculate the statistics described in Section IV.D. Since the number of observations is large, we use one simulation per observation ($S = 1$).

Minimization Throughout the minimization of the objective function, the underlying random draws are kept constant and only shifted and/or rescaled as the parameters change. Nonetheless, the simulated objective is not continuous with respect to φ due to discrete changes in whether some simulated potential claims are filled or not. The large number of potential sequences of claims makes smoothing the objective function difficult. Instead, we use a minimization algorithm that is robust to poorly behaved objectives, the covariance matrix adaptation evolution strategy (CMA-ES) of Hansen (2006). Like simulated annealing and various genetic algorithms, CMA-ES incorporates randomization, which makes it effective for global minimization. Like quasi-Newton methods, CMA-ES also builds a second order approximation to the objective function, which makes CMA-ES much more efficient than purely random or pattern based minimization algorithms. In comparisons of optimization algorithms, CMA-ES is among the most effective existing algorithms, especially for non-convex non-smooth objective functions (Hansen et al. 2010; Rios and Sahinidis 2013). Andreasen (2010) shows that CMA-ES performs well for maximum likelihood estimation of

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

DSGE models. As discussed by Hansen and Kern (2004), an important parameter for the global convergence of CMA-ES is the population size. We initially set the population size to the default value of 15 (which is proportional to the logarithm of the dimension of the parameters), and then increased it to 100. The computation is primarily CPU bound. The estimation takes roughly four days to run on a server with two Intel Xeon E5-2670 eight-core processors.

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

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

where the subscripts denote plan j and type m. The expectation is subscripted by m to emphasize that it depends on the type-specific distribution of θ, ω, and λ'. Given that $v_{jm}(x,0,\lambda) = 0$, we can compute an approximation to v_{jm} sequentially. First, we approximate $v_{jm}(x,1,\lambda)$. Then, we use that approximation to compute $v_{jm}(x,2,\lambda)$, and so on. To be more specific, let $\{x_{k,j}\}_{k=1}^K$ be a large set of values of x that cover the support of x. Then, given some approximation to $v_{jm}(x,t-1,\lambda)$, say $\tilde{v}_{jm}(x,t-1,\lambda)$, we compute

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

We then calculate $\tilde{v}_{jm}(x,t,\lambda)$ using linear interpolation between the $\{(x_{k,j}, v_{k,jm\lambda})\}$ values.[3] We allow $x_{k,j}$ to differ for each plan. For each plan, $x_{k,j}$ is set to 20 evenly spaced points between 0 and the deductible amount, 20 evenly spaced points between the deductible amount and the kink location, 20 evenly spaced points in the gap, and only 2 points above the catastrophic limit. Thus, plans with a deductible use $K = 62$ interpolation points and plans without a deductible use $K = 42$ interpolation points. Above the catastrophic limit, $c(\theta,x) = C\theta$ for some constant C, so the value function is constant and two interpolation points suffice.

To calculate $v_{k,jm\lambda}$, we must integrate over θ, ω, and λ'. λ' is discrete, so integrating over its distribution is straightforward. For θ and ω, we must compute

