Dynamic Moral Hazard in Nonlinear Health Insurance Contracts*

Cecilia S. Diaz Campo†

University of Western Ontario, Department of Economics

November 9, 2021

Link to the latest version

Abstract

Standard health insurance contracts generate nonlinear pricing through the presence of deductibles and caps on out-of-pocket spending, in which the out-of-pocket price paid by consumers decreases as the cumulative use of health care increases. This nonlinear benefit structure, coupled with the uncertainty intrinsic to future health care demand, provides dynamic incentives for consumers' choices: health care utilization today reduces future expected prices. Standard analyses of insurance contracts study the trade-off between the welfare gains from risk protection and the welfare losses from moral hazard, which I relabel as static moral hazard. In this paper, I study a new source of moral hazard, *dynamic* moral hazard, which I define as the additional health care utilization when individuals internalize that current utilization lowers future expected prices via the nonlinearities of the contract. By leveraging the random assignment of families to health plans from the RAND Health Insurance Experiment, I am able to focus specifically on moral hazard, avoiding the typically confounding adverse selection present in insurance markets. I develop and estimate a dynamic, stochastic model of weekly health care utilization at the family level that incorporates the dynamic pricing effects. My estimation framework allows for flexibly-correlated multidimensional unobserved heterogeneity related to family health risk, preferences for visiting a doctor, and price sensitivity. I document that 40 percent of total moral hazard is attributed to dynamic moral hazard. Using my estimated model, I study the welfare implications of dynamic moral hazard in the setting of employer-sponsored health insurance. My results show that the presence of dynamic moral hazard can severely dampen the welfare gains associated with higher cost-sharing and plays a crucial role, distinct from static moral hazard, in determining optimal insurance contract design.

*I am greatly indebted to my advisor David Rivers and committee members Salvador Navarro and Tim Conley for their support and guidance. I want to thank Audra Bowlus, Kenneth Judd, Paul Schrimpf, and Corina Mommaerts for valuable comments and suggestions. All errors are my own.

†Email: cdiazcam@uwo.ca; Website: https://sites.google.com/view/ceciliadiazcampo

1 Introduction

Typical health insurance contracts increasingly include sizable annual deductibles and caps on consumers' out-of-pocket expenditures. In the context of employer-sponsored health insurance in the United States, 83 percent of covered workers had a deductible in 2020 and all had a plan with a cap on out-of-pocket expenditures.[1] The presence of deductibles and caps give rise to nonlinear health insurance contracts, where the out-of-pocket price decreases as the cumulative use of health care (over the covered year) increases. In a typical nonlinear contract, families pay the full price of care below the deductible. After the deductible is exhausted, families pay only a portion of the bill equal to the coinsurance rate, and, once they reach the cap, they face no cost-sharing and have complete insurance coverage for the remainder of the year. These nonlinear benefit structures, coupled with the uncertainty surrounding future health care demand, create dynamic incentives for consumers because current health care utilization reduces future expected prices.

Standard analyses of insurance contracts study the trade-off between the welfare gains from risk protection and the welfare losses when consumers do not face the full cost of their care (Arrow, 1963; Pauly, 1968). In the health insurance literature, the term "moral hazard", which I relabel as *static moral hazard*, is used to capture the notion that insurance coverage may increase health care use by lowering the out-of-pocket price of care to the individual (Einav and Finkelstein, 2018).[2] In this paper, I study a new source of moral hazard, *dynamic moral hazard*, which I define as the extra health care utilization when individuals internalize that the more they consume today, the closer they move towards the deductible and the cap, and the higher the probability they enjoy lower prices for the remainder of the year. Thus, an additional benefit of health care utilization today is lower future expected out-of-pocket prices.

The presence of dynamic moral hazard has implications for the standard analysis of moral hazard. For example, a standard approach to reducing moral hazard is to increase consumer cost sharing. But, if individuals anticipate a lower future price, they respond to a shadow price lower than the spot price in the range below the deductible. Thus, much of the savings thought-to-be-achieved in this range will not actually be realized. Given the concern about the size and rapid growth of the health care sector, there is considerable academic and public policy interest in a better understanding of moral hazard and the ways to mitigate its impact on social welfare. Dynamic moral hazard is particularly relevant in this context, because nonlinear contracts are widely popular not only in private health insurance but also, increasingly, in public health insurance programs, such as Medicare Part D.

[1] These numbers are up from 70 percent and 82 percent a decade ago, respectively. Source: Employer Health Benefits Survey 2020, Kaiser Family Foundation.

[2] As emphasized by Einav and Finkelstein (2018), the use of the term "moral hazard" is an abuse of the "hidden action" origin of the term. In the health insurance literature, the "action", i.e., the agent's health care utilization decision, is observed and contractible. The asymmetric information problem may be more accurately described as a problem of "hidden information" regarding the agent's health risk.

Recognition of the possibility of dynamic moral hazard highlights potentially important limitations of standard models of health insurance that have traditionally overlooked these dynamic incentives and their implications in health insurance.[3] Most empirical papers study health care utilization decisions through the lens of annual models, which aggregate health care decisions up to the annual level (see e.g., Einav et al. (2013), Kowalski (2015), Ho and Lee (2021), and Marone and Sabety (2021)). In these models, individuals make a one-shot decision under full certainty about the complete sequence of health shocks within the year. If future health care demand could be predicted with certainty, the sequential decision problem would be the same as in the one-period case. When uncertainty is present, any health care utilization in the range below the cap has the additional benefit of reducing the remaining cap and, hence, reducing the expected costs of future health care. Annual models thus abstract, by design, from the uncertainty intrinsic to health care demand and the dynamic incentives induced by the nonlinearities of the contracts, which could have important implications in health insurance design. This limitation is not unique to annual models of health care utilization. More broadly, the dynamic pricing incentives are neglected whenever the frequency of the consumer's decision in the model coincides with the length of the contract or when the individuals' behavior is assumed to respond only to the spot price of care.

In this paper, I begin by providing compelling evidence showing that consumers respond to the dynamic incentives of nonlinear contracts and I explore the nature of this response. I then use my findings to develop a tractable, within-year model of health care demand that incorporates the dynamic pricing effects via the nonlinearities of the contract. I estimate the model primitives using a state-of-the-art technique that allows for multidimensional, flexibly-correlated unobserved heterogeneity in health risks, preferences for doctor visits, and price sensitivity. Finally, I use my model and estimates to study the implications of dynamic moral hazard for the optimal design of nonlinear contracts in the context of employer-sponsored health insurance. To do so, I explore the interplay between four contract features: (1) the deductible size, (2) the coinsurance rate after the deductible, (3) the cap on out-of-pocket expenditures, and (4) the resetting time for deductibles and caps. While the first three features are standard in the literature, the optimal resetting time for deductibles and caps has not been studied before.[4]

I use rich, individual, line-item records from the RAND Health Insurance Experiment (New-

[3] Two exceptions are the early theoretical works of Keeler et al. (1977a) and Ellis (1986). Empirically, Cronin (2019) allows the nonlinearities of the contract to affect the number of monthly health care visits, but not the dollar amount consumed. Similarly, in Einav et al. (2015) individuals decide weekly whether to fill a prescription drug, internalizing the impact on expected future prices. Neither of these papers quantifies dynamic moral hazard and its implications for the design of health insurance contracts.

[4] I recently became aware of the concurrent work of Hong and Mommaerts (2021), which explores the implications of deductibles that reset over 6 months versus 12 months. I also note that the early theoretical work of Keeler et al. (1977b) analyzing individual versus family deductibles has a parallel with the optimal resetting time for deductibles. A separate literature studies the optimal *contract* length (and therefore the optimal frequency of open enrollment), holding the timespan over which deductibles and OOP limits aggregate to be annual (see e.g., Darmouni and Zeltzer (2017) in health care, Ghili et al. (2020) and Atal et al. (2020) in long-term care, and Cabral (2017) in dental care).

house and The Insurance Experiment Group, 1993). The RAND experiment is a large randomized field trial of alternative insurance plans offered to approximately 2,500 families representing the non-elderly U.S. population. The experiment randomly assigned families to one of 14 different fee-for-service insurance plans that varied along two principal dimensions: the coinsurance rate and the annual cap on out-of-pocket expenditures. By leveraging the random assignment of families to plans, I can focus on the problem of moral hazard, avoiding the typically confounding adverse selection present in insurance markets (Akerlof, 1970).[5]

To provide evidence of whether consumers internalize the dynamic incentives induced by typical health insurance contracts, I first derive four testable implications within a linear regression of weekly health care utilization on cumulative utilization (over the coverage year), contract week, and their interaction. Using family fixed effects, I exploit the within-family variation in the shadow price of care that comes from two different sources: the distance to the cap on out-of-pocket expenditures and the number of weeks left in the contract before the price schedule resets. I also leverage the staggered enrollment dates from the experimental design to separate seasonal variation in health care demand from the dynamics of my model. Intuitively, the first two implications capture that the further away from the cap and the closer to the end of the contract, the lower is the likelihood of reaching the cap before the contract resets, which increases the shadow price and discourages current utilization. The last two implications check that none of the first two effects survive once families exceed their cap. Using two measures of health care utilization, I show that the implications hold, suggesting that the observed behavior is consistent with forward-looking families who internalize the dynamic pricing effects by updating their expected future prices over the course of the year.

Informed by this evidence, I build a single-agent, finite-horizon, dynamic, stochastic model of health care utilization at the family level combining elements of the annual model of health care demand from Einav et al. (2013) and the within-year model of internet demand from Nevo et al. (2016). I model families' health care utilization decisions at the weekly level where they respond to the *shadow* (or effective) price of health care, rather than the *spot* price or the realized end-of-year price.[6] Thus, families in my model act as though they face a shadow price lower than the spot price in the range below the cap. In line with my regression estimates, the model implies that the shadow price of care is weakly decreasing in the proportion of the cap consumed and in the number of weeks left before the contract resets. In this way the model accounts for the fact that decisions

[5] Because of their nonlinear cost-sharing features, the RAND plans anticipated the design of modern health insurance plans and still receive much theoretical and empirical attention (see e.g., Lin and Sacks (2019), Aron-Dine et al. (2015), and Vera-Hernández (2003)).

[6] In general, nonlinear cost-sharing features of health insurance contracts imply that the out-of-pocket price of health care declines as total utilization accumulates. Thus, at any point in time, the *shadow* price of a unit of health care is the marginal (or *spot*) price minus the bonus for moving closer to the next kink in the budget set, past which cost-sharing by the individual falls or is even eliminated. There are some exceptions. In the case of Medicare Part D, where the coinsurance rates faced by the patients are not monotonically decreasing as total health care utilization accumulates, the *shadow* price can potentially exceed the *spot* price of care.

are made sequentially throughout the year and information is obtained gradually as health shocks arrive and families move along their nonlinear budget set.

I then estimate my model by adapting the approach proposed by Ackerberg (2009), Bajari et al. (2007), Fox et al. (2011), and Fox et al. (2016); and recently applied by Nevo et al. (2016) in the context of demand for residential broadband and Blundell et al. (2020) in firms' investment decisions in pollution abatement technologies. This approach allows me to incorporate flexibly-correlated unobserved heterogeneity in four dimensions related to family health risk (two dimensions), preferences for visiting the doctor, and price sensitivity, and a fifth partially-observed dimension that captures family income. The existence of multiple dimensions of individuals' private information is well documented in the literature (see e.g., Finkelstein and McGarry (2006) in the long-term care insurance market and Fang et al. (2008) in the Medigap insurance market). Nevertheless, previous literature has been constrained regarding the amount of heterogeneity they could allow for and their correlation structure, mainly due to computational reasons. I overcome these constraints by using the computationally advantageous estimator of Fox et al. (2011).

The estimator recovers the nonparametric distribution of unobserved heterogeneity using inequality constrained least squares on a fixed grid. While applying the method of Fox et al. (2011) helps reducing the computational burden, there is still a curse of dimensionality as the number of dimensions in the grid increases. I determine the grid of family types by adapting the method of good lattice points (*glp*) introduced in economics by Judd (1998) in the context of integration and simulation. Good lattice point sets have better space-filling properties than standard tensor product point grids or random sequences and can produce more accurate approximations. As far as I know, this paper is the first to apply this method to dynamic programming problems. Results from a Monte Carlo exercise suggest that it improves computational efficiency by more than a factor of ten. Fox et al. (2011) approach together with the *glp* method enables me to introduce considerable unobserved heterogeneity in health care demand.

In order to decompose static moral hazard from dynamic moral hazard, I simulate a version of my model in which families are myopic, and thus do not respond to the dynamic price incentives. In this model, families respond only to the current spot price of care. I document that 40 percent of total moral hazard is attributed to the dynamic moral hazard component, whereas the rest is standard moral hazard. Moreover, I find that certain contract features exacerbate the impact of dynamic moral hazard on utilization. This highlights the importance of accounting for the dynamic pricing effect when thinking about the optimal cost-sharing features in health insurance.

To analyze the impact of dynamic moral hazard on welfare, I extend the welfare decomposition of Azevedo and Gottlieb (2017) and Marone and Sabety (2021) and provide a novel decomposition in three terms: the value of risk protection, the social cost of static moral hazard, and the social cost of dynamic moral hazard. Using full insurance as a benchmark, I study the impact of dynamic moral hazard on welfare under alternative contract designs not observed in the data. First, I find that for

low caps dynamic moral hazard is particularly strong. So while welfare losses from higher caps due to risk protection are also increasing with the cap, the presence of dynamic moral hazard implies larger optimal caps than would be predicted otherwise. Second, zero deductibles are optimal for low caps, but high deductibles are welfare-maximizing for high caps. In other words, what matters is the distance between the deductible and the cap. So while a high deductible increases the gains due to static moral hazard, a deductible too close to the cap exacerbates the losses due to dynamic moral hazard. This implies, for example, that pure stop-loss contracts in which the deductible and the cap coincide are never optimal. Finally, I find that longer resetting times increase the probability of hitting the cap at some point during the coverage period, so resetting times shorter than twelve months are welfare-maximizing.

Beyond the work noted above, my paper relates to several strands of literature. First, it adds to the sparse literature that test whether individuals respond to the within-year dynamic incentives induced by nonlinear health insurance contracts.[7] The closest to my paper is Aron-Dine et al. (2015) which exploits quasi-experimental variation due to timing of new hires enrolling in employer-provided health insurance plans. The focus on dynamic incentives relates more generally to empirical tests of forward-looking behavior, which plays a key role in many economic problems. Outside the context of health insurance, two works related to mine are Nevo et al. (2016) who analyze the effect of nonlinear pricing schedules in the context of residential broadband use, and Chevalier and Goolsbee (2009) who investigate whether durable goods consumers are forward looking in their demand for college textbooks.

Second, my paper is one of the very few empirical papers which considers health care utilization decisions for periods shorter than the standard contract length of twelve months. By doing so, I incorporate the uncertainty intrinsic to the nature of health care demand and the dynamic incentives throughout the year induced by the nonlinear pricing of typical health insurance contracts. I also model explicitly how health care utilization decisions change with the number of periods left until the contract resets. Thus, I leverage my structural model to explore contracts with shorter resetting times for deductibles and caps, their impact on welfare, and the unique role played by dynamic moral hazard.

My paper also relates to the literature that studies optimal design of health insurance contracts emphasizing the trade-off between welfare gains from risk protection and welfare losses from moral hazard. My paper is closest in spirit to the work of Kowalski (2015), Ho and Lee (2021), and Marone and Sabety (2021), which propose a coherent and unified framework to evaluate risk protection and moral hazard simultaneously. However, none of these papers study the dynamic moral hazard component and its policy implications. Finally, my paper adds to the methodological literature for estimating demand under rich and flexibly-correlated multidimensional unobserved heterogeneity by adapting the methodology from Fox et al. (2011) and Nevo et al. (2016), combined with the use

[7] See e.g., Aron-Dine et al. (2015), Einav et al. (2015), Keeler and Rolph (1988), and Guo and Zhang (2019).

of good lattice points.

The remainder of the paper proceeds as follows. Section 2 describes the data and Section 3 presents descriptive evidence of responses to the shadow price variation. Section 4 details the dynamic model of weekly health care utilization. In Section 5, I present the econometric specification of my model and describe its estimation and identification. Section 6 presents the main results. Section 7 examines optimal contract design within the setting of employer-sponsored health insurance. The last section concludes.

2 Data and Sample

I use rich, individual, line-item records from the RAND Health Insurance Experiment (hereafter, HIE). The RAND HIE is a randomized field trial of alternative insurance plans offered to approximately 2,500 non-elderly families in the U.S. Each line-item record contains information on the total line-item cost, out-of-pocket expenses, insurance payment, date and place of service, and procedure codes. The data are particularly suitable for the study of moral hazard because insurance plans were randomly assigned to families. This forestalls the possibility that less-healthy people, anticipating large health care expenditures, buy more generous insurance coverage (Akerlof, 1970). Specifically, these data are ideal for studying dynamic moral hazard given the unique cross-randomization design of nonlinear cost-sharing features. In what follows I describe the experimental design and the analysis sample.

2.1 Experimental design and randomization

The RAND HIE is a large social experiment conducted between 1974 and 1982 in four urban and two rural sites, chosen to be broadly representative of the nonelderly U.S. population.[8] Families offered enrollment in the experiment represent a random sample from each site, subject to certain eligibility criteria. The criteria excluded those whose health care delivery systems differed from options available to the general population.[9] At a given site and enrollment date, families were randomly assigned to one of 14 different fee-for-service insurance plans or to a prepaid group practice.[10] In

[8]The sites were: Dayton, Ohio; Seattle, Washington; Fitchbury-Leominster and Franklin County, Massachusetts; and Charleston and Georgetown County, South Carolina.

[9]The experiment excluded people age 62 or over at the time of enrollment since they were or would become eligible for Medicare during the experiment or people under age 62 who were eligible for the Medicare program; those with family incomes greater than $25,000 (in 1973 dollars); those who were institutionalized (jail or long-term hospital); those in the military and their dependents; and veterans with service-connected disabilities.

[10]The RAND HIE assigned families to treatments using the Finite Selection Model (Morris, 1979), which explicitly balanced a subset of observable characteristics across plans. Potential selection bias can be introduced if there is differential refusal or attrition across plans. To reduce refusals, families were given a Participation Incentive (PI) if their experimental plans provided less coverage than their existing health insurance policies. For details about the PI payments see Appendix A of Codebook 203 (Newhouse, 1999). A Completion Bonus was offered to reduce

addition, each family was randomly assigned to either three or five years of participation. Families were enrolled in the experiment as a unit, with only eligible family members participating.[11]

The fee-for-service plans varied along two principal features: (1) the coinsurance rate, which is the fraction of billed charges paid by the participant, and (2) the maximum dollar expenditure (MDE), which is the cap on family out-of-pocket expenditures. The coinsurance rates were set at either 0 (free care), 25, 50, or 95 percent. Except for the free care plan, each plan had a MDE of 5, 10, or 15 percent of family income in the previous year (hereafter, PY).[12] To further limit participants' financial exposure, the MDE was capped at $1,000 in 1973 dollars, unadjusted for inflation.[13] Beyond the MDE, the insurance plan paid all covered health care utilization in full for the remainder of the contract year. Associated with the MDE, it will become useful to introduce the concept of a Total Annual Threshold (hereafter, TAT), defined as the ceiling placed on accumulated health care utilization during the coverage period above which health care is free to the family members. The TAT differs from the MDE because the former includes both the portion paid by the insurance and the portion paid out-of-pocket by the patient, while the latter only includes the patient's portion.

All experimental plans feature a zero-dollar deductible, a coverage length of 12 months, and no premiums. Every plan covered inpatient and outpatient health care, as well as vision, prescription drugs, medical supplies, and mental and dental health.[14] A contract year was defined as the 12-month period following each anniversary of the enrollment date, which was not always January 1. New families were enrolled over several start dates, but all members of a given family shared the same enrollment date, even those added later.[15]

Figure 1 shows an example of a health insurance plan offered by the RAND HIE. The total dollar amount of annual health care utilization is summarized on the horizontal axis, as the sum of both insurer payments and out-of-pocket payments by the beneficiary family. The vertical axis indicates how this particular insurance contract translates total utilization into out-of-pocket spending. The figure illustrates the case of a family assigned to a 25 percent coinsurance rate plan with a MDE

withdrawal from the experiment subsequent to enrollment.

[11] After the enrollment date, families could not incorporate new members into the insurance plan. The only exceptions were newborns and adopted children under one year of age. Families who either lost or acquired members during a given contract year were given a new identifier in the following contract year to reflect their change in composition. Hence, a RAND HIE participant might belong to different families throughout the experiment.

[12] There was a group of mixed plans, in which the coinsurance rate differed between medical services and dental or outpatient psychiatric services. Regarding the MDE, one plan limited the out-of-pocket expenditure to either $150 (individual) or $450 (family). These plans are not analyzed in the present study. Since the MDE is tied to the family's previous year's income, it varied from year to year. Families with zero income in the previous year received *de facto* free care, regardless of the plan.

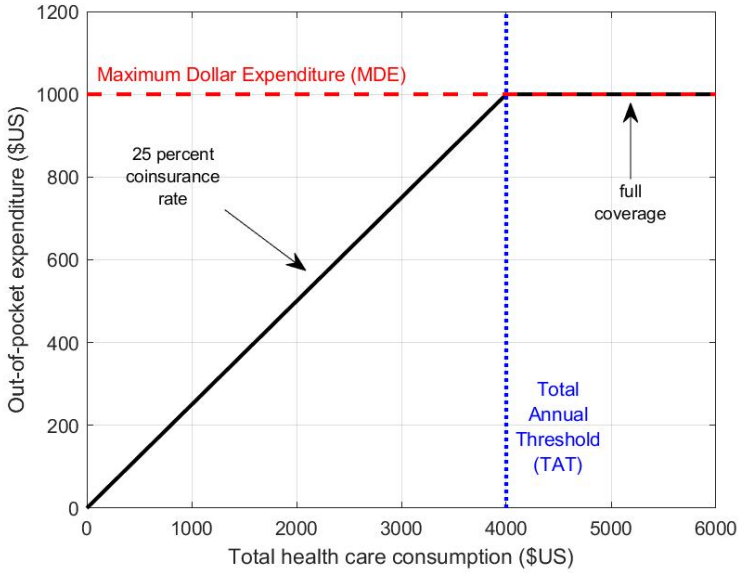
[13] An MDE of $1,000 in 1973 dollars would correspond to about $6,000 in 2020 dollars, based on the U.S. Consumer Price Index (CPI-U). Source: U.S. Bureau of Labor Statistics.

[14] The following services were excluded: non-preventive orthodontic services, cosmetic surgery for pre-existing conditions, and outpatient mental health visits exceeding 52 per contract year. See Appendix D of Codebook 203 (Newhouse, 1999) for a thorough list of possible reasons for noncoverage of a service by the RAND HIE.

[15] Table A1 in the Appendix shows enrollment dates by site.

equals to 10 percent of PY income. Since family income in the previous year is greater than \$10,000, the MDE is capped at \$1,000 (the horizontal red dashed line). In this case, the family would pay 25 percent of the first \$4,000 in health care utilization, and \$0 beyond that for the current contract year. Hence, a 25 percent coinsurance rate coupled with a MDE of \$1,000 has an associated TAT of \$4,000 (the vertical blue dotted line).

Figure 1: Health insurance plan with 25 percent coinsurance rate and MDE=min$\{0.10 \times \text{income}, \$1,000\}$ for a family with PY income greater than \$10,000



Except for the absence of a deductible, Figure 1 shows a stylized example of a typical health insurance contract in the U.S. This example shows a concave, piece-wise linear schedule with two "arms". First, the "coinsurance" arm, where the family faces a price of 25 percent for every dollar of health care utilization, and second, the "catastrophic" arm that provides full coverage. While the plans in my data do not include deductibles, the same forces that govern health care utilization behavior below the cap also apply to the deductibles, and allow me to use my estimated model to study behavior under plans with deductibles (see Section 7).

2.2 Sample

For the purpose of studying family behavior within a contract year, I aggregate health care utilization to the family-week-year level.[16] The variable *date of service* defines whether a line-item claim belongs to one week or another.[17] In addition to the claims data, I use the *eligibility* file to record

[16] A week refers to a contract week, as opposed to a calendar week. The same applies to a year.

[17] Each claim has multiple dates, including the admission and discharge dates in case of hospitalization, the date of service and the date filed.

coverage and family structure, and the *episodes of care* file to recover the family MDE by year. The weekly consumption totals represent only the health care utilization that was covered by the RAND HIE and for which claims were submitted. The weekly covered health care utilization used for the analysis includes both the portion paid out-of-pocket by the family and the portion paid by the insurer. If no family member used covered health care services during a given week-year pair, the utilization value for that family-week-year observation equals zero. Note that observations with zero utilization are kept in the sample.

I make five restrictions to create my baseline sample. First, I exclude fee-for-service plans with different coinsurance rates for different providers (i.e., the so-called mixed plans) and the prepaid group practice.[18] Second, I exclude the first contract year from Dayton, Ohio.[19] Third, I exclude family-year observations in which no member enjoyed coverage for the whole contract year. Fourth, for families reassigned to a different plan after a relocation, I exclude the year of the move as well as the following years. Finally, I exclude family-year observations with missing MDE information.[20] After these exclusions, my baseline sample consists of 2,145 families and 4,763 family-years. Table 1 provides details about the remaining sample size after sequentially applying each exclusion criterion.

Table 1: Analysis sample derivation

Row	Description	Sample Size	Percent
1	Family-years (end-of-experiment point of view)	9,388	
2	and not in mixed-plans or HMO (after exclusion 1)	5,279	56.23% of row 1
3	and not first year in Dayton (after exclusions 1-2)	4,891	92.65% of row 2
4	and full year participation (after exclusions 1-3)	4,790	90.74% of row 2
5	and plan change (after exclusions 1-4)	4,767	90.30% of row 2
6	and complete MDE data (after exclusions 1-5)	4,763	90.23% of row 2
7	Family-years in analysis sample	4,763	

Table 2 provides summary statistics by plan type for the baseline sample. I group the insurance plans into four categories based on the coinsurance rate. In the first panel, I explain the features of the plans. In the second panel, I provide statistics at the annual level as well as a break down by category of expenditure. Finally, in the third panel I present statistics regarding the behavior above the cap. Between 17 and 36 percent of families in the cost-sharing plans hit the TAT in a given year. This is important, as my identification strategy relies on having enough families with a positive probability of exceeding the TAT during the contract year. Families in the less generous plans are more likely to hit the TAT due to the higher cost sharing, but end-of-year prices are

[18] My model does not distinguish between providers of service (e.g., physician versus dentist) or whether the provider belongs to a prepaid group network.

[19] Dental and mental health services were treated differently in the first year of the experiment in Dayton, Ohio. Dental services for adults were covered only on the free-care plan (dental services for children were covered on all plans). Outpatient mental services were not covered.

[20] For details about each exclusion step, see Appendix H.

increasing with the coinsurance rate of the plan. The reported average realized end-of-year price per dollar of health care consumption is the coinsurance rate times the fraction of family-years who do not hit the TAT by the end of the coverage year. It varies between 0 in the free-care plan and 0.61 in the 95 percent coinsurance rate plans.

Table 2: Plan characteristics and descriptive statistics of annual health care utilization

Coinsurance rate	0 percent	25 percent	50 percent	95 percent
Number of family-years	2,393	832	499	1,248
Plan features				
Premium ($)	0.00	0.00	0.00	0.00
Deductible ($)	0.00	0.00	0.00	0.00
Contract length (in years)	1.00	1.00	1.00	1.00
Coinsurance rate below TAT	0.00	0.25	0.50	0.95
Mean Total Annual Threshold ($)	0.00	2557.45	1560.65	744.54
Family total annual consumption				
Percentage with zero claims	4.47	6.13	8.22	12.74
Mean annual consumption ($)	1851.01	1284.92	1303.34	1159.95
Median annual consumption ($)	1022.35	598.37	436.57	329.305
Mean realized end-of-year OOP price	0.00	0.21	0.40	0.61
Mean outpatient share	0.34	0.37	0.40	0.35
Mean inpatient share	0.15	0.13	0.12	0.13
Mean dental share	0.30	0.27	0.235	0.25
Mean drug and supply share	0.14	0.15	0.15	0.13
Mean mental share	0.02	0.02	0.01	0.01
Family total annual consumption above TAT				
Proportion of family-years over TAT	1.00	0.1683	0.2084	0.3566
Mean share of TAT used	1.00	0.9559	1.8100	4.0620
Median share of TAT used	1.00	0.2468	0.2851	0.4917
Family-years over TAT | any inpatient	1.00	0.5150	0.6484	0.9040
Mean consumption above TAT ($)	1851.01	2428.09	3260.28	2219.70
Median consumption above TAT ($)	1022.35	1120.715	1130.375	1246.49

Notes: All expenditures are in dollars and cents for the year of service, unadjusted for inflation.

These differences in plan generosity translate into differences in average annual health care utilization. An average family in the free-care plan consumes $1,851 in health care services during a contract year. At the other extreme, an average family in the 95 percent coinsurance rate plan consumes $1,160 in health care services during a contract year. This implies that average health care utilization in the most generous plan is about 60 percent higher than that in the least generous plan. Table 2 also shows a steady increase of the percentage of family-years with zero annual health care utilization as the coinsurance rate increases. On average, the zero annual utilization rate is

almost 3 times higher in the 95 percent coinsurance rate plans versus the free-care plan.

3 Evidence of Response to the Shadow Price of Care

Under standard nonlinear health insurance plans, an additional benefit of health care utilization today is lower future expected prices within the coverage period (typically one year). This section examines whether families' health care utilization decisions within the year respond to these dynamic pricing incentives. To do so, I derive a set of four testable implications for families who face a nonlinear pricing scheme coupled with uncertainty regarding future health care demand. The implications derived have only one key unobserved mechanism at play: the change in the shadow price of care within the year. Specifically, the higher the cumulative consumption and the more weeks left in the coverage period, the stronger are these incentives, i.e., the lower the shadow price of consumption. However, this is only true while families are below the cap. After they hit the cap they enjoy free care, and these dynamic incentives disappear.

In order to test these implications, I estimate linear regressions of the following form, where the dependent variable, y_{jtq}, is a measure of family j's health care utilization in contract week t of experimental year q:

$$\begin{aligned} y_{jtq} \quad = \quad & NonLastYear_{jq} \sum_{s=\substack{\text{below cap,}\\ \text{above cap}}} \left[\beta_0^s + \beta_1^s \left(T-t+1\right) + \beta_2^s \frac{Cum_{jtq}}{TAT_{jq}} + \beta_3^s \left(T-t+1\right) \frac{Cum_{jtq}}{TAT_{jq}} \right] \\ & + LastYear_{jq}\, \alpha^{\text{above cap}} + Family_j + \boldsymbol{\gamma}\, \mathbf{x}_{jtq} + \epsilon_{jtq}. \end{aligned} \tag{1}$$

The variable $(T-t+1)$ is the number of weeks left in the coverage period and the variable Cum_{jtq} is family j's cumulative health care utilization up to the beginning of week t of experimental year q. The variable $LastYear_{jq}$ is an indicator variable that equals 1 if q is the last year of participation in the experiment for family j and 0 otherwise (the opposite is true for $NonLastYear_{jq}$). As defined in the previous section, the variable TAT_{jq} stands for "Total Annual Threshold" and captures the family- and year-specific level of cumulative health care utilization above which the family enjoys free care for the remainder of the year. I also include family fixed effects, $Family_j$, to remove persistent heterogeneity across families, and dummy variables for calendar month, $\mathbf{x}_{jtq}$. By leveraging the experimental design, I can separate seasonality in health care demand from the dynamics of my model, since families have staggered enrollment dates.[21]

The key coefficients of interest are β_1^s, β_2^s, and β_3^s, which capture the response of weekly health care demand to the variation of the shadow price of care. Using equation (1), I derive four testable

[21] This is a unique feature since standard health insurance coverage begins and ends on the same date for almost all individuals (unless it is terminated due to job separation or death).

implications to evaluate whether families are sensitive to the dynamic incentives generated via the nonlinearities of health insurance plans.[22] The first two implications examine the behavior when families are below the cap on OOP spending, where the spot price is fixed but the shadow price varies. Focusing on the behavior after families exceed the cap, the last two implications address potential threats to identification of the response to the shadow price captured by the first two implications.

$$\frac{\partial\left(y_{jtq} \,\middle|\, Cum_{jtq} < TAT_{jq},\, NonLastYear_{jq}\right)}{\partial(T-t+1)} = \beta_1^{\text{below cap}} + \beta_3^{\text{below cap}} \frac{Cum_{jtq}}{TAT_{jq}} \qquad > 0, \tag{2a}$$

$$\frac{\partial\left(y_{jtq} \,\middle|\, Cum_{jtq} < TAT_{jq},\, NonLastYear_{jq}\right)}{\partial(Cum_{jtq}/TAT_{jq})} = \beta_2^{\text{below cap}} + \beta_3^{\text{below cap}} (T-t+1) \qquad > 0, \tag{2b}$$

$$\frac{\partial\left(y_{jtq} \,\middle|\, Cum_{jtq} \geq TAT_{jq},\, NonLastYear_{jq}\right)}{\partial(T-t+1)} = \beta_1^{\text{above cap}} + \beta_3^{\text{above cap}} \frac{Cum_{jtq}}{TAT_{jq}} \qquad = 0, \tag{2c}$$

$$\frac{\partial\left(y_{jtq} \,\middle|\, Cum_{jtq} \geq TAT_{jq},\, NonLastYear_{jq}\right)}{\partial(Cum_{jtq}/TAT_{jq})} = \beta_2^{\text{above cap}} + \beta_3^{\text{above cap}} (T-t+1) \qquad = 0. \tag{2d}$$

Implication (2a) states that a forward-looking family, whose accumulated health care utilization is held fixed at some level below the cap, should increase its utilization as the number of weeks left in the contract increases. This is because the family has more remaining opportunities to consume within the current contract, and, therefore, a higher probability of hitting the cap on OOP spending and enjoying free care for the remainder of the contract, all else equal. This situation drives the shadow price of care down and, therefore, encourages current utilization. An empirical test for Implication (2a) amounts to a test on the sign of the coefficients $\beta_1^{\text{below cap}}$ and $\beta_3^{\text{below cap}}$. This is because the ratio Cum_{jtq}/TAT_{jq} is always between 0 and 1. Therefore, a sufficient condition for Implication (2a) to hold is that $\beta_1^{\text{below cap}}$ and $\beta_3^{\text{below cap}}$ be jointly positive.