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

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

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

closed form. In particular,

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

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

where

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

and

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

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

C. More details about model extensions

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

C.1. Allowing for risk aversion

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

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

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

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

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

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

Let

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

The Bellman equation for v is then

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

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

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

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

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

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

To calculate the value function we must compute,

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

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

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

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

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

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

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

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

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

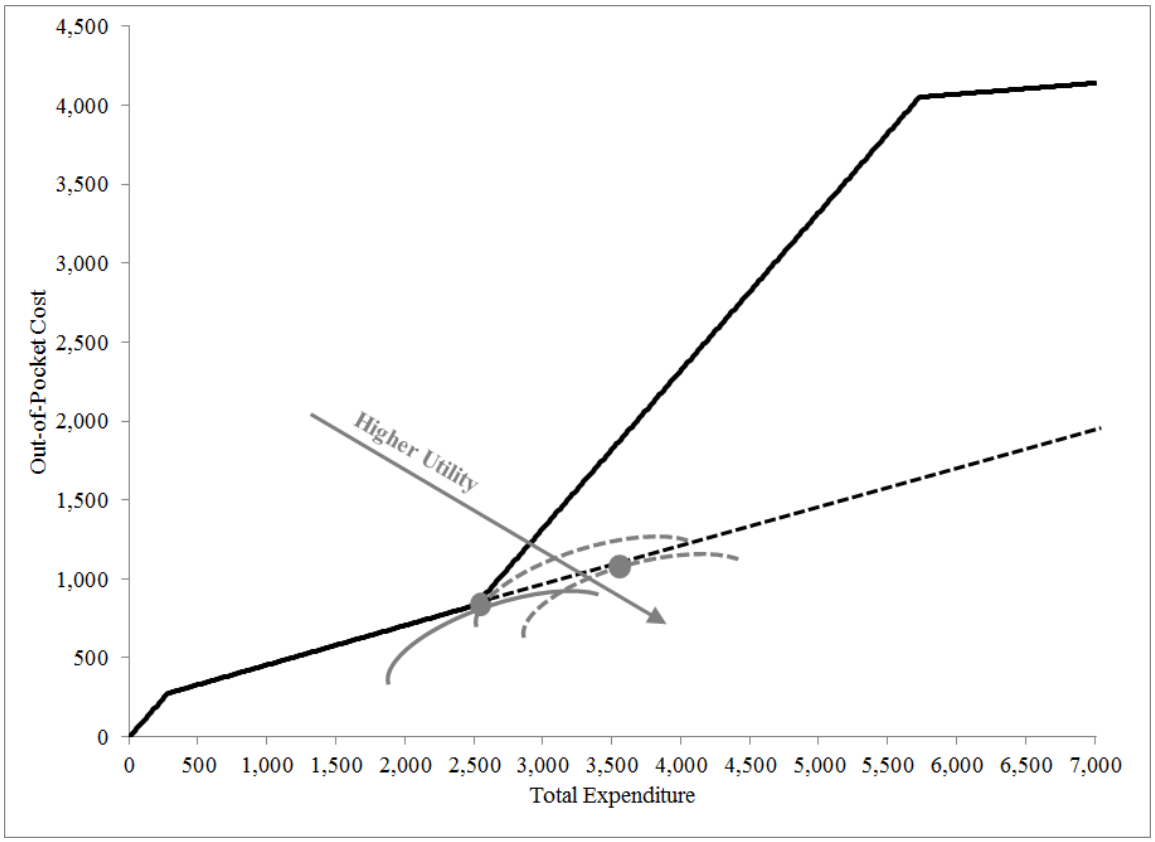
Other cases are treated similarly.

Additional references mentioned only in the appendix

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

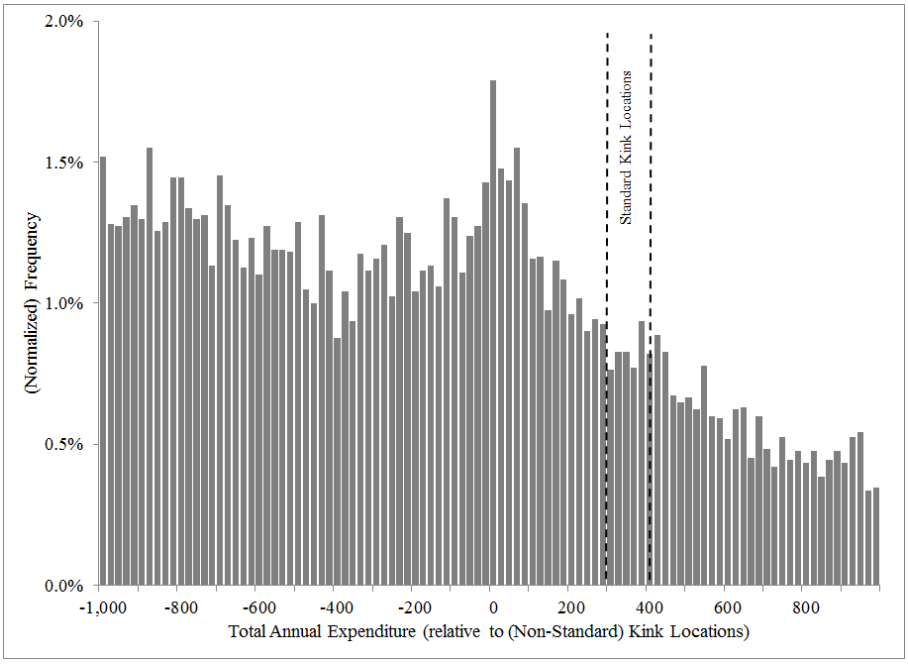
Lo Sasso, Anthony, Lorens Helmchen and Robert Kaestner. 2010. "The Effects of Consumer Directed Health Plans on Health Care Spending." *Journal of Risk and Insurance* 77(1): 85-103.

Figure A1: A Graphical Illustration for The Rationale to Observe Bunching at The Kink



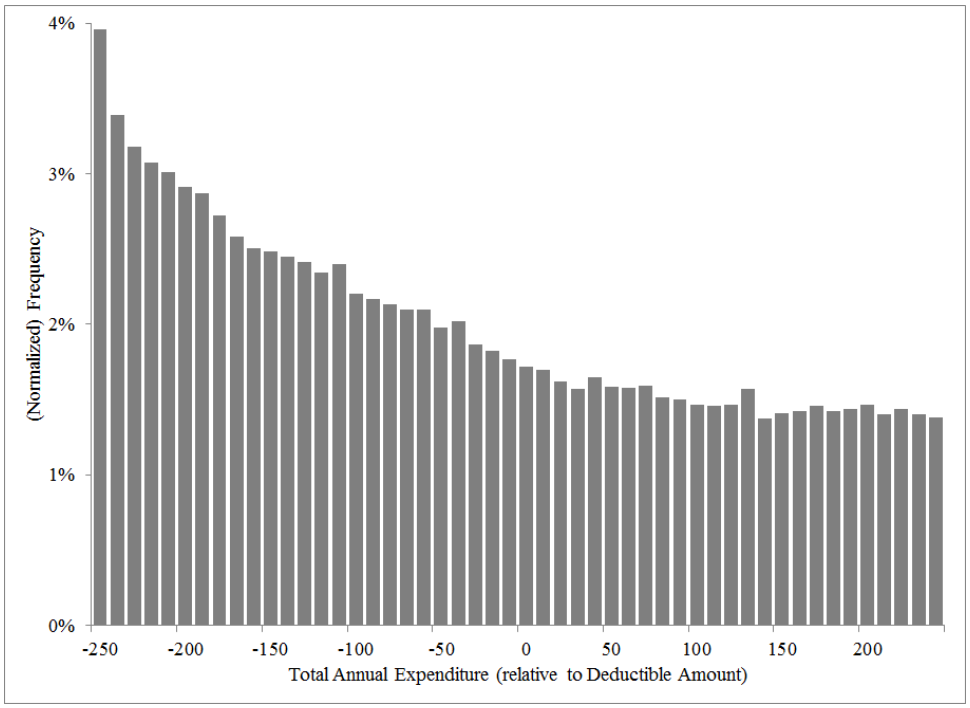
This figure illustrates graphically the theoretical prediction that individuals will bunch at the convex kink point in their budget set. The solid line illustrates the budget set of the same standard benefit design as in Figure I; the standard budget set has a kink (price increase) at $2,510 in total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

Figure A2: Distribution of Annual Drug Expenditure for Individuals with Non-Standard Kink Location



Our baseline sample consists of individuals with a standard kink location. A small sample of individuals excluded from the baseline sample have a kink at an amount that is different from the standard level. The modal non-standard kink amount is $2,100; most of these plans are in 2007 or 2008. The figure displays the histogram of total annual prescription drug spending (in $20 bins) for individuals with the modal ($2,100) non-standard kink location in 2007 or 2008. Such individuals are not in our baseline sample. The x-axis reports total spending relative to the $2,100 kink location. The dashed vertical lines indicate the level of the standard kink locations in 2007 ($2,400) and 2008 ($2,510). Frequencies are normalized to sum to 1 across the displayed range. N =12,189. The figure shows that for individuals in plans with the $2,100 kink location, there is evidence of excess mass around $2,100 but not at the standard kink locations. Naturally, the figure is somewhat noisier than the baseline analyses that use the considerably larger baseline sample.