Implication (2b) states that, as the share of the cap consumed moves closer to 1, current health care utilization should increase holding the contract week fixed. The rate at which it increases depends on the number of weeks left in the contract. This response is driven by a decrease in the shadow price of care by taking into account the benefits of moving closer to the cap and potentially enjoying free care for the remainder of the contract. An empirical test for Implication (2b) accords naturally to a test on the sign of the coefficients $\beta_2^{\text{below cap}}$ and $\beta_3^{\text{below cap}}$. Notice that the variable $(T-t+1)$ varies between 1 and 52. Therefore, a sufficient condition for Implication (2b) to hold is that $\beta_2^{\text{below cap}}$ and $\beta_3^{\text{below cap}}$ be jointly positive.

Up to this point, the key identifying assumption that allows me to interpret the estimates of Implication (2a) and (2b) as evidence of dynamic response to the shadow price of care is that

[22] The implications are tailored to the specific cost-sharing features and nature of the RAND HIE, in that they assume zero-dollar deductibles and exclude the last experimental year to avoid the end-of-experiment effect. They can easily be adapted to more general nonlinear contracts and insurance settings.

there are no confounding effects on health care demand as the family approaches the end of the coverage horizon and the cap on OOP spending, conditional on being below the cap. In other words, any differential patterns of weekly health care demand that I observe across contracts weeks and proportion of the cap consumed are caused by differences in the shadow price. However, this identifying assumption might not be correct if health care shocks are correlated across time (even after controlling for persistent heterogeneity at the family level) or if families increase health care utilization toward the end of the experiment in anticipation of being enrolled in a less generous plan after the experiment ends. The first scenario will bias my estimate of Implication (2b) away from 0, while the second scenario will bias my estimate of Implication (2a) toward 0. To investigate the validity of the identifying assumption, next I discuss the last two implications using families' health care utilization behavior once they exceed the cap on OOP spending.

Conditional on being above the cap, having more remaining opportunities to consume within the contract year should not impact current behavior. In other words, there are no further dynamics coming from the nonlinearities of the contract once above the cap, so weekly utilization should remain constant. An empirical test for Implication (2c) amounts to a joint test of the coefficients $\beta_1^{\text{above cap}} = 0$ and $\beta_3^{\text{above cap}} = 0$. This could fail in the presence of an *end-of-experiment* effect, for example, which I document in Appendix B, and constitutes the reason why I condition on non-last experimental years.[23] Implication (2c) could also fail if seasonality in health care demand cannot be separated from the dynamics via the nonlinear pricing (e.g., seasonal flu). To overcome this, I exploit the experimental design and use the variation in families' enrollment month to isolate seasonal demand.

After hitting the cap, families enjoy free care for the remainder of the contract year. Since current utilization does not affect within-the-year future prices, an increase in the share of the cap consumed should not impact current behavior. An empirical test for Implication (2d) amounts to a joint test of the coefficients $\beta_2^{\text{above cap}} = 0$ and $\beta_3^{\text{above cap}} = 0$. This could fail in the presence of correlated health shocks across time, even after removing persistent heterogeneity.[24] In order to capture this potential week-to-week correlation, I include family fixed effects in equation (1). The correlation I fail to capture with family fixed effects will show up in Implication (2d), and this is why this implication is central to my identification strategy. Since there is no shadow price variation left to exploit after exceeding the cap, any effect captured by $\beta_2^{\text{above cap}}$ points to week-to-week correlation in health shocks.

Table 3 displays the results of equation (1) using two measures of health care demand: an

[23] Families participating in the RAND HIE knew that the experiment would end either after 3 or 5 years (randomly assigned before enrollment). This might induce an increase in health care utilization toward the end of the experiment if families anticipate being enrolled in less generous coverage plans after the experiment ends.

[24] Consider the situation of a family member who undergoes a hip replacement surgery at some point during the year. This situation most probably creates demand for further care (e.g., follow-up visits or further tests), which could generate positive serial correlation. This would contaminate the estimate of $\beta_2^{\text{below cap}}$ because I would attribute to a shadow price response a merely week-to-week correlation in health shocks.

indicator for whether the family had any claim over the week and the level of health care utilization (in dollars) over the week. The first three rows show the estimates related with behavior below the cap on OOP spending, and the subsequent three rows show their counterpart once families exceed the cap. As predicted, families increased their weekly health care utilization level the further they were from the end of the contract year and the closer they were to the cap on OOP spending, conditional on being below the cap. This is consistent with families facing uncertainty about future health care demand, updating their probability of exceeding the cap over the course of the contract year, and having a positive discount factor. Once families exceed the cap, I find no statistically significant relationship between health care demand and weeks left or share of the cap consumed. I interpret this as supportive of the identifying assumption. Table 3 also confirms that the four testable implications hold as shown by the p-values from the joint F-tests.

To further delineate the dynamic response to the shadow price variation, I evaluate Implication (2a) at the mean level of the ratio Cum_{jtq}/TAT_{jq} in the middle of the contract year (i.e., at the beginning of week $t = 27$). I estimate that moving one month farther away from the end of the contract year is associated with a 1 percentage point increase in the probability of any weekly claim and a 2.8 percent increase in weekly family health care utilization, on average.[25] This estimate is in line with Aron-Dine et al. (2015) who find that enrollment a month earlier (and thus having one month more to reach the unadjusted deductible) is associated with a 1 percentage point increase in the probability of any claim and a 2.2 to 7.5 percent increase in health care utilization in the first three contract months.

Regarding Implication (2b), a 10 percentage point increase in the ratio Cum_{jtq}/TAT_{jq} is associated with a 1 percentage point increase in the probability of any weekly claim and a 5.9 percent increase in weekly family health care utilization, on average, when evaluated at the beginning of week $t = 27$.[26] To my knowledge, these are the first estimates of the relationship between within-year health care demand and proportion of the cap consumed using variation in the shadow price of care. These effects are exacerbated the further the family is from the end of the contract year or, in other words, when families have more opportunities remaining to consume. For example, when there are 11 months remaining, a 10 percentage point increase in the ratio Cum_{jtq}/TAT_{jq} is associated with a 2 percentage point increase in the probability of any weekly claim and a 10.9 percent increase in weekly health care utilization, on average.

Collectively, my results provide support for the hypothesis that families internalize the nonlinear nature of the incentive scheme and are responsive, in an statistically and economically meaningful way, to variation in the shadow price of care within the coverage period. While these regression models are useful for showing associations between variables, they are less useful for predicting

[25] $\%\Delta y \approx 100 \times \frac{52}{12} \times \left[\beta_1^{\text{below cap}} + \beta_3^{\text{below cap}} \times \overline{\frac{Cum_{j(27)q}}{TAT_{jq}}}\right]$, with $\overline{\frac{Cum_{j(27)q}}{TAT_{jq}}} = 0.1778$.

[26] $\%\Delta y \approx 100 \times 0.10 \times [\beta_2^{\text{below cap}} + \beta_3^{\text{below cap}} \times (T - 27 + 1)]$, with $T = 52$.

Table 3: Weekly response to the variation in the shadow price of care

	(1) Share with any utilization	(2) Utilization in 2019 $ (in logs)
$\beta_1^{\text{below cap}}$: weeks left, non-last year	0.001***	0.003***
	(0.000)	(0.001)
$\beta_2^{\text{below cap}}$: cum to TAT ratio, non-last year	0.017	0.023
	(0.020)	(0.100)
$\beta_3^{\text{below cap}}$: interaction, non-last year	0.004***	0.021***
	(0.001)	(0.004)
$\beta_1^{\text{above cap}}$: weeks left, non-last year	0.000	-0.002
	(0.001)	(0.004)
$\beta_2^{\text{above cap}}$: cum to TAT ratio, non-last year	0.000	0.003
	(0.000)	(0.002)
$\beta_3^{\text{above cap}}$: interaction, non-last year	0.000	-0.000
	(0.000)	(0.000)
Cost Sharing Plans	Y	Y
Free Care Plan	N	N
Family fe	Y	Y
Clustered se	Y	Y
Families	1791	1791
Family-weeks	160784	160784
Adjusted R^2	0.23	0.22
p-value Implication 1: $\beta_1^{\text{below cap}} = 0$ and $\beta_3^{\text{below cap}} = 0$	0.000	0.000
p-value Implication 2: $\beta_2^{\text{below cap}} = 0$ and $\beta_3^{\text{below cap}} = 0$	0.000	0.000
p-value Implication 3: $\beta_1^{\text{above cap}} = 0$ and $\beta_3^{\text{above cap}} = 0$	0.898	0.675
p-value Implication 4: $\beta_2^{\text{above cap}} = 0$ and $\beta_3^{\text{above cap}} = 0$	0.638	0.279

Standard errors in parentheses

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

Notes: The table reports selected least-squares coefficients estimates from equation (1). Log variables are defined as log(var + 1) to accommodate zero values. I inflate health care utilization to 2019 prices using the monthly CPI-U. Standard errors clustered at the family level are in parentheses below the coefficients.

how behavior may change in response to exogenous changes in policy. Estimation of the structural parameters of the explicit optimization problem provides for a better understanding of factors affecting health care demand and for the evaluation of alternative health insurance contracts. Informed by these findings, in the next section I develop a dynamic model of weekly health care utilization decisions at the family level.

4 Model

My model is built around the problem of a forward-looking family who is enrolled in a general nonlinear health insurance plan for a given contract length. The nonlinearity of the plan arises from deductibles, coinsurance rates, and maximum out-of-pocket expenditures. In order to study the family's problem, I develop a single-agent, finite-horizon, dynamic, stochastic model of health care utilization at the family level combining elements of the annual model of health care demand from Einav et al. (2013) and the within-year model of internet demand from Nevo et al. (2016). An important feature of my model is that it explicitly incorporates the possibility of zero health care utilization as the optimal choice for a given period. Moreover, the generosity of insurance coverage can potentially affect the decision of whether or not to consume any health care.[27]

A period in my model is a contract week. After observing its realized health state, an expected-utility-maximizing family makes an optimal health care utilization decision every period. Since families are forward-looking, they form expectations on future health care utilization and internalize the dynamic pricing effect induced by the nonlinearities of the plan. This way, the model incorporates the fact that utilization decisions are made throughout the coverage period, before the uncertainty about subsequent health states is fully resolved.

4.1 Preliminaries

Families in my model are heterogeneous along several dimensions, which are unobserved to the econometrician and potentially correlated. For clarity of exposition, I omit the family subscript for now, and then in Section 5, I describe how families vary.

At the time of each weekly utilization choice, a family is characterized by its current health state realization ν, the beliefs about its subsequent health realizations $F_\nu(.)$, and its price sensitivity ω. The random variable ν captures the uncertain aspect of demand for health care, with higher ν representing sicker family members who demand greater health care utilization. The parameter ω determines how responsive health care utilization decisions are to insurance coverage. In other words, ω affects the family's price elasticity of demand for health care. Families with higher ω increase their utilization more sharply in response to more generous insurance coverage.

4.2 Utility Function

From the family's point of view, insurance coverage, denoted by k, is taken as given, and its health care utilization decision maximizes a trade-off between health and money. Following Einav et al.

[27] The proportion of zeros is not a prominent feature in annual models of health care demand, so the literature has traditionally avoided corner solutions or defined a plan-invariant proportion of zeros.

(2013), the family's per-period utility is separable in health and money and can be written as follows:

$$u(c_t;\nu_t,\omega,k) = \underbrace{\left[(c_t-\nu_t) - \frac{1}{2\,\omega}\,(c_t-\nu_t)^2\right]}_{b(c_t-\nu_t;\omega)} + \underbrace{\left[y_t - premium_{tk} - OOP(c_t, C_{t-1};k)\right]}_{x(c_t)}, \qquad (3)$$

where $c_t \geq 0$ represents the dollar consumption of covered health care goods and services for contract week t, including both the portion paid out-of-pocket by the family (if any) and the part paid by the insurance company; ν_t is the monetized health realization; and C_{t-1} represents accumulated health care utilization entering week t. I explicitly write the per-period residual income, $x(c_t)$, as the initial period-income y_t minus the per-period premium associated with coverage k and the out-of-pocket expenditure $OOP(c_t, C_{t-1};k)$ associated with utilization c_t under coverage k.[28] Naturally, $x(c_t)$ is (weakly) decreasing in c_t at a rate that depends on coverage k.[29]

The first term $b(c_t-\nu_t;\omega)$ is quadratic in its first argument, with ω affecting its curvature. It is increasing for low levels of utilization, when treatment improves health, and is decreasing eventually, when there is only marginal health benefit from treatment and time costs dominate. Thus, the marginal benefit from incremental utilization is decreasing. Using this formulation, the underlying health realization ν_t plays the role of shifting the level of optimal health care utilization, $c_t^\star$. Since ω is constrained to be strictly positive, period utility is increasing in ω.

To facilitate intuition, I consider here optimal utilization under a linear coverage contract where the OOP price remains constant throughout the year irrespective of past cumulative utilization. Thus, $OOP(c,.;k) = coins_k \times c$, where $coins_k$ represents the constant coinsurance rate of coverage k, $coins_k \in [0,1]$. Per-period optimal health care utilization is given by

$$c^\star(\nu,\omega;k) = \max\big[0, \nu + \omega \times (1 - coins_k)\big]. \qquad (4)$$

Abstracting from the potential truncation of utilization at zero, the family optimally chooses $c^\star = \nu$ under no insurance, i.e., when $coins_k = 1$, and $c^\star = \nu+\omega$ under full insurance, i.e., when $coins_k = 0$. Thus, ω can be thought of as the incremental utilization attributed to the change in coverage from no insurance to full insurance or, in other words, the full scope of moral hazard (per period).

As can be seen from equation (3), families enrolled in a general nonlinear health insurance plan do not always pay the total price of health care because of the plan's cost-sharing arrangement. Rather, a family pays a dollar amount out-of-pocket that is determined by the total price of health care, insurance plan characteristics, and accumulated health care utilization during the coverage

[28] I denote the remaining consumption before the TAT is reached as $\overline{C_t} = \overline{C_t}(C_{t-1};k) \equiv \max\{\text{TAT}_k - C_{t-1}, 0\}$. Then, under zero-deductible contracts, I can define $OOP(c_t, C_{t-1};k) \equiv coins_k \times \min\{c_t, \overline{C_t}\}$.

[29] This structure assumes that a family consumes all per-period income by the end of each contract week, as saving decisions are not observed in the data. This is a standard assumption in this literature.

period. As a consequence, the family faces a nonlinear budget set. The out-of-pocket expenditure function, $OOP(.)$, contains these nonlinearities.

Consider the case of a family enrolled in a plan with no deductible, a 25 percent coinsurance rate, and maximum out-of-pocket expenditure of \$750. Entering a given week with \$0 accumulated health care consumption, this family is charged \$1,000 for a medical visit. In this case, the family pays \$250 out of pocket (i.e., $\min[0.25 \times 1,000, 750]$). However, if the same family were to have accumulated \$2,500 in health care utilization prior to the visit, including both the portion paid out-of-pocket and the part paid by the insurance company, then it would pay only \$125, which is the minimum between \$250 and what is left to hit the family's maximum out-of-pocket expenditure of \$750.[30]

The timing of the model is as follows. At the beginning of contract week t, families learn their realization of the period-t health state, ν_t. Taking into account the plan characteristics, the accumulated health care utilization, and the expected future health risk, families choose the optimal level of health care utilization for period t. By the end of contract week t, families update their accumulated health care utilization level, which under a general nonlinear contract determines the price of health care for the subsequent contract week.