Figure A3: Distribution of Annual Drug Expenditure Around The Deductible Amount



The figure displays the histogram of total annual prescription drug spending (in $10 bins) for individuals in our baseline sample in plans with the (year-specific) standard deductible amount (which was $265 in 2007, $275 in 2008, and $295 in 2009). The x-axis reports total spending relative to the (year-specific) deductible amount. Frequencies are normalized to sum to 1 across the displayed range. N =186,548.

Figure A4: The Relationship between Excess Mass around The Kink and The Price Change at The Kink

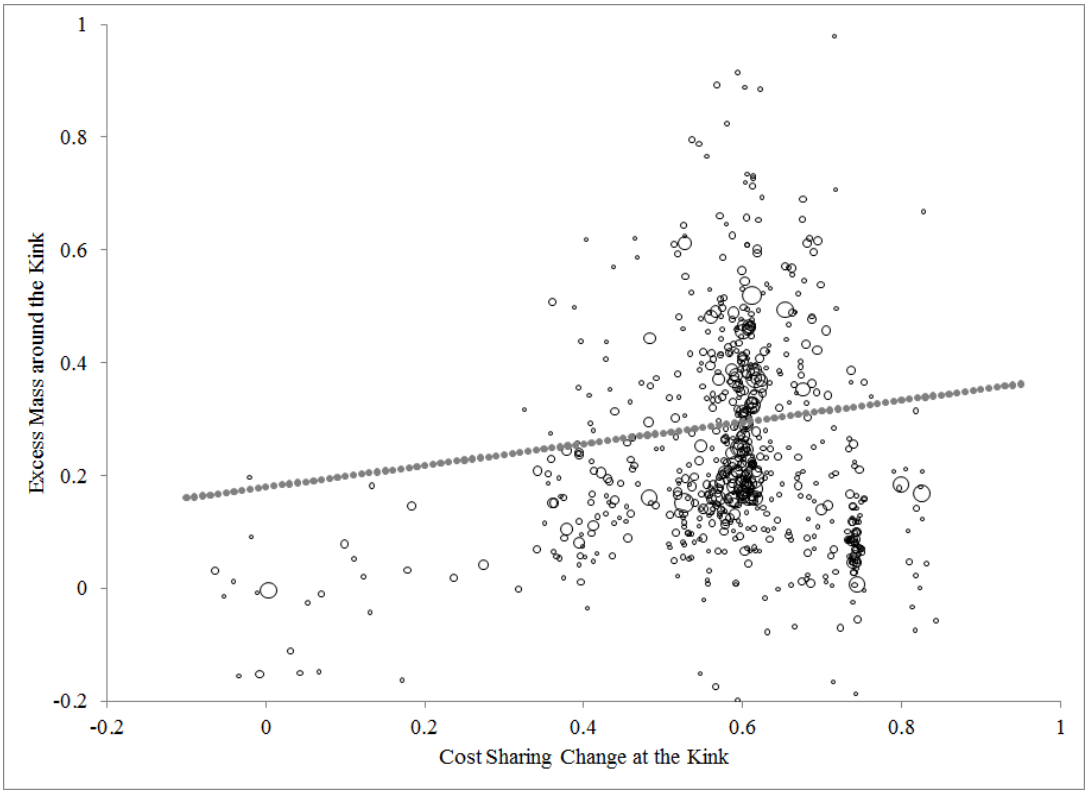


Figure graphs the excess mass in different plans against the size of the kink (i.e. the size of the price increase faced by the consumer as she moves into the gap). The size of the circles is proportional to the number of beneficiaries in the plan. Analysis is limited to the approximately 80% of our baseline sample who are in plans with at least 1,000 beneficiaries within $2,000 of the kink. Excess mass is calculated separately for each plan using the exact same procedure described for Figure IV. The dashed line in the figure represents the enrollee-weighted regression line of the relationship between excess mass and kink size; the slope of the line is 0.19 (standard error = 0.08) and the regression has an R-square of 0.011. N =1,985,697.