Utility from covered health care services is assumed to be additively separable over all weeks in the coverage period. For any given health insurance plan k, denote the number of weeks in the coverage period by T_k. Conditional on being enrolled in insurance plan k, the family's problem is as follows:

$$\max_{\{c_1,\ldots,c_{T_k}\}\in\mathbb{R}_+^{T_k}} \quad \sum_{t=1}^{T_k} \delta^{t-1}\mathbb{E}[u(c_t;\nu_t,\omega,k)], \quad \text{s.t.} \begin{cases} OOP(C_{T_k},0;k) + Y_{T_k} + premium_k \leq I\,, \\ C_{T_k} = \sum_{t=1}^{T_k} c_t, \quad Y_{T_k} = \sum_{t=1}^{T_k} y_t\,, \end{cases} \tag{5}$$

where δ represents the weekly discount rate. From a period-t point of view, the expectation is taken with respect to the uncertainty involving the future health realizations ν_m, $m = \{t+1,\ldots,T_k\}$. I assume that wealth, I, is large enough so that it does not constrain covered health care utilization decisions.[31]

4.3 The Dynamic Optimization Problem

The family's objective is to maximize the expected discounted future utility by selecting the optimal sequence of health care utilization, c_t, for $t = 1,\ldots,T_k$. In this subsection, I describe the family's

[30] A family enrolled in a zero-deductible plan, with a 25 percent coinsurance rate, and maximum out-of-pocket expenditure of \$750 has an associated TAT of \$750/0.25 = \$3000. Following the notation of equation (3), $OOP(1000, 2500; k) = 0.25 \times \min[1000, (3000 - 2500)] = \$125$.

[31] This is a reasonable assumption in the context of my data since the RAND HIE features low caps relative to income.

dynamic optimization problem that captures the health care utilization decisions made repeatedly over the course of a coverage period, taking into account the uncertainty about subsequent health states.

In the last contract week of the coverage period, T_k, the model becomes static: cumulative health care utilization resets to zero at the beginning of the following coverage period despite the period-T_k decision. Denote the period-T_k optimal level of covered health care utilization by the function $c^{\star}_{T_k} = c^{\star}_{T_k}(C_{T_k-1}, \nu_{T_k}; k)$. The family's utility in the terminal period is then given by

$$V_{T_k}(C_{T_k-1}, \nu_{T_k}; k) = (c^{\star}_{T_k} - \nu_{T_k}) - \frac{1}{2\,\omega}\,(c^{\star}_{T_k} - \nu_{T_k})^2 + x(c^{\star}_{T_k})\,. \tag{6}$$

For any other week $t < T_k$, covered health care utilization counts toward the family's TAT and affects the next period's state, so the optimal policy function for a family incorporates this. I therefore solve for the optimal health care utilization decision recursively. Then the family's optimal decision in period t satisfies

$$c^{\star}_t(C_{t-1}, \nu_t; k) = \max\left[0, \nu_t + \omega\left(1 - \frac{\partial OOP(c_t, C_{t-1}; k)}{\partial c_t} + \delta\frac{\partial \mathbb{E}\big[V_{t+1}(C_{t-1} + c_t, \nu_{t+1}; k)\big]}{\partial c_t}\right)\right], \tag{7}$$

where the term $\partial OOP(.)/\partial c_t$ represents the spot price of care and the term $\partial \mathbb{E}\big[V_{t+1}(.)\big]/\partial c_t$ captures the reduction in future expected prices via the nonlinearities of the contract.

In each decision period t, the state is defined by three components. First, the contract week t which determines the number of weeks left until the end of the coverage period, $T_k - t + 1$. Second, the accumulated health care utilization up until period t, C_{t-1}. And third, the stochastic health state ν_t which is known to the family at the beginning of period t. So the vector (t, C_{t-1}, ν_t) provides a complete description of the state at time t. For brevity, I describe the state vector as (C_{t-1}, ν_t) and index the policy and value functions by t.

The value function for each ordered pair (C_{t-1}, ν_t) and for any $t < T_k$ is given by

$$V_t(C_{t-1}, \nu_t; k) = \max_{c_t}\Big[u\big(c_t; \nu_t, \omega, k\big) + \delta\mathbb{E}\big[V_{t+1}(C_t, \nu_{t+1}; k)\big]\Big]\,, \tag{8}$$

where $c_t \geq 0$ and $C_t = C_{t-1} + c_t$.

To provide a clear understanding of the dynamic pricing effect, abstract for a moment from the discontinuous nature of the nonlinear price structure and the potential truncation of utilization at zero. Then, the solution to equation (8) would be characterized by the following first-order condition:

$$\frac{\partial u}{\partial c_t} + \frac{\partial u}{\partial x_t}\,\frac{\partial x_t}{\partial c_t} + \delta\mathbb{E}\left[\frac{\partial V_{t+1}}{\partial C_t}\right] = 0\,. \tag{9}$$

The first term reflects the consumption value of health care. The second term reflects the direct monetary cost of that consumption, expressed in utility terms. Finally, the third term reflects the effect of current health care utilization on future expected prices. It is this last effect that is of central interest in this paper.

4.4 The Shadow Price of Health Care

The shadow price of health care is a combination of the spot price and the option value associated with lower future expected prices. In a multi-period model with uncertain future health needs and a nonlinear price schedule, any health care expenditure below the TAT reduces the remaining distance to the TAT and, hence, the future expected prices. I define the *shadow price* of covered health care as

$$\widetilde{sp}_t(c_t, C_{t-1}; k) = \begin{cases} \dfrac{\partial OOP(c_t, C_{t-1}; k)}{\partial c_t} - \delta \dfrac{\partial \mathbb{E}\big[V_{t+1}(C_{t-1} + c_t, \nu_{t+1}; k)\big]}{\partial c_t} & , \text{ if } \quad C_{t-1} + c_t < TAT_k \\ 0 & , \text{ if } \quad C_{t-1} + c_t \geq TAT_k \end{cases}, \tag{10}$$

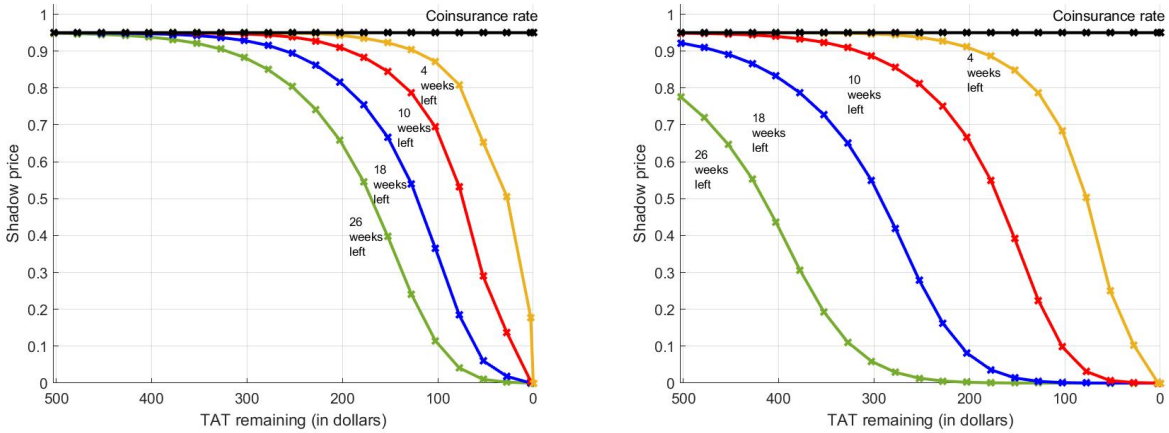
where the first part of equation (10) represents the shadow price for families who have not reached their caps on OOP spending yet, while the second part shows the shadow price once they hit the cap. The shadow price of care is equal to the marginal out-of-pocket price, $\partial OOP(.)/\partial c_t$, minus the marginal value of reducing the remaining distance to hit the cap. This latter value is the rate a family would pay (ex ante) to exchange the current insurance policy for one with the TAT reduced by one dollar.

The presence of the term $\partial \mathbb{E}\big[V_{t+1}(.)\big]/\partial c_t$ is the crucial distinction between my model and annual models or multi-period, static models of health care demand. In annual models, the shadow price of care always coincides with the end-of-year price. This is because families make a one-shot decision regarding their total annual health care utilization, which place them either below the deductible, between the deductible and the cap, or above the cap with certainty. In other words, annual models remove the uncertainty about future health care needs, so that families have perfect foresight about their end-of-year price and adjust their annual health care utilization decision accordingly. Multi-period static models assume that families are myopic and respond only to the out-of-pocket price of the period they are taking the decision. So in these models, the shadow price always coincides with the spot price of health care.

Figure 2 provides a graphical illustration of the properties of the shadow price for a family with a zero-deductible plan, a 95 percent coinsurance rate, and a \$1,000 cap on OOP spending. In order to compute the option value component of the shadow price I need to specify values for the distribution of health shocks F_ν and the price sensitivity ω, although the patterns displayed in the figure hold generally. In panel 2(a) I use values corresponding to a low mean of F_ν and in panel

2(b) a high mean. There are three main properties to highlight. First, the shadow price ranges from 0 (once the cap is hit) to the coinsurance rate of 0.95. Second, as cumulative expenditures increase and the TAT remaining falls, the shadow price also falls. Third, the rate at which it falls is increasing in the amount of time left. These last two properties capture the mechanisms I document using data from the RAND Health Insurance Experiment (see Implications (2a) and (2b) in Section 3).

Figure 2: Model properties of the shadow price of care



(a) Shadow price, low mean of F_ν

(b) Shadow price, high mean of F_ν

The shadow price of care also depends on expected future health care needs. Intuitively, the marginal value of reducing the remaining distance to hit the cap is lower for relatively healthy people, whose probability of becoming sick in the future is low, compared to people with poor health, *ceteris paribus*. Figure 2(b) shows the shadow price for a family sicker than the one considered in Figure 2(a), in that it has a higher mean of the health shocks distribution. As can be seen, the shadow price curves shift inwards, implying that the shadow price falls as the mean health risk increases. It can also be shown that if the variance of the shocks rises, given a constant mean, the shadow price curves become less steep (see Appendix C).

5 Econometric Specification

I estimate the model developed in Section 4 by extending the approach proposed by Ackerberg (2009), Bajari et al. (2007), Fox et al. (2011), and Fox et al. (2016); and recently applied by Nevo et al. (2016) in the context of demand for residential broadband and Blundell et al. (2020) in firms' investment decisions in pollution abatement technologies. This framework allows me to incorporate flexibly-correlated unobserved heterogeneity in several dimensions related to family health risk, preferences for visiting a doctor, and price sensitivity, without requiring parametric

assumptions. The structural estimation of the model proceeds by combining a method-of-moments approach with a simple nonparametric estimator for the distribution of the correlated random coefficients. This section presents the estimation approach and discusses identification.

5.1 Parameterization

Families in my model are defined by their beliefs about their subsequent health status $F_\nu(.)$, their price sensitivity parameter ω, and their previous year (PY) income. I allow all these objects to flexible vary across families, but assume they remain constant within a contract year. Yet the family type can change across experimental years to capture cross-year differences in family composition that may affect health care demand.

The health state ν_{th} is a time-varying and type-specific health shock, which represents the period-t shock to the family's health capital stock.[32] Health realizations ν_{th} are assumed to be independently and identically distributed and drawn from a (shifted) log-normal distribution with support (κ_h, ∞). The assumption of no cross-week correlation in health shocks after conditioning on family type is in line with my estimates of implication (2d) in Section 3 and the findings in previous literature.[33]

Before the uncertainty is resolved, families believe that

$$\log(\nu_{th} - \kappa_h) \sim N(\mu_h, \sigma_h^2)\,, \tag{11}$$

and these beliefs are correct. Assuming a log-normal distribution for ν is natural, as the distribution of weekly health care utilization is highly skewed. The additional parameter κ_h is used to capture the significant fraction of families who have zero health care utilization within a week. When κ_h is negative, the support of the implied distribution of ν_{th} is expanded, allowing for ν_{th} to obtain negative values, which may lead to zero health care demand. Therefore, expected health care needs for a week are given by

$$\bar{\nu}(\mu, \sigma, \kappa) = \exp(\mu + 0.5\,\sigma^2) + \kappa\,. \tag{12}$$

Finally, I include PY income as a component of the family type to capture the variability of the cap on out-of-pocket expenditures within each experimental plan.[34]

[32] ν_{th} captures the composite shock from illnesses at period t plus health capital depreciation from period $t-1$ to period t.

[33] See e.g., Einav et al. (2015) in the context of Medicare Part D, who find that conditional on allowing for unobserved heterogeneity across individuals in their permanent health state, the remaining week-to-week correlation is not very important.

[34] Income is the only dimension of family heterogeneity in the model that is observable to the econometrician, although only partially. The income distribution is censored from above, because I can only recover income from those families with MDEs strictly smaller than \$1,000.

5.2 Estimation

I estimate the joint distribution of unobserved heterogeneity using a method-of-moments approach similar to the two-step algorithms proposed by Ackerberg (2009), Bajari et al. (2007), and Fox et al. (2011); and first applied by Nevo et al. (2016). This estimator is flexible, easy to program, and computationally advantageous compared to alternative estimators for random coefficient models. For complex structural dynamic models, one does not need to nest a solution to the economic model during optimization. The estimator uses a finite and fixed grid of random coefficient vectors as mixture components to construct the distribution from the estimated probability weight of every component. The methodology exploits a re-parametrization of the underlying model so that the new parameters of interest (the weights on each type) enter linearly. Because of this linearity, the model can be estimated using inequality constrained least squares (ICLS). The ICLS minimization problem is convex, so a standard least squares algorithm will find a global optimum. By reducing the computational burden, the methodology allows me to relax several strong assumptions frequently imposed on the joint distribution of random coefficients. I do not need to assume that the random coefficients are mutually independent or that they are symmetrically distributed. The statistical and shape properties of the distributions are learned directly from the data once the parameters are estimated.

These advantages of the Fox et al. (2011) estimator are in contrast to previous approaches in the literature, which are highly nonlinear and computationally expensive. Researchers have tended to specify a parametric distribution and estimate its parameters. Estimation usually proceeds by simulation: maximum likelihood or the method of moments. These methods are computational demanding, specially for high-dimensional vectors of random coefficients (Bajari et al., 2007). Moreover, the specified distributions usually feature undesired properties.[35] Nonparametric methods offer the possibility of not being as constrained by distributional assumptions. The most common frequentist, mixtures estimator is nonparametric maximum likelihood (Heckman and Singer, 1984). Often the expectation-maximization (EM) algorithm is used for computation, which is sensitive to its starting values and is not guaranteed to converge to a global optimum. Moreover, the number of support points allowed is generally small, often only two or three, and can take the wrong sign if estimation is not constrained. Hierarchical Bayesian estimation is an alternative (Rossi et al., 1996). For example, Einav et al. (2013) employ a Bayesian hierarchical model to approximate the random coefficients' distribution. The estimator uses a Markov Chain Monte Carlo Gibbs sampling, which requires training and monitoring by the user. Moreover, the procedure usually involves evaluating the objective function many times, which is computational demanding specially in complex dynamic models.

[35] The normal distribution is probably the most widely used; however, its support on both sides of zero makes it problematic for coefficients that are necessarily signed. Lognormal distributions are usually used to avoid wrong signs. Yet they have relatively thick tails extending without bound, which implies that a share of the population has implausibly large values for the relevant coefficients (Train, 2008).

I overcome these limitations by using the methodology of Fox et al. (2011), which consists of two steps: a computational step and an estimation step. For the computational step, I fix a large but finite grid of H types in the five-dimensional space, where a family type is characterized by the vector $\boldsymbol{\beta}_h = (\mu_h, \sigma_h, \kappa_h, \omega_h, PY income_h)$. Then, for each plan k and family type h, I solve the finite-horizon dynamic programming problem described in Section 4 recursively, starting from the last period T_k, and solving backwards period-by-period until the first contract week, and collecting the sequence of decision rules in each t. This way I construct the optimal policy given initial condition $C_0 = 0$ and any realization of $\{\nu_t\}_{t=1}^{T_k}$. Because a family does not know the realization of the health demand shock ν_t prior to period t, I integrate over its support. The solution to the dynamic programming problem for each plan and family type can be characterized by the expected value functions, $\mathbb{E}[V_t(C_{t-1}; k, h)]$, and expected policy functions, $\mathbb{E}[c_t^\star(C_{t-1}; k, h)]$. The solution to the dynamic program implies a distribution for the number of weeks spent in particular states (t, C_{t-1}) over a coverage period.