Figure A5: An Alternative Measure to Compute Predicted Claim Propensity around The Kink

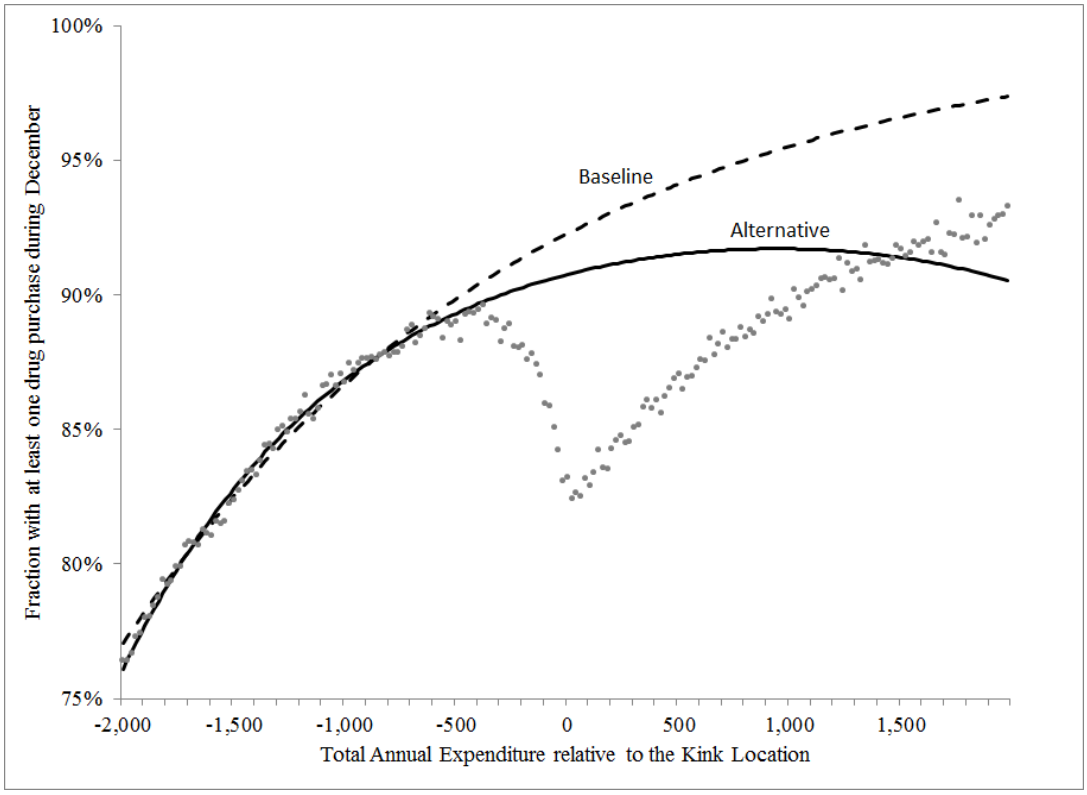


Figure is the same as the December (bottom right) panel of Figure V of the main text, in which the dashed ("baseline") line is generated by regressing the logarithm of the share of individuals with no purchase in December in each $20 spending bin, using only individuals with annual spending (relative to the kink location) between -$2,000 and -$500, on the mid-point of the spending amount in the bin, weighting each bin by the number of beneficiaries in that bin. This prediction forces the predictive line to be monotone in spending, and to asymptote to one as annual spending approach infinity. The solid ("alternative") line shows how the prediction would change if the restriction is not imposed. It is generated by regressing the share of individuals with purchase in December in each $20 spending bin, using only individuals with annual spending (relative to the kink location) between -$2,000 and -$500, on a quadratic function of the mid-point of the spending amount in the bin, weighting each bin by the number of beneficiaries in that bin. Online Appendix Table A1 shows how this alternative prediction affects the quantitative results.

Figure A6: Out-of-Sample Model Predictions for The Distribution of Annual Drug Expenditure