In the second stage, I estimate the weight associated with each family type, θ_{kh}, to match the weighted average of the behavior predicted by the model to moments from the data using inequality constrained least squares. Each moment $G_j(\boldsymbol{\theta}_k)$ can be written as the difference between some moment in the data and the weighted average of the type moments predicted by the model:[36]

$$G_j(\boldsymbol{\theta}_k) = m_{kj}^{\text{data}} - m_{kj}^{\text{model}}(\boldsymbol{\theta}_k) = m_{kj}^{\text{data}} - \sum_{h=1}^{H} \theta_{kh} \times m_{kjh}^{\text{model}}(\boldsymbol{\beta}_h). \tag{13}$$

The key insight from equation (13) is that the new parameters of interest $\boldsymbol{\theta}_k$ enter linearly, irrespective of the highly nonlinear model used to compute the type-specific moments $m_{kjh}^{\text{model}}(\boldsymbol{\beta}_h)$. Formally, for each plan k, the methodology chooses weights $\widehat{\boldsymbol{\theta}}_k$ to satisfy

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}} = \text{argmin}_{\boldsymbol{\theta}_k} G'(\boldsymbol{\theta}_k) G(\boldsymbol{\theta}_k) \quad \text{subject to} \quad \begin{cases} \theta_{kh} \geq 0 \quad \forall h \\ \sum_{h=1}^{H} \theta_{kh} = 1 \\ \sum_{h=1}^{H} \theta_{kh} \times \mathbb{1}[\text{MDE}_{kh} \in \text{bin}_r] = \pi_{kr}^{\text{data}} \end{cases} \tag{14}$$

The first two sets of constraints in equation (14) restrict the weights to be non-negative and to sum up to 1 for each plan. The last set of constraints uses the plan-specific empirical distribution of OOP limits to partition the type distribution according to the type-specific PY income.[37] For

[36] Following Bajari et al. (2007), the second term of equation (13) is a series estimator that approximates an unknown function $\widetilde{m}_{kj}^{model}$ with the approximation $\widetilde{m}_{kj}^{model} \approx \sum_{h=1}^{H} \theta_{kh} \times m_{kjh}^{model}(\boldsymbol{\beta}_h)$. The basis functions are not the flexible mathematical functions from traditional series estimators, but the predictions of a single-agent, finite-horizon, dynamic, stochastic model of health care demand for a family of type h enrolled in health insurance plan k. The unknown frequencies θ_{kh} are structural objects, not just the approximation weights from series estimation.

[37] For each plan, I partition the empirical distribution of OOP limits in 11 bins. The first ten bins correspond to the deciles of the distribution. The last bin collects all types with caps equal to \$1,000.

each plan k, the number of unknown parameters is the number of types, $dim(\boldsymbol{\theta}_k) = H$. Then, I construct the estimated cumulative distribution function for the random coefficients as

$$\widehat{F}(\boldsymbol{\beta}; k) = \sum_{h=1}^{H} \widehat{\boldsymbol{\theta}}_k \times \mathbb{1}[\boldsymbol{\beta}_h \leq \boldsymbol{\beta}], \tag{15}$$

where $\mathbb{1}[\boldsymbol{\beta}_h \leq \boldsymbol{\beta}] = 1$ when $\boldsymbol{\beta}_h \leq \boldsymbol{\beta}$. Thus, this method provides a structural estimator for the distribution of random parameters for each plan k. This estimator is consistent under standard regularity conditions.[38]

Following Nevo et al. (2016), I choose the following moments because they have a clean connection to weekly and cumulative utilization. For each plan k, I match three sets of moments. The first set of moments is related to the mass of families at a particular state (C_{t-1}, t) or, in other words, the fraction of observations at each state. These moments capture the distribution of cumulative utilization at the end of each contract week. For example, the end-of-year distribution of cumulative utilization, which reflects the distribution of annual utilization, is a subset of moments in this set. Formally, moments in this set are given by

$$m^{model}_{kjh,\text{set }1}(\boldsymbol{\beta}_h) = \sum_{h=1}^{H} \gamma_{kjh}(C_{t-1} = C_s), \tag{16}$$

where j indexes combinations of time t and thresholds s, and $\gamma_{kjh}(C_{t-1} = C_s)$ represents the probability that a family of type h enrolled in plan k reaches contract week t with C_s dollars in accumulated health care utilization within the coverage period.

For the second set of moments, I use the mean health care utilization at each state (C_{t-1}, t), which captures the weekly utilization level and how it varies with the cap remaining and the weeks left:

$$m^{model}_{kjh,\text{set }2}(\boldsymbol{\beta}_h) = \sum_{h=1}^{H} \gamma_{kjh}(C_{t-1} = C_s) \times \mathbb{E}\left[c^{\star}_{kht} | C_{t-1} = C_s\right], \tag{17}$$

where $\mathbb{E}\left[c^{\star}_{kht} | C_{t-1} = C_s\right]$ is the mean weekly utilization in contract week t for a family of type h enrolled in plan k, conditional on past accumulated utilization level $C_{t-1} = C_s$.

Finally, the third set of moments is the mean probability of zero utilization at each state (C_{t-1}, t). These moments are different from Nevo et al. (2016). Unlike internet usage and annual health care demand, zeros are a much more prominent feature in my weekly health care utilization

[38] See Andrews (2002) who show that consistency is not affected by linear inequality constraints.

data. Formally,

$$m^{model}_{kjh,\text{set }3}(\boldsymbol{\beta}_h) = \sum_{h=1}^{H} \gamma_{kjh}(C_{t-1} = C_s) \times \Pr\big[c^{\star}_{kht} = 0 | C_{t-1} = C_s\big], \tag{18}$$

where $\Pr\big[c^{\star}_{kht} = 0 | C_{t-1} = C_s\big]$ is the probability that a family of type h enrolled in plan k does not consume covered health care services in contract week t, conditional on past accumulated utilization level $C_{t-1} = C_s$. Note that in all three sets of moments, the average is taken across all types of families in the plan, not just those that arrive at the particular state (C_{t-1}, t) with positive probability. This keeps the moments linear in the parameters $\boldsymbol{\theta}_k$, which is particularly attractive from the perspective of computational ease. Appendix D provides additional details regarding how I construct the data counterparts of these three sets of moments.

Similar to Nevo et al. (2016) and Blundell et al. (2020), I use a nonparametric block-resampling procedure to obtain standard errors for my structural parameters estimates, $\boldsymbol{\theta}_k$. Specifically, I sample the original data by family-year with replacement, keeping all 52 weeks for each of the family-years drawn. For each of 1,000 bootstrap samples, I recalculate the three sets of moments and then re-estimate the weights separately for each plan. I calculate confidence intervals for subsequent statistics and counterfactual analyses by repeating the calculation using the 1,000 different estimates of the weights.[39]

5.3 Choice of Grid Points

As mentioned earlier in this section, the methodology treats the grid of random coefficients as known and fixed. Thus, it requires the ex-ante specification of parameter grid values. In order to choose the points for the discrete five-dimensional family-type space, I follow the method of good lattice points (hereafter, *glp*). This approach has been proposed in economics by Judd (1998) in the context of integration and simulation, but to the best of my knowledge has not been applied to the estimation of dynamic programming problems. The *glp* method generates a finite set of "quasi-random" points with the property of low *discrepancy*.[40] The discrepancy of a set is a measure of how dispersed a collection of points is. Essentially, it measures the deviations from uniformity of different sets of points and provides a formal way to rank them.[41] A small discrepancy says that the collection of points evenly fills up the hypercube I^d, where d is the dimension of the grid. More details about the *glp* method can be found in Appendix E.

[39] A step-by-step description of the bootstrap procedure is provided in Appendix D.4.

[40] Blundell et al. (2020) choose their grid values by using co-prime Halton sequences, an alternative method to generate quasi-random numbers.

[41] Roughly speaking, the discrepancy of a collection of points in the interval $[0, 1]$ is low if the proportion of points in the set falling into an arbitrary subinterval $[a, b]$, $0 \leq a \leq b \leq 1$, is proportional to the length of that interval. For a more formal treatment of low-discrepancy methods, see Section 9.2 of Judd (1998).

A more common approach for the choice of grid points is the one followed by Nevo et al. (2016), which I call the method of tensor product points (hereafter, *tpp*). This standard approach consists of choosing the support points in each dimension separately, and then building the grid with all the possible combinations of them. For example, Nevo et al. (2016) choose seven points of support for each of their five dimensions of unobserved heterogeneity and build a grid with $7^5 = 16,807$ types. The *tpp* approach generates a lot of overlap between the points. I find that the *tpp* approach needs at least 50 times more points than the *glp* approach to achieve the same fit. Results of this Monte Carlo exercise are presented in Appendix F.

In order to assess the discrepancy of the sets produced by these two methods, I apply a simple discrepancy test.[42] I generate 1,000 random hypercubes of dimension five, check their discrepancy, and report the maximum discrepancy as the desired statistic. It is relevant to emphasize that, the smaller the discrepancy, the more uniformly distributed the points are inside the hypercube, and the more accurate mass points can be captured. I calculate the discrepancy for 1,000 sets of 1,069 *glp* and 1,000 sets of 16,807 *tpp* coming from seven uniformly distributed points on [0,1] in each of the 5 dimensions. The maximum discrepancies are 0.0135 and 0.3996, respectively. Therefore, the *glp*'s measure is 30 times smaller than the *tpp*'s measure. This means that in the worst scenario, the *glp* method produces points that are 30 times more uniformly distributed than the ones generated by the *tpp* method. From a statistical standpoint, a collection of points with the smaller discrepancy captures the joint distribution of random coefficients more accurately.

5.4 Identification

Conditional on the model described in Section 4, the objective is to identify the joint distribution of the parameters governing the health risk distribution, $F_\nu(.)$, the price sensitivity, ω, and the PY income. Following Fox et al. (2011), I use a nonparametric finite mixture model by fixing a large but finite grid of five-dimensional points. This way, the support points of the multivariate distribution, $\{\boldsymbol{\beta}_h\}_{h=1}^H = \{(\mu_h, \sigma_h, \kappa_h, \omega_h, \text{PY income}_h)\}_{h=1}^H$, are treated as known, and the parameters to identify are the weights, $\{\theta_{kh}\}_{h=1}^H$, on the support points. Since equation (14) accords to a linear regression subject to inequality constraints, the weights are identified as long as the matrix of model moments has full rank. In other words, the weights are identified as long as the behavior predicted by different types is not collinear over all the moments and all states used in estimation. Next I discuss how each parameter in the type vector $\boldsymbol{\beta}_h = (\mu_h, \sigma_h, \kappa_h, \omega_h, \text{PY income}_h)$ impacts the variation in predicted behavior across moments and states.

The intuition works similarly to Nevo et al. (2016), although I adapt and extend it to my context.[43] In Nevo et al. (2016), all dimensions of the space of types are unobserved by the

[42] I thank Ken Judd for suggesting this empirical test.

[43] The logic behind the identification follows closely the formal argument in Kasahara and Shimotsu (2009).

econometrician. In my context, four out of the five dimensions are unobserved, while family's PY income is censored from above. As a consequence, my identification strategy differs according to the observability of the source of permanent heterogeneity. In what follows, I first discuss the identification of the distribution of each unobserved dimension and then explain how I can provide additional identification by exploiting the variation of PY income. The identification of unobserved dimensions relies on matching the moments related to the timing and the level of weekly health care consumption observed in the data with their counterpart moments predicted by the model. The identification of the income dimension comes through the constraints imposed on the weights.

Fixing σ_h and κ_h, a higher value of μ_h generates a higher value of the weekly health shock ν_h, which in turns induces a (weakly) higher average weekly health care consumption.[44] Here, the panel dimension of my data emerges as an important determinant for identification. By modeling weekly decisions, I can identify the persistence in health care consumption that comes through the time-invariant health risk distribution of the family.

Fixing μ_h and κ_h, a change in σ_h impacts the variance of weekly consumption and therefore the likelihood of reaching certain states. This is because the lower the variance, the lower the likelihood of reaching extreme states (i.e., cumulative consumption states far apart from what the mean consumption would dictate).

The parameter κ governs the shift of the log-normal distribution of weekly health shocks. When κ is negative, the weekly shocks can take negative values, which can lead to zero consumption. Hence, κ affects primarily the extensive margin of weekly health care consumption. In particular, the set of moments related to the probability of zero consumption in each state aids in the identification of the distribution of κ.

Identification of the price sensitivity parameter ω exploits the nonlinearity of the contracts induced by the presence of the cap on annual out-of-pocket expenditures. Within plan and fixing ω, changes in the shadow price of care generate variation in the consumption level that aids in the identification of ω. Hence, how likely and early a family hits the cap provides a dynamic source of identification for ω. Table 2 in Section 2 showed that between 17 and 36 percent of family-years on plans with nonlinear pricing exceed their cap. This is important, as one source of variation needed to identify the distribution of ω relies on having enough families with a positive probability of exceeding the TAT during the contract year.

On top of the four unobserved dimensions of the type space, each type has one observed component: the family's PY income, which impacts behavior only through the cap on out-of-pocket expenditures. Hence, even though the distribution of income is censored from above in the data, the distribution of MDE is fully observed. I divide the possible MDE values in eleven groups, in

[44] Since $\log(\nu_h - \kappa_h) \sim N(\mu_h, \sigma_h^2)$, the average health shock not only depends on μ_h but also σ_h as well as κ_h. For the exact formula, see equation (12).

an attempt to balance the trade-off between adding more sources of identification versus putting more pressure on the fit. The share of family-years in each MDE bin provides information on the type distribution. In the model, each type belongs to one and only one MDE bin. So family's PY income and plan characteristics split the type space into distinct groups. In other words, they put a weight on each group of types equal to the share of family-years who belong to each MDE bin.

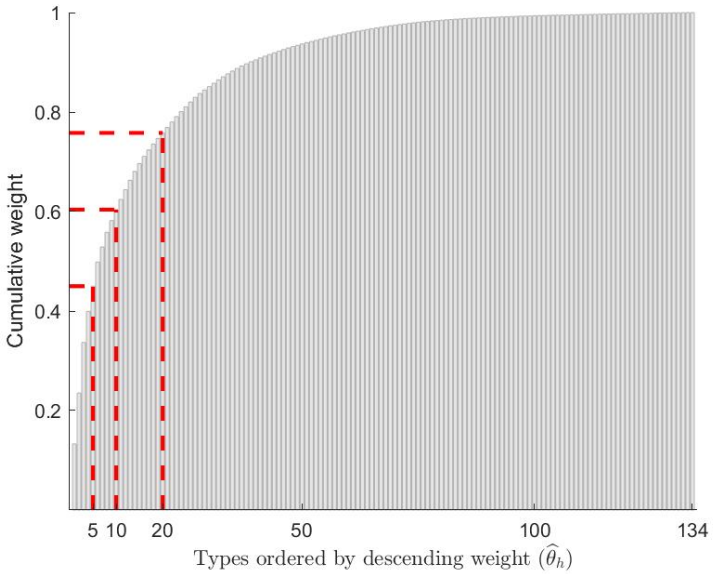
6 Results

This section begins with a discussion of the estimated type distribution, implied quantities, and model fit. I then provide estimates of the impact that moral hazard has on health care consumption and compare these estimates to the literature.

6.1 Type Distribution

I estimate a weight greater than 0.01 percent (i.e., $\theta_h > 0.0001$) for 134 types out of 1,069 considered. The first feature I find is substantial heterogeneity in the distribution of weights. The most common type accounts for 13.3 percent of the total mass, the top 5 types account for 45 percent, the top 10 for 60.4 percent, and the top 20 for 75.8 percent. Figure 3 shows the cumulative distribution of weights ordered from the most to least common type.

Figure 3: Estimated distribution of types



The second feature I find is that this heterogeneity drives a wide variety of health care utilization behavior. To get an idea of what the results imply, Table 4 presents selected statistics for the top 5

types with the highest estimated mass. The most frequent type ($h = 1$) is the healthiest, but likes going to the doctor the most. Indeed, there is only 1 percent probability that this type ends the year with zero accumulated health care demand, despite being particularly healthy (i.e., low mean of the shock ν). The second most common type is the sickest and the poorest, which results in a lower TAT, so it exceeds the cap with probability close to one even in the least generous plan. The third and fourth most common types are relatively healthy and have a strong distaste for going to the doctor, so they rarely have positive annual health care demand. The fifth type has the largest variance of the health shocks distribution, which implies a large dispersion in the distribution of annual health care demand.

Table 4: Estimates of type weights

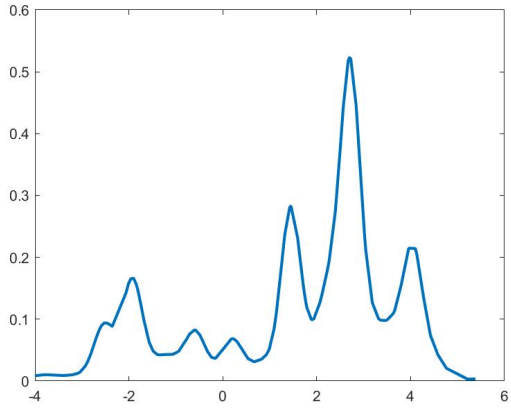
		Top 5 types				
		$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$
	Mean of the shock, with $\lambda \sim LN(\mu_h, \sigma_h)$	11.24	66.73	22.85	61.22	54.28
	Standard deviation of the shock	20.4	162.4	24.4	11.0	376.2
	κ_h: shift of the LN distribution	-20.9	-132.8	-177.8	-203.0	-172.0
	ω_h: static moral hazard	17.7	72.4	121.0	140.0	141.9
	annual income$_h$	10,945	1,459	16,558	11,646	2,862
Free care plan	$\mathbb{E}$[annual utilization]: free	457	2,168	131	187	2,583
Least generous plan	$\mathbb{E}$[annual utilization]	231	2,161	1	0	2,509
	TAT: Total Annual Threshold	1,053	230	1,053	1,053	452
	$\mathbb{P}$[hit TAT]:	<0.001	0.995	<0.001	<0.001	0.742
	$\mathbb{P}$[annual utilization = 0]:	0.012	<0.001	0.989	1	0.061
	θ_h: type weight	0.13	0.10	0.10	0.06	0.05
	Characteristics	healthiest highest κ	sickest poorest	wealthiest	lowest κ	highest SD highest ω

In Figure 4, I present the estimated marginal distributions of unobserved heterogeneity. Overall, the estimates imply an average health risk $\mathbb{E}(\nu_t)$ of \$26.33 per family-week. I estimate an average price sensitivity parameter ω of \$26.89. I estimate large heterogeneity in both health risk and price sensitivity. One standard deviation of expected health risk $\mathbb{E}(\nu_t)$ is equal to \$32.74, or a coefficient of variation of 1.24. Price sensitivity ω is also estimated to be highly heterogeneous, with a standard deviation across families of \$15.82, or a coefficient of variation of 0.59.

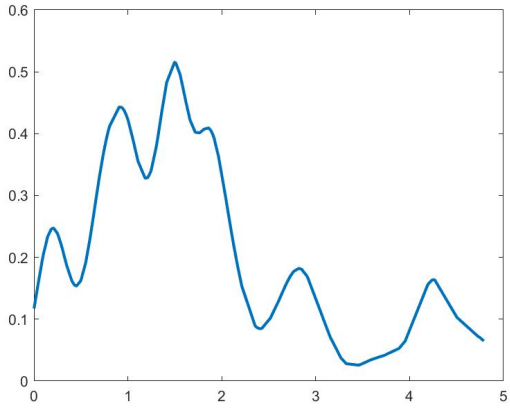
I also find that allowing for flexible-correlated heterogeneity is important. As an illustration, in Figure 5 I plot the price sensitivity ω separately for families with a low versus high preference for doctor visits (κ). First, there is a big range in terms of how price sensitive families are. Families increase their weekly health care utilization between 0 and 180 dollars when moved from no insurance to full insurance. Second, there is a strong negative correlation between price sensitivity and preference for doctor visits. Families who like to go to the doctor are less likely to increase spending due to moral hazard.

In Table 5, I report the unconditional correlations implied by the estimated type distribution. As shown in Figure 5, the unconditional correlation between ω and κ is -0.41. I also find that

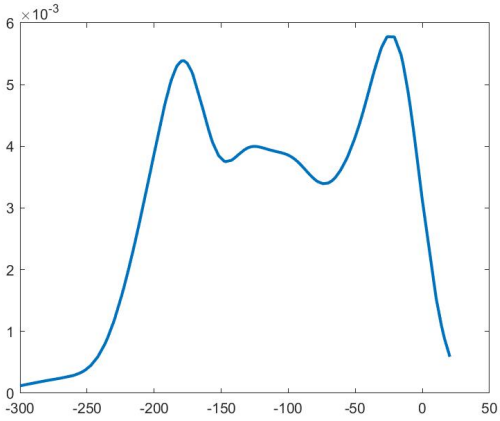
Figure 4: Estimated marginal distributions of unobserved type heterogeneity



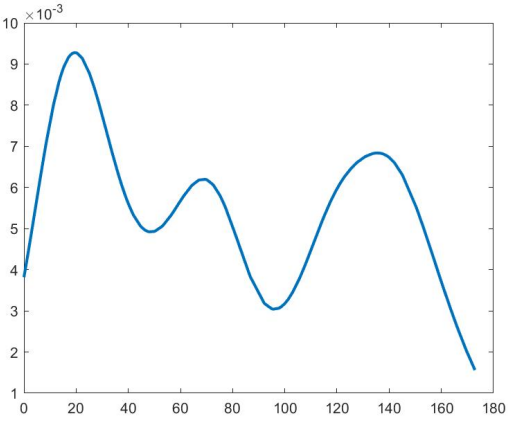
(a) Marginal distribution of μ



(b) Marginal distribution of σ

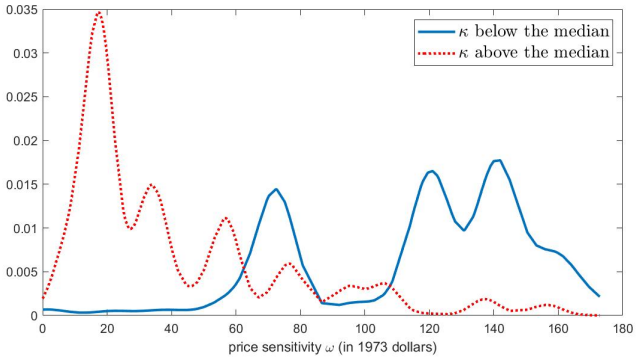


(c) Marginal distribution of κ



(d) Marginal distribution of ω

Figure 5: Price sensitivity ω, by preference for doctor visits κ



Price sensitivity ω, by preference for doctor visits κ

the unconditional correlation between μ and σ is negative and sizeable (-0.45). This implies that the health shocks of sicker individuals are less volatile and more concentrated around the mean. However, this is only part of the picture. The correlation between μ and κ is also negative and important (-0.33). Thus, higher means of the normal distribution of $\log(\nu + \kappa)$ are also associated with higher shifts towards the left.

	μ	σ	κ	ω	income
μ	1.00	-0.45	-0.33	-0.26	0.31
σ	-0.45	1.00	0.08	0.30	-0.03
κ	-0.33	0.08	1.00	-0.41	-0.10
ω	-0.26	0.30	-0.41	1.00	-0.09
income	0.31	-0.03	-0.10	-0.09	1.00

Table 5: Unconditional correlations

6.2 Model Fit

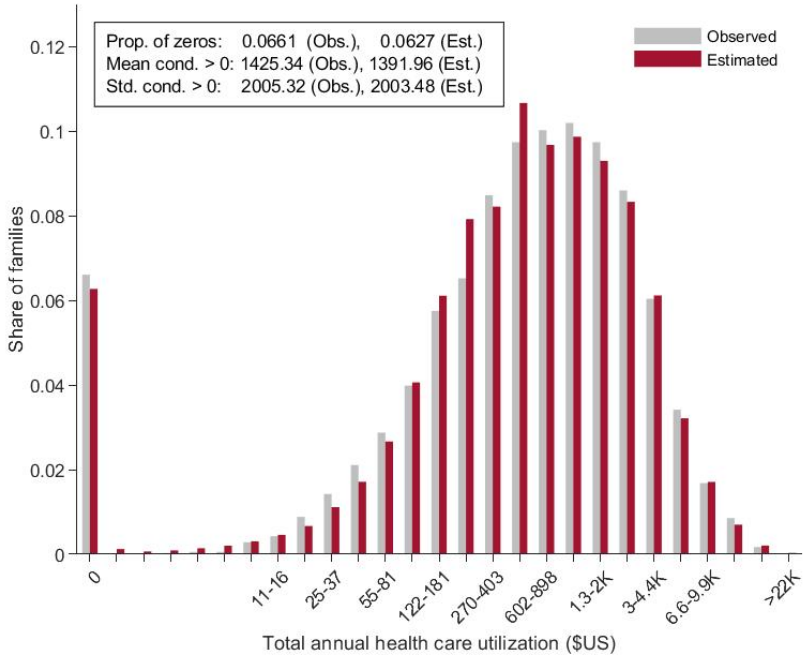
Figure 6 reports the actual and predicted distributions of annual health care utilization for the overall sample. This measure includes both the portion paid out-of-pocket by the family (if any) and the portion covered by the insurance company. Overall, the fit is quite good. For example, actual average annual health care utilization is \$1,188, while the estimate from the model is \$1,217, a difference of 2 percent. The fraction of families who have zero annual health care demand is also tightly fitted, 0.066 in the data versus 0.063 in the model. This is a very nice feature of the model, since previous papers either abstract from the corner solution at zero or display poor fit on this dimension.

Figure 7(a) shows the observed and estimated proportion of families with zero accumulated health care demand as a function of weeks left until the end of the contract. Assessing whether the model is able to accurately replicate these proportions is key for studying the optimal resetting time for deductibles and caps, which I do in the following section. As can be seen, the model fits these proportions remarkably well. Figure 7(b) reports the observed and predicted probabilities of hitting the annual OOP limit by the beginning of each contract week for the overall sample. These moments are not targeted in the estimation procedure. The model fits these probabilities remarkably well.

6.3 Dynamic Moral Hazard Estimates

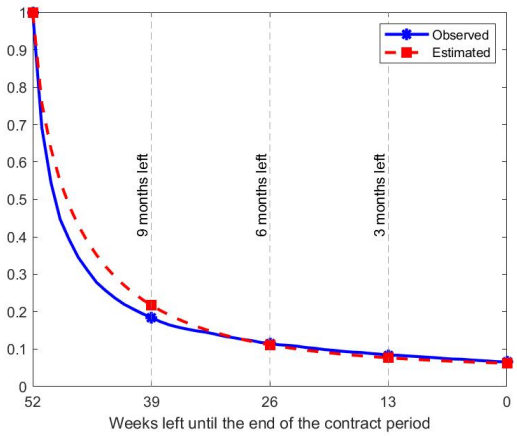
In my model, ω captures the full scope of moral hazard per week. The estimated average of ω_h is about \$77, which induces a 45 percent increase in annual utilization from no insurance to

Figure 6: Observed and Estimated Distribution of Annual Health Care Utilization

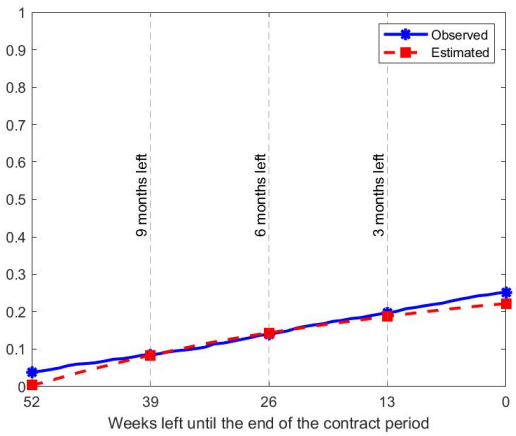


Notes: Observed (light) and estimated (dark) annual health care utilization in dollars. This figure uses a log scale: each bin $k = 1, \ldots, 26$ corresponds to utilization in the range $exp(0.4 \times (k-1)) - exp(0.4 \times k)$, with all utilization above $exp(0.4 \times 26) \approx 22K$ contained in last bin. The labels on the x-axis show the corresponding dollar amounts for each bin.

Figure 7: Model Fit



(a) Proportion of families starting each contract week with zero accumulated health care consumption



(b) Proportion of families starting each contract week above the annual cap on OOP spending

full insurance, relative to full insurance. However, knowing the full scope of moral hazard is not very useful for policy makers and contract designers given the wide popularity of nonlinear health insurance contracts. Under a typical nonlinear contract, forward-looking families internalize that current health care utilization reduces future expected prices. As a consequence, the distribution of ω does not provide a complete picture of moral hazard.

At the core of this paper is the concept of dynamic moral hazard. To disentangle the contribution of dynamic moral hazard to total moral hazard, I simulate the weekly health care utilization behavior using my estimated model but assuming families behave myopically.[45] Under this assumption, families respond only to the current spot price of care, thus shutting down dynamic moral hazard.

The additional health care utilization due to the presence of dynamic moral hazard comes from two sources: (1) the difference in weekly utilization between myopic and forward-looking families in weeks in which both are below the cap, and (2) the difference in the number of weeks spent above the cap between myopic and forward-looking families. Equation (19) below shows these two sources, and a third scenario where dynamic moral hazard is zero because both the myopic family and its forward-looking counterpart are above the cap.

$$\text{Dynamic MH}_t = \begin{cases} \omega\delta \dfrac{\partial E[V_{t+1}(C_{t-1}+c_t, \nu_{t+1}; k)]}{\partial c_t} & , \quad \text{if} \quad C_t^{\text{myopic}} < TAT_k \quad \text{and} \quad C_t < TAT_k \\ \omega \dfrac{\partial OOP(c_t^{\text{myopic}}, C_{t-1}^{\text{myopic}}; k)}{\partial c_t^{\text{myopic}}} & , \quad \text{if} \quad C_t^{\text{myopic}} < TAT_k \quad \text{and} \quad C_t \geq TAT_k \\ 0 & , \quad \text{if} \quad C_t^{\text{myopic}} \geq TAT_k \quad \text{and} \quad C_t \geq TAT_k \end{cases} \tag{19}$$

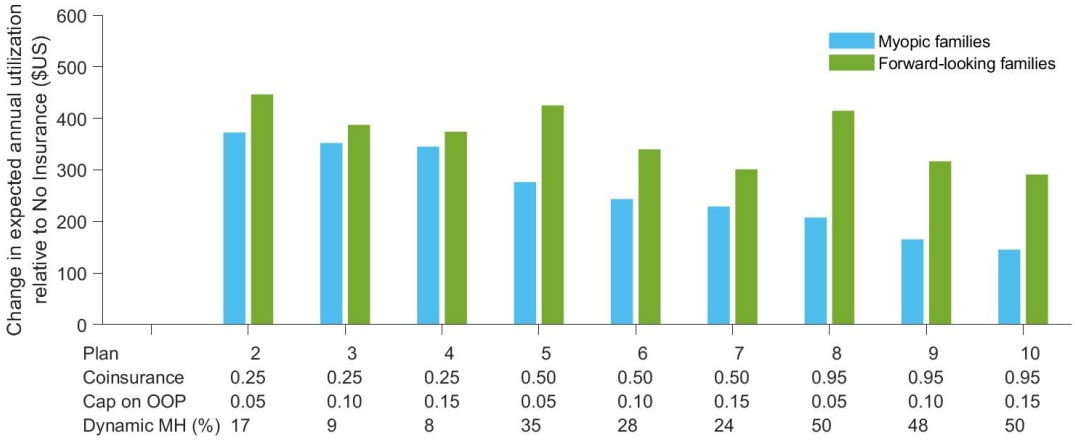
Using the estimated distribution of types in the overall sample, I simulate the behavior of forward-looking and myopic families. I decompose annual moral hazard in the portion explained by static moral hazard and the portion explained by dynamic moral hazard. As is standard in the literature, moral hazard is defined as the change in annual health care utilization in dollars relative to no insurance. To see the impact of different contract features on the importance of dynamic moral hazard, I perform this decomposition for each of the nonlinear experimental plans studied here.

Figure 8 shows the difference in average annual health care utilization comparing no insurance to one of the nonlinear experimental plans. Light-blue bars correspond to the estimates for myopic families while green bars correspond to forward-looking families. I list the plan features below the horizontal axis and the estimated contribution of dynamic moral hazard to total moral hazard in the

[45] I randomly draw 20,000 sequences of 52 health shocks for each family type with estimated weight greater than or equal to 0.01 percent, i.e., $\widehat{\theta}_h \geq 0.0001$. I simulate the two models forward for each type and nonlinear plan in the sample. Then I average across the 20,000 simulations to compute annual health care utilization for myopic and forward-looking families for each type and plan.

last row. According to my estimates, dynamic moral hazard explains between 8 and 50 percent of total moral hazard, depending on the plan, and 40 percent on average across plans. Dynamic moral hazard is more important when [1] the coinsurance rate is higher, and [2] the cap on out-of-pocket spending is smaller. This is because lower caps are more likely to be reached and higher coinsurance rates increase OOP spending relative to total utilization, which increase the likelihood of hitting the cap. This analysis highlights that health care utilization under plans with high coinsurance rates and/or low caps is particularly affected by the dynamic moral hazard. Abstracting away from these dynamic pricing incentives underestimates the cost from moral hazard, and may likely lead towards socially inefficient levels of coverage. In the next section I examine the impact of each contract feature thoroughly and study the mechanisms at play.

Figure 8: Dynamic Moral Hazard by Experimental Plan



7 Optimal Design of Employer-Sponsored Health Insurance

I now use my model and estimates to explore the impact of dynamic incentives and the associated dynamic moral hazard on welfare and optimal health insurance design. I do so in the context of employer-sponsored health insurance, in which a hypothetical employer offers a single plan in which all employees are enrolled. This allows me to focus on the welfare trade-off between risk protection and (static and dynamic) moral hazard across different plan structures, abstracting away from questions related to competition across plans. Moreover, this setting also describes a reasonable proportion of 43 percent in the U.S. population[46] and is consistent with recent papers finding that the optimal menu to offer features a single plan.[47] In the interest of exploring broader questions

[46] Proportion of employees in private-sector establishments offered only one health insurance option through their employer. Source: Agency for Healthcare Research and Quality, Medical Expenditure Panel Survey (MEPS) Insurance Component National-Level Summary Tables, 2020.

[47] See Ho and Lee (2021) and Marone and Sabety (2021).

related to the current discussion about insurance market, the counterfactual analysis considers the impact of dynamic moral hazard outside the scope of the plans offered in the RAND HIE. The contracts I study in this section vary across four features: (1) the size of the deductible, (2) the coinsurance rate after the deductible, (3) the cap on OOP spending, and (4) the resetting time for deductibles and caps.[48]

In order to rank alternative contracts, I follow the recent literature and use a measure of welfare that incorporates the benefits of risk protection and the social costs of utilization induced by insurance in a consistent framework.[49] In particular, I use the welfare metric from Einav et al. (2013) and extend the welfare decomposition in Marone and Sabety (2021) by incorporating the impact of dynamic moral hazard separate from traditional (i.e., static) moral hazard.

7.1 Measuring Welfare

Following Einav et al. (2013), I assume that families have constant absolute risk aversion (CARA) preferences and measure welfare using a certainty equivalent approach. This approach equates the expected utility for each health plan with a certain monetary payment at the beginning of the coverage year. Formally, for a family defined by the type vector $\boldsymbol{\beta}_h = (F_\nu(.;h), \omega_h, \text{PY income}_h)$, the certainty equivalent to a plan j, e_{hj}, is determined by solving

$$-\exp(-\psi\, e_{hj}) = -\int \exp(-\psi\, u^\star(\boldsymbol{\nu}; \boldsymbol{\beta}_h, j)) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}), \tag{20}$$

where ψ is the annual coefficient of absolute risk aversion and $u^\star(.)$ is the maximum annual utility for a given sequence of realized health shocks $\boldsymbol{\nu}$.[50] The assumption of CARA preferences over the annual utility implies that (1) risk preferences only impact plan utilities but not within-the-year health care utilization decisions, and (2) family income does not impact relative plan utilities.

[48] Even though all plans in the RAND HIE feature zero deductibles, the 95 percent coinsurance plans in the study closely resemble the high-deductible catastrophic plans offered today (Brook et al., 2006). Regarding the variation needed to study shorter resetting times for deductibles and caps, I leverage my within-year dynamic model of health care utilization, in which demand is affected by the number of weeks left in the contract, among other things. The key identifying variation comes from the variation in the proportion of families in the data that starts each contract week with zero accumulated health care utilization. For example, a family with zero accumulated health care demand at the beginning of contract week 27 is *de facto* facing a resetting time of six months.

[49] See e.g., Kowalski (2015), Ho and Lee (2021), and Marone and Sabety (2021).

[50] The risk aversion parameter is difficult to identify in the absence of insurance plan choice data. Hence, for the counterfactuals, I borrow from Einav et al. (2013) the estimate of the average coefficient of absolute risk aversion, $\psi = 0.0019$. This value implies that to make families indifferent between (i) a payoff of zero and (ii) an equal-odds gamble between gaining \$100 and losing \$X, the mean value of \$X is \$84.0.

Hence, the certainty equivalent for a type-h family and plan j can be written as:[51]

$$e_{hj}(\boldsymbol{\beta}_h) = -\frac{1}{\psi}\ln\left[\int \exp(-\psi\,\widetilde{u}^{\star}(\boldsymbol{\nu};\boldsymbol{\beta}_h,j))dF_{\boldsymbol{\nu}}(\boldsymbol{\nu})\right] + (Y_h - premium_j) \equiv \widetilde{e}_{hj}(\boldsymbol{\beta}_h) + (Y_h - premium_j), \tag{21}$$

where $\widetilde{e}_{hj}(\boldsymbol{\beta}_h)$ captures the family's welfare from coverage, and residual income, $Y_h - premium_j$, enters additively.[52] Using this notation, differences in $\widetilde{e}_h(.)$ across contracts with different coverages capture the willingness to pay for coverage. For example, a type-h family is willing to pay at most $\widetilde{e}_{hk}(\boldsymbol{\beta}_h) - \widetilde{e}_{hj}(\boldsymbol{\beta}_h)$ in order to increase its coverage from j to k.

I further assume that insurance providers are risk neutral. Thus, the provider's welfare when a type-h family is enrolled in contract j is given by his expected profits, or

$$\pi_{hj}(\boldsymbol{\beta}_h) = premium_j - \mathbb{E}_{\boldsymbol{\nu}}\Big[k_j\big(C_T^{\star}(\boldsymbol{\nu};\boldsymbol{\beta}_h,j)\big)\Big], \tag{22}$$

where $C_T^{\star}(.)$ represents the optimal annual health care utilization and the function $k_j(.)$ maps this dollar amount to the portion covered by the provider under the rules of contract j.

Finally, the social welfare generated by allocating a type-h family to contract j is the sum of family and provider welfare. To define a measure of social welfare that does not depend on family income, I follow the literature (see e.g., Ho and Lee (2021) and Marone and Sabety (2021)) and measure social welfare relative to full insurance, which can be expressed as the difference between willingness to pay and expected insurer cost:[53]

$$RSS_{hj}(\boldsymbol{\beta}_h) = WTP_{hj}(\boldsymbol{\beta}_h) - \mathbb{E}_{\boldsymbol{\nu}}\Big[k_j\big(C_T^{\star}(\boldsymbol{\nu};\boldsymbol{\beta}_h,j)\big) - C_T^{\star}(\boldsymbol{\nu};\boldsymbol{\beta}_h,\text{full})\Big]. \tag{23}$$

To aggregate the welfare measure across all families according to a utilitarian social welfare function, I define the average relative social welfare across all families for contract j:

$$\overline{RSS}_j(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{h=1}^{H}\theta_h \times RSS_{hj}(\boldsymbol{\beta}_h). \tag{24}$$

Decomposition of Relative Social Welfare. I extend the discussion in Azevedo and Gottlieb (2017) and the generalization in Marone and Sabety (2021) and show that relative social welfare, $RSS_{hj}(\boldsymbol{\beta}_h)$, can be decomposed in three terms: (1) the value of risk protection, (2) the social cost of *static* moral hazard, and (3) the social cost of *dynamic* moral hazard. This last component is novel

[51] $\widetilde{u}^{\star}(\boldsymbol{\nu};\boldsymbol{\beta}_h,j) = \sum_{t=1}^{T}((c_t^{\star}-\nu_t) - \frac{1}{2\omega}(c_t^{\star}-\nu_t)^2 - OOP(c_t^{\star},C_{t-1};j))$.

[52] Equation (21) highlights that in this setup, the design of an optimal employer-sponsored insurance plan and the choice of how much to contribute towards employee premiums are separable problems for the employer, as long as employer contributions take the form of a fixed dollar amount.

[53] Note that because of the CARA assumption, premiums are transfers that do not affect social welfare.

to the literature and captures that a potentially important part of the welfare gains thought to be achieved by moving families away from full insurance are not actually realized because families anticipate reaching the deductible and/or the cap and adjust utilization accordingly.

The social welfare generated by allocating a type-h family to contract j (relative to allocating the same family to the free-care contract) can also be written as:

$$RSS_{hj}(\boldsymbol{\beta}_h) = \underbrace{\Psi(j, \boldsymbol{\beta}_h)}_{\substack{\text{Relative value of} \\ \text{risk protection}}} - \Bigg[\underbrace{\mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} \frac{\omega_h}{2} p^{\star}_{thj}\left(p^{\star}_{thj} - 2\right) \right]}_{\substack{\text{Relative social cost of} \\ \text{static moral hazard}}} + \underbrace{\mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} \frac{\omega_h}{2} \left((\widetilde{sp}^{\star}_{thj} - 1)^2 - (p^{\star}_{thj} - 1)^2 \right) \right]}_{\substack{\text{Relative social cost of} \\ \text{dynamic moral hazard}}} \Bigg], \tag{25}$$

where $p^{\star}_{thj}$ represents the spot price of care and $\widetilde{sp}^{\star}_{thj}$ is the shadow price of care from equation (10), evaluated at the optimal health care utilization level.[54]

Premium Setting. As in Ho and Lee (2021), I require that insurance premiums cover families' total expected health care utilization, net of out-of-pocket payments that they make in the form of deductibles or coinsurance.[55]

Following the welfare decomposition in equation (25), three forces shape the design of health insurance contracts. Two of them are standard in the literature: risk protection and static moral hazard. In general, these forces go in opposite directions: more generous contracts provide higher protection against health risks but at the same time induce consumers to purchase additional care that they would not have bought had they faced the full cost. The third force is dynamic moral hazard, which is new to the discussion of how to manage the spending coverage trade-off. As the probability of exceeding the deductible/OOP limit increases, families internalize this and purchase more health care than they otherwise would. As I illustrate below, the presence of dynamic moral hazard can severely dampen the welfare gains associated with higher cost-sharing and plays a crucial role, distinct from static moral hazard, in determining optimal insurance contract design. In what follows, I analyze the deductible, coinsurance rate, OOP limit, and resetting time (for deductibles and OOP limits) that maximize average welfare when the employer only offers a single plan. I define the annual deductible in \$250 increments between \$0 and \$2000, the level of coinsurance to vary from 0-100 percent in 10 percent increments, annual OOP maximum ranges from \$0 (full insurance) to \$2000 in \$250 increments, and resetting times at either 3, 6, or 12 months. All dollar amounts are expressed in 1973 dollars.[56]

[54] In Appendix G, I provide details about how to arrive from the definition of relative social welfare in equation (23) to the decomposition in equation (25).

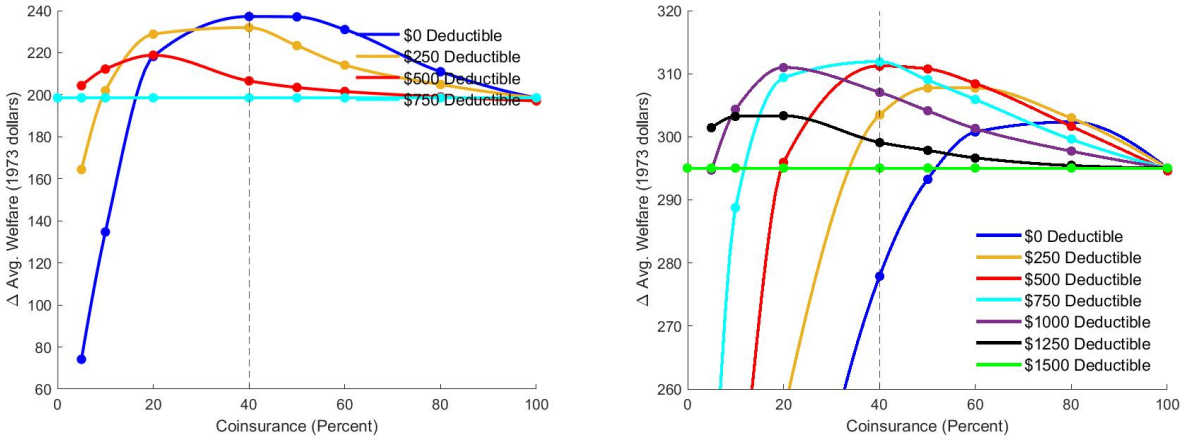
[55] These premiums are before loading costs. When loading costs are passed on to consumers, they are just transfers from the agents to the insurer. Therefore, loading costs do not affect average welfare.

[56] I refer to deductibles and OOP limits corresponding to those in the metal-tiered plans (i.e., Gold, Silver, Bronze) offered on Affordable Care Act exchanges as "low", "middle", and "high", respectively. The deductibles, coinsurance rates, and out-of-pocket maximums in 2016 were \$1,169, 21%, \$2,564 for Gold; \$3,060, 34%, \$4,872 for Silver; and

7.2 The determinants of the optimal deductible size

High deductible health plans have become increasingly common. In 2020, 31 percent of workers in the U.S. with employer-sponsored insurance were in such plans (Kaiser Family Foundation). To better understand the role played by the deductible, I fix the resetting time at the standard twelve-months level and study the optimal deductible size for different caps on OOP spending. I analyze two levels of OOP limits: a low-cap scenario, where I set the annual cap at \$750, and a high-cap scenario, where I set the cap at \$1500. In Figure 9, I plot the average social welfare from low-cap (left panel) and high-cap (right panel) plans with different deductible sizes as a function of the coinsurance rate, relative to full insurance. I find that under a low-cap plan, a zero-dollar deductible would be optimal, paired with a 40 percent coinsurance rate after the deductible. However, under a high-cap contract, a \$750 deductible would be optimal, also paired with a 40 percent coinsurance rate after the deductible.

Figure 9: Change in Average Social Welfare, Relative to Full Insurance



(a) Low cap on OOP spending

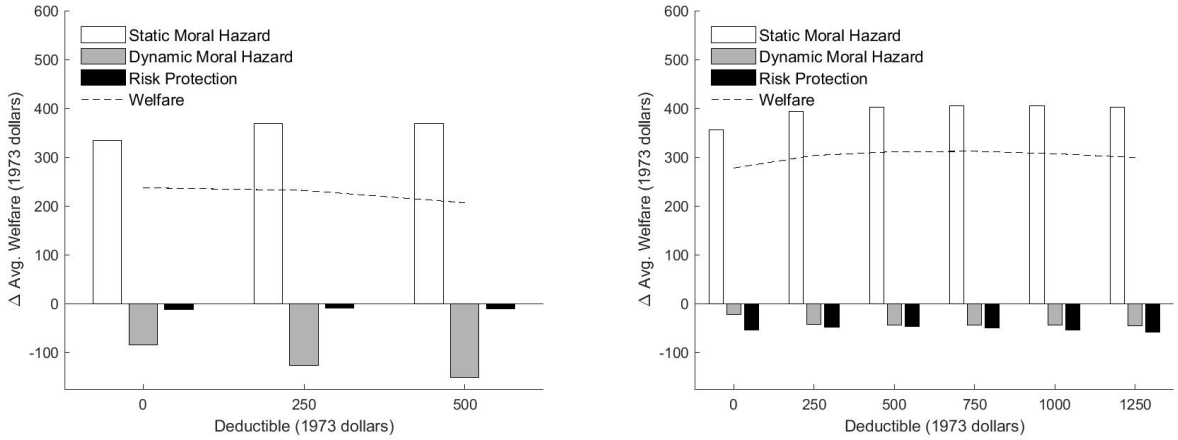
(b) High cap on OOP spending

To assess the role that each welfare component plays behind these findings, I fix the coinsurance rate at 40 percent and take advantage of the welfare decomposition in equation (25). Figure 10(a) shows each welfare component as a function of the deductible size under the low-cap scenario. Clear bars represent the average welfare due to static moral hazard, relative to full insurance. Since full insurance is the most generous plan, this component is always non-negative. As one would expect, the gains from static moral hazard are increasing in the deductible as families bear a higher portion of the bill.

Gray bars represent the average welfare due to dynamic moral hazard, relative to full insurance. This component is always non-positive under typical health insurance contracts, as consuming more

\$5,771, 48%, \$7,436 for Bronze. In 1973 dollars, the deductibles and out-of-pocket maximums are \$210, \$461 for Gold; \$550, \$876 for Silver; and \$1,038, \$1,337 for Bronze.

Figure 10: Welfare Decomposition Relative to Full Insurance



(a) Low cap on OOP spending

(b) High cap on OOP spending

today reduces future expected prices. It has the biggest gradient with respect to the deductible due to two mechanisms that reinforce each other. Higher deductibles increase OOP spending relative to total utilization, which increases the likelihood of hitting the cap. Forward-looking families respond by increasing utilization. In turn, both of these forces increase the expected number of periods spent above the deductible/cap which in turn increases utilization due to the lower prices faced after exceeding the deductible/cap. Therefore, a zero-dollar deductible, which minimizes these forces, achieves the minimum welfare loss due to dynamic moral hazard.

Lastly, black bars represent the average welfare due to risk protection, relative to full insurance. Since full insurance provides full protection against risks, this component is always non-positive. The loss due to decreased risk protection is small and roughly constant across deductible sizes because most of the risk protection comes from the low cap.

Combining the three welfare components, the dashed line represents average relative social welfare as a function of the deductible size. The big losses from dynamic moral hazard more than offset the gains from static moral hazard as the deductible increases, which makes zero-dollar deductibles optimal under low-cap contracts. Average social welfare increases $237 per year relative to full insurance, which represents 16 percent of the annual premium in full insurance, and average annual health care utilization decreases 20 percent compared to full insurance. To further highlight that dynamic moral hazard plays a key role in making zero deductibles optimal under a low-cap scenario, I shut down dynamics and re-rank plans according to welfare. I find that a zero-dollar deductible is no longer optimal. The optimal deductible is $250, and even a $500 deductible achieves higher welfare than a zero deductible (absent dynamic moral hazard).