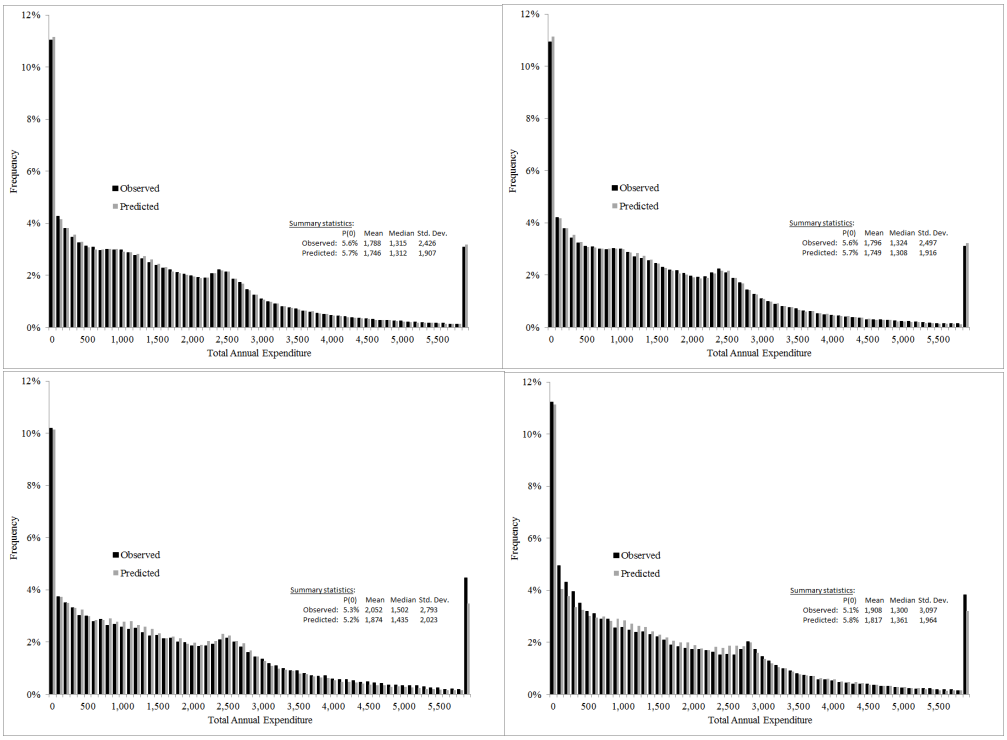


Figure presents the out-of-sample fit of the model. Top left panel replicates the in-sample fit of the model as in the top panel of Figure VI in the main text. Recall (see footnote 23 in the main text) that for estimation we limit the baseline sample to the 500 most common plans, and then only use a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

Figure A7: Out-of-Sample Model Predictions for The Distribution of Annual Drug Expenditure around The Kink

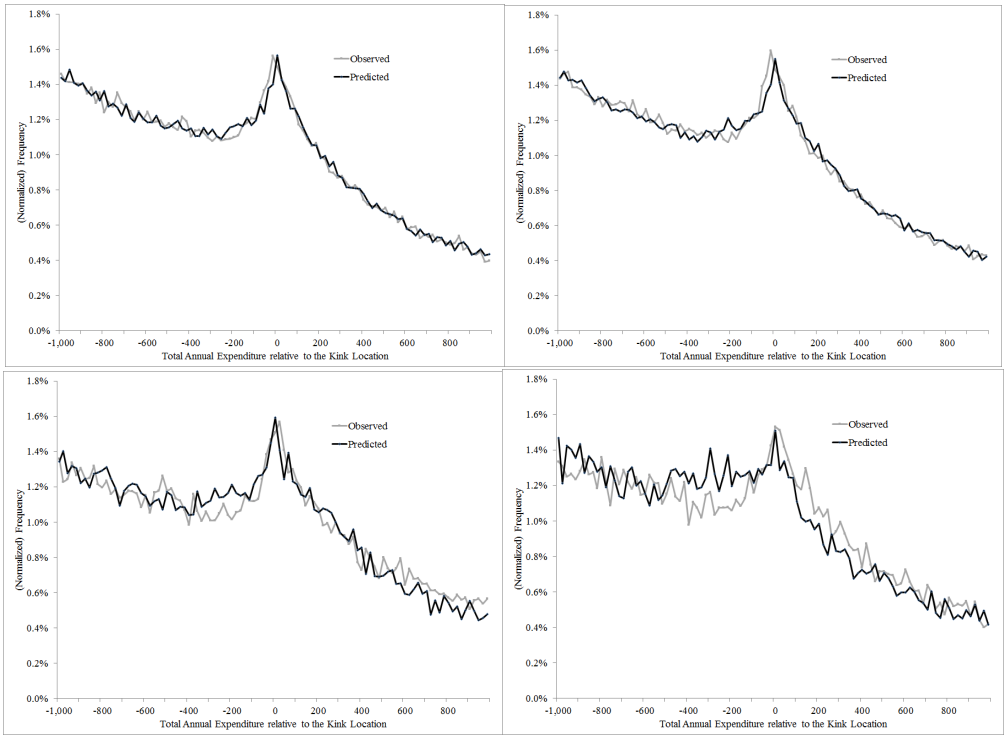


Figure presents the out-of-sample fit of the model. Top left panel replicates the in-sample fit of the model as in the bottom panel of Figure VI in the main text. Recall (see footnote 23 in the main text) that for estimation we limit the baseline sample to the 500 most common plans, and then only use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

Figure A8: Out-of-Sample Model Predictions for The Monthly Propensity to Claim

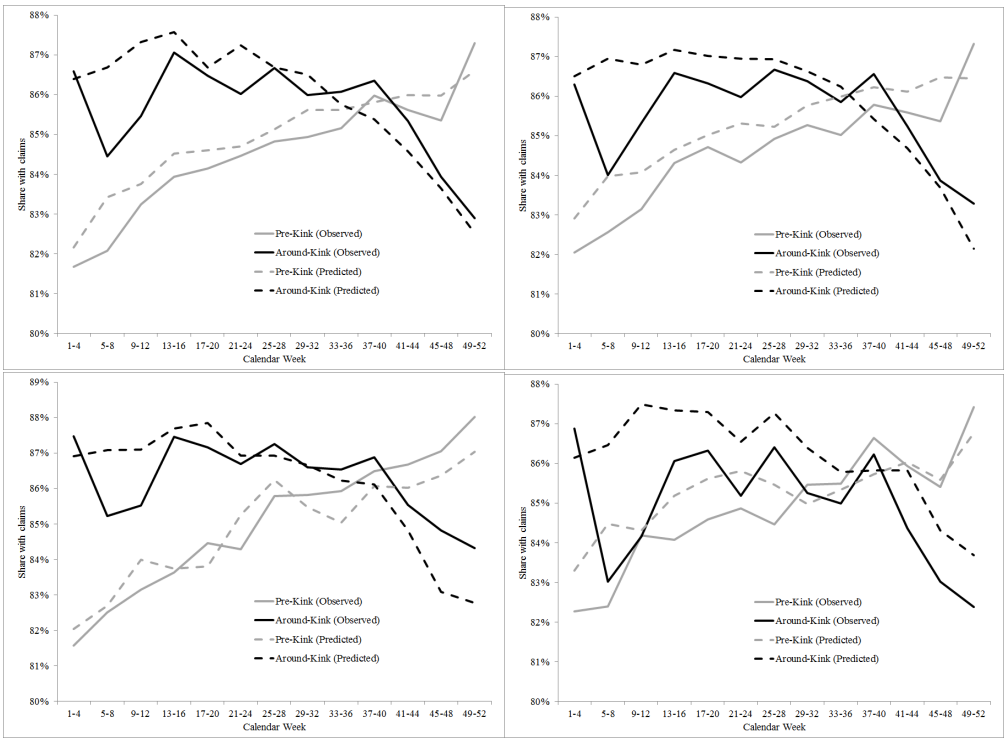


Figure presents the out-of-sample fit of the model. Top left panel replicates the in-sample fit of the model as in Figure VII of the main text. Recall (see footnote 23 in the main text) that for estimation we limit the baseline sample to the 500 most common plans, and then only use a 10% random subsample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

Figure A9: Relative January Drug Expenditure for Deductible and No-Deductible Plans