In contrast, for high caps on OOP spending, the optimal deductible size is no longer zero. Figure 10(b) presents the welfare decomposition for the high-cap scenario as a function of the deductible

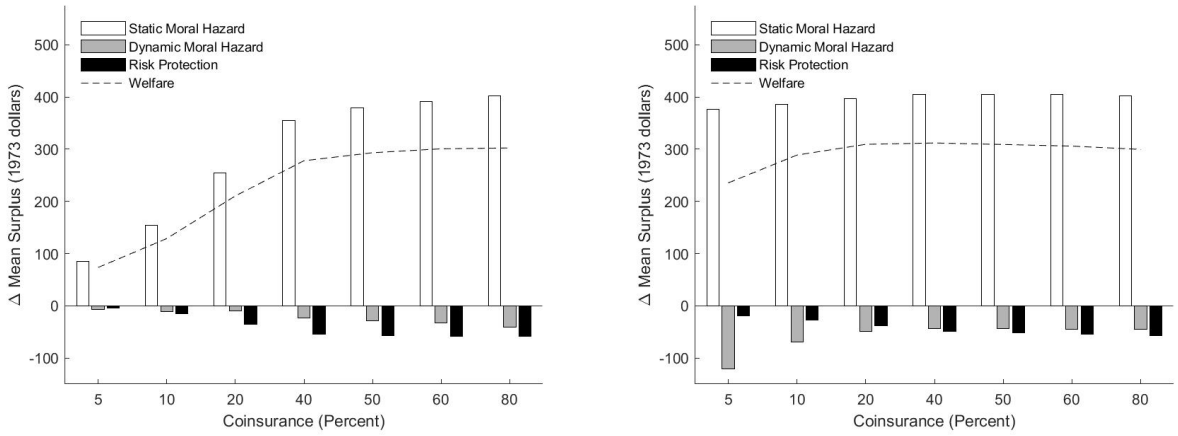
size. The main difference with the low-cap scenario is that under the high cap, the likelihood of hitting the cap is significantly smaller. This reduces the difference between the spot price and the shadow price, thus lowering the losses due to dynamic moral hazard, which were penalizing higher deductibles in the low-cap scenario. The losses from dynamic moral hazard are still increasing in the deductible, but smaller. The losses from decreased risk protection are more important now, and even bigger than the losses from dynamic moral hazard. As a result, the static moral hazard gains from higher deductibles drive the optimal deductible to \$750. Social welfare is \$312 higher and annual health care utilization is 33 percent lower compared to full insurance.

In summary, I find that zero-dollar deductibles are optimal for low to medium caps, while for high caps on OOP spending, high deductibles are welfare-maximizing. The first result coincides with the findings in Ho and Lee (2021), who focus on fairly generous (low) OOP maximums. The second result highlights that the optimal deductible size depends importantly on the cap on OOP spending, precisely due to the presence of dynamic moral hazard.

7.3 The determinants of the optimal coinsurance rate

To study how average welfare changes with the coinsurance rate, I continue to fix the resetting time at twelve months and set the cap at the high level (\$1,500). Figure 11 depicts the welfare decomposition from equation (25) as a function of the coinsurance rate, under a zero-dollar deductible (left panel) and a high deductible (\$750, right panel).

Figure 11: Welfare Decomposition Relative to Full Insurance



(a) Zero deductible

(b) High deductible

Under a zero-dollar deductible (Figure 11(a)), the gains from static moral hazard are increasing in the coinsurance rate because consumers' cost-sharing is increasing. The welfare losses from dynamic moral hazard are also increasing in the coinsurance rate, because higher coinsurance rates

allow forward-looking families to approach the cap at a faster rate. However, their magnitude is small due to the combination of zero deductible and high cap, and thus static moral hazard plays the dominant role. The losses due to decreased risk protection increase with the coinsurance rate as families pay a higher proportion out-of-pocket, but the magnitudes are also small relative to static moral hazard. Thus, static moral hazard drives the optimal coinsurance rate, which is 80 percent.

For the high-deductible scenario (Figure 11(b)), the picture is quite different. First, the gains from static moral hazard are significantly flatter across coinsurance rates, compared to the zero-deductible scenario. The reason is that the high deductible discourages health care demand, so even very low coinsurance rates after the deductible accrue big gains from static moral hazard. The loss in risk protection is increasing in the coinsurance rate, which pushes the optimal coinsurance rate down. However, for very low coinsurance rates, the losses due to dynamic moral hazard are now sizable. This is because now the dynamic moral hazard effect operates through two thresholds, the deductible and the cap, as opposed to just the cap in Figure 11(a). Under this new mechanism, low coinsurance rates make hitting the deductible very appealing. As a consequence, the shadow price decreases and utilization increases. The optimal coinsurance rate is 40 percent.

In summary, there are two main takeaways from this section. First, the optimal coinsurance rate and deductible are negatively related, as they are substitutes in providing the incentives to achieve welfare gains from static moral hazard, relative to full insurance. Second, under typical health insurance contracts that feature a deductible and a cap, dynamic moral hazard operates distinctively through these two thresholds. As a consequence, very low and very high coinsurance rates exacerbate the losses from dynamic moral hazard.

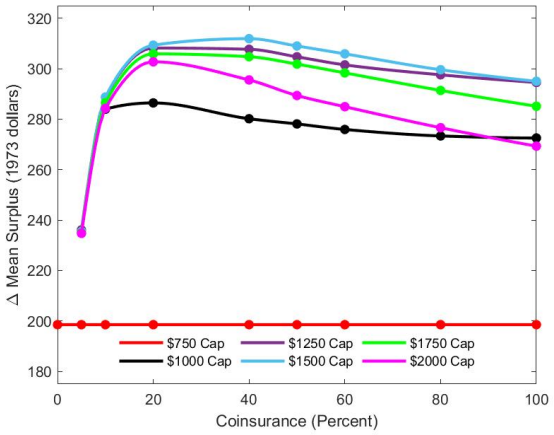
7.4 The determinants of the optimal cap on OOP spending

I now turn to the determinants of the optimal cap on OOP spending again fixing the resetting time at twelve months. I only focus on the high-deductible ($750) scenario, but the results for other deductibles are similar both qualitatively and quantitatively. In Figure 12(a), I plot the average social welfare from high-deductible plans with different caps on OOP spending as a function of the coinsurance rate, relative to full insurance. The first takeaway is that pure stop-loss contracts[57] are never optimal, irrespective of the coinsurance rate (and the deductible size, not shown). This can be seen by comparing the welfare level of the straight horizontal line, with the other curves in the plot. The second takeaway is that the optimal cap on OOP spending lies within the range of high caps for all coinsurance rates, but not very high (below $1750). The optimum across coinsurance rates occurs at a cap of $1500.

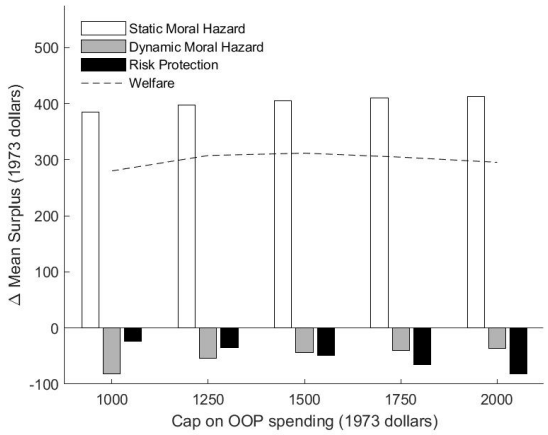
To unpack the contribution of each welfare component, I fix again the coinsurance rate at 40

[57] Pure stop-loss contracts have no insurance up to a point, and full insurance thereafter, so that the deductible and the cap on OOP spending coincide.

Figure 12: Change in Average Social Welfare Relative to Full Insurance



(a) High deductible



(b) High deductible - 40 percent coinsurance rate

percent, and present in Figure 12(b) the welfare decomposition as a function of the cap on OOP spending. The losses due to decreased risk protection are increasing in the cap since large caps leave families more exposed to health risks. Welfare gains from static moral hazard are also increasing in the cap because consumers' cost-sharing is increasing. Finally, the losses due to dynamic moral hazard are decreasing in the cap, as higher caps are less likely to be reached, thus decreasing the incentives to increase spending. Interestingly, the cap on OOP spending is the only contract feature for which the impact of static and dynamic moral hazard on welfare go in the same direction. This is why pure stop-loss contracts, in which the deductible and the cap coincide, are never optimal. Ultimately, the trade-off between net moral hazard and risk protection determines the optimal cap, achieved at $1500 (classified as high) under the high-deductible scenario.

So far I have studied the three popular cost-sharing features while fixing the resetting time at the standard twelve-months level. The overall optimal plan features a high deductible, 40 percent coinsurance rate after the deductible, and a high cap on OOP spending (equal to twice the deductible).[58] Under this optimal plan, average social welfare increases $312 per year relative to full insurance, which represents 21 percent of the annual premium under full insurance. Moreover, annual health care utilization is 33 percent lower compared to full insurance, on average.

7.5 Optimal resetting time for deductibles and OOP limits

Finally, I examine the impact of varying the timespan over which deductibles and OOP limits reset (six versus twelve months) holding fixed the contract length at the annual level. When I compare

[58] The optimal deductible size is consistent with the average family deductible in the so-called High-Deductible Health Plans (HDHP) in employer-sponsored settings in 2020, which was $4552 in 2020 dollars (or $752 in 1973 dollars). Source: Employer Health Benefits, 2020 Survey, Kaiser Family Foundation.

welfare between these policies, I adjust the deductible and cap proportionally to the resetting time. For instance, if I fix the standard deductible at \$1000, I compare welfare between a \$500 deductible that resets every six months and a \$1000 deductible that resets every twelve months.

Shorter resetting times provide more risk protection through smaller caps. The consumers' financial losses from medium to severe health shocks are more likely to be capped under shorter resetting times.[59]

A change in the resetting time for deductibles and caps may exacerbate or hinder static moral hazard. Intuitively, in the first six-month period there is (weakly) more overspending due to static moral hazard under a six-month resetting contract than a standard resetting contract. In the last six-month period, static moral hazard in the six-month resetting policy is the same as in the first six-month period, on average. However, under a standard resetting policy, the overspending due to static moral hazard in the last six-month period is (weakly) higher than in the first six-month period. Which resetting policy leads to higher overspending due to static moral hazard ultimately depends on the empirical distribution of health shocks. Relatively healthy people overspend more due to static moral hazard under shorter resetting times, while the opposite is true for relatively sick families.

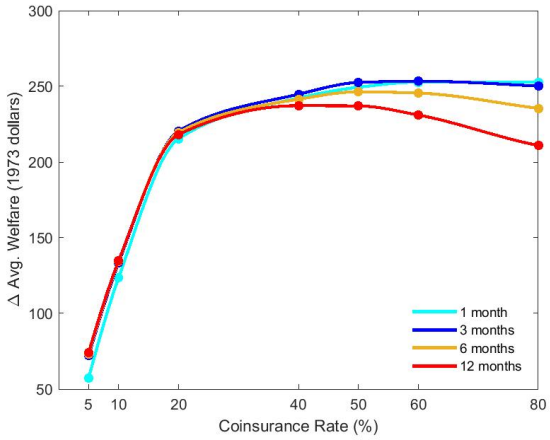
At the family level, the optimal resetting time for deductibles depends on (a) the deductible size, and (b) the family's health status and price sensitivity. In general, relatively healthy families achieve higher welfare in longer resetting times, while relatively sick families have higher welfare in shorter resetting times. This is because sick families, who anticipate reaching the deductible with high probability, induce bigger losses from dynamic moral hazard in longer resetting times.

At the aggregate level, I find that the six-months resetting deductible policy is welfare-maximizing because it limits the escalation of dynamic moral hazard. Figure 13(a) shows the average welfare for pure-deductible plans as a function of the deductible size, relative to full insurance. The figure shows that the six-months deductible achieves higher welfare for all deductible sizes, on average. In Figure 13(b) I fix the annual deductible at \$500 (medium size) and plot the welfare decomposition from equation (25) for the two resetting policies. In this case, the gains from static moral hazard (clear bars) and the losses from risk protection (black bars) are similar between the two policies. However, the gray bars show that the twelve-months deductible is associated with bigger losses due to dynamic moral hazard, so the six-months deductible achieves higher welfare. In general, the gains from static moral hazard increase in the resetting time, while both the losses from dynamic moral hazard and the losses from risk protection worsen with the resetting time. This trade-off assures an interior solution for the optimal resetting time.

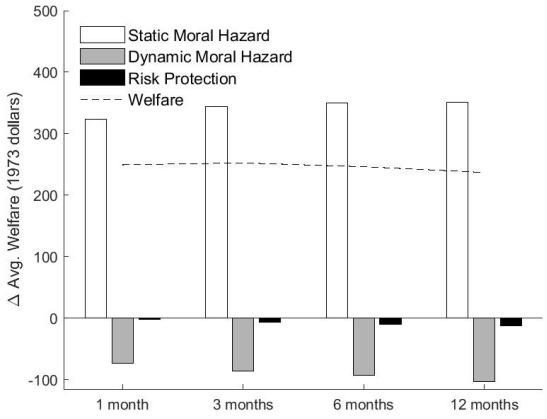
Longer resetting times induce bigger welfare losses due to dynamic moral hazard because families

[59] In general, the financial losses from severe health shocks are equally likely to be capped under shorter or standard resetting times.

Figure 13: Welfare under deductibles that reset after six versus twelve months



(a) High deductible, high cap on OOP spending



(b) High deductible, high cap on OOP spending

enjoy more periods above the deductible per year, on average. The longer the resetting time for the deductible, the more appealing it is to reach it soon because the family will enjoy the benefits of being above the deductible for more periods.[60] As a consequence, in the standard deductible policy, families spend more in health care when they are under the deductible, which puts them above the deductible proportionally earlier than in the six-months resetting policy, enjoying more periods in free care. For example, families enjoy 10.4 weeks above the $500 deductible in the standard policy, and only 9.1 weeks above the $250 deductibles in total in the two consecutive six-months resetting policies. This behavior leads to bigger welfare losses from dynamic moral hazard in the twelve-months resetting deductible policy.

To summarize, I find that the optimal resetting time is heterogeneous across the distribution of health risks and that shorter-than-standard resetting times are associated with higher welfare precisely due to their smaller losses from dynamic moral hazard. This last finding suggests that abstracting from dynamic moral hazard would favor longer resetting times, which provides an answer to the more fundamental question of why deductibles and OOP limits at shorter frequencies are not offered. Finally, whether or not different resetting times might coexist in the market would be an interesting avenue for further research.

8 Conclusion

In this paper, I study a new source of moral hazard that has been overlooked in most of the prior literature, mainly due to the use of annual models to study health care utilization decisions. I label

[60] Abstracting from truncation at zero, the utility function in equation (3) implies that for every period above the cap, a family enjoys positive utility equal to $\omega/2$.

this source as *dynamic* moral hazard to contrast with the traditional moral hazard present in static and annual models of health care demand. Under the typical nonlinear pricing scheme generated by the presence of deductibles and consumers' out-of-pocket limits in health insurance contracts, current health care utilization lowers future expected prices. Using data from the RAND Health Insurance Experiment, I show that families respond to these dynamic pricing incentives through two mechanisms: distance to the cap and time left until the contract resets. I then build and estimate a dynamic model of health care demand that incorporates consumer heterogeneity along multiple and flexible-correlated dimensions. Using my model and estimates, I document that 40 percent of total moral hazard can be attributed to the dynamic moral hazard component.

Finally, I explore the impact of these dynamic incentives and the associated dynamic moral hazard on welfare and optimal health insurance design in the context of employer-sponsored health insurance. I show that the presence and significance of dynamic moral hazard have important implications, distinct from static moral hazard, for health care demand, optimal health insurance design, and costs of the health care sector. For example, a standard approach to curbing the social cost of moral hazard is to increase consumer cost sharing. Following this rationale, there is a recent trend towards offering high deductible plans. My results imply that introducing high deductibles is an efficient measure of cost containment only if the caps on consumers' out-of-pocket spending are not too close to the deductibles. Otherwise, dynamic moral hazard severely dampens the savings thought-to-be-achieved due to static moral hazard.

References

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. *Quantitative Marketing and Economics*, 7(4):343–376.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*.

Andrews, D. K. W. (2002). Generalized method of moments estimation when a parameter is on a boundary. *Journal of Business & Economic Statistics*.

Aron-Dine, A., Einav, L., and Finkelstein, A. (2013). The rand health insurance experiment, three decades later. *Journal of Economic Perspectives*, 27(1):197–222.

Aron-Dine, A., Einav, L., Finkelstein, A., and Cullen, M. (2015). Moral hazard in health insurance: do dynamic incentives matter? *Review of Economics and Statistics*, 97(4):725–741.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*.

Atal, J. P., Fang, H., Karlsson, M., and Ziebarth, N. R. (2020). Long-term health insurance: Theory meets evidence. Technical report, National Bureau of Economic Research.

Azevedo, E. M. and Gottlieb, D. (2017). Perfect competition in markets with adverse selection. *Econometrica*, 85(1):67–105.

Bajari, P., Fox, J. T., and Ryan, S. P. (2007). Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. *American Economic Review*, 97(2):459–463.

Blundell, W., Gowrisankaran, G., and Langer, A. (2020). Escalation of scrutiny: The gains from dynamic enforcement of environmental regulations. *American Economic Review*, 110(8):2558–85.

Bourdeau, M. and Pitre, A. (1985). Tables of good lattices in four and five dimensions. *Numer. Math.*

Brook, R. H., Keeler, E. B., Lohr, K. N., Newhouse, J. P., Ware, J