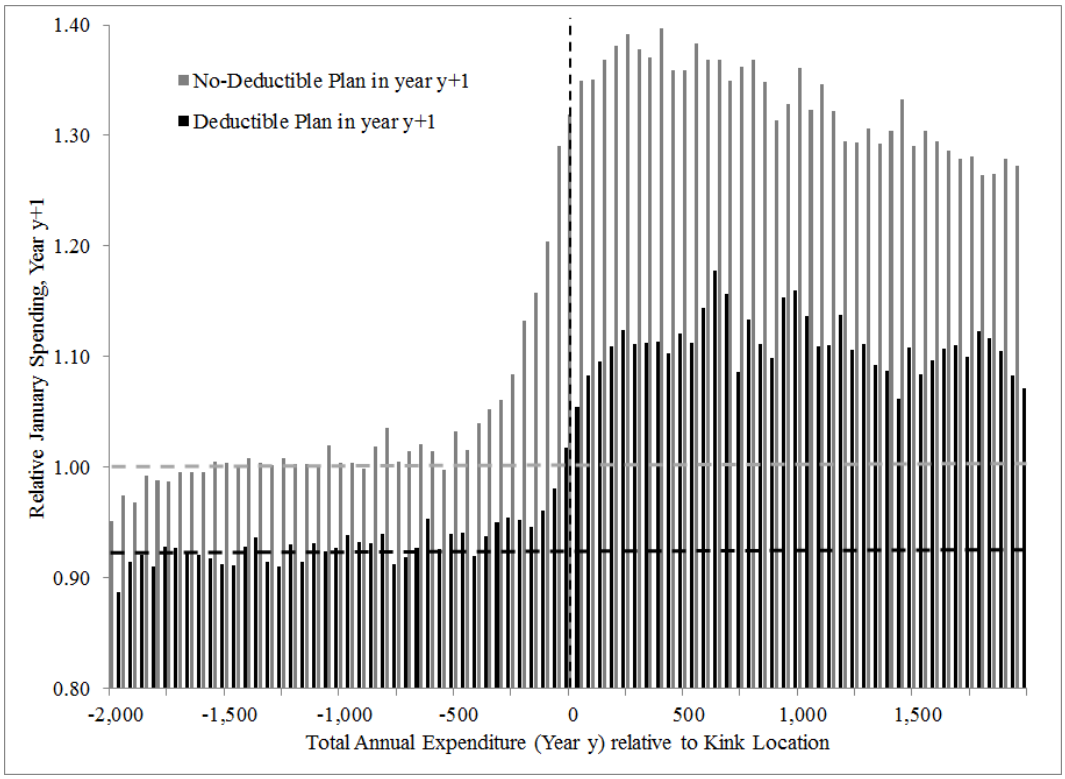


Figure replicates the top panel of Figure IX, but does it separately for beneficiaries who are enrolled in a deductible plan (black bars, $N = 305,437$) and no-deductible plan (gray bars, $N = 1,148,102$) in year $y + 1$. Figure shows the individual's relative January spending in year $y + 1$ as a function of her total annual spending (relative to the kink location, which is normalized to 0) in the prior year (year y). "Relative" January spending in year $y + 1$ is defined as the ratio of January spending in year $y + 1$ to average monthly spending in March through June (of year $y + 1$). Each bar on the graph represents individuals within \$50 above the value on the x-axis. The y-axis reports the average, for each year t spending bin, of the "relative January spending" measure. The dashed, horizontal "counterfactual" relative January spending is calculated as the average relative January spending for people -500 to -2000 below the kink in year y.

Table A1: Using an Alternative Measure to Quantify The Heterogeneity across Types of Drugs in The Response to The Kink

Drug Type	Percent of purchases (1)	Percent of spending ($) (2)	Actual P(Dec. Purchase) (3)	Predicted P(Dec. Purchase) (4)	Percent decrease in purchase probability (5)	Excess January spending (6)
All	100.0	100.0	0.846	0.907	0.068 (0.002)	1.333
Chronic	69.6	77.1	0.762	0.841	0.094 (0.003)	1.323
Acute	30.4	22.9	0.532	0.612	0.129 (0.005)	1.335
Maintenance	85.1	90.1	0.807	0.879	0.082 (0.002)	1.350
Non-Maintenance	14.9	9.9	0.303	0.338	0.104 (0.008)	1.151
Brand	33.4	74.8	0.624	0.744	0.161 (0.003)	1.362
Generic	66.6	25.2	0.731	0.819	0.107 (0.003)	1.235
"Inappropriate" [a]	2.7	1.3	0.075	0.091	0.178 (0.018)	1.262

Table replicates Table IV in the main text, but uses a less restrictive way to generate "predicted" probabiltiies. Online Appendix Figure A5 provides more details about the comparison between the primary and alternative way to construct these predictions.

Table A2: Estimated Price Elasticities as Implied by The Estimated Model Parameters

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

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

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

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

Table A3: Parameter Estimates from The Extension of The Model that Allows Cross-Year Substitution

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

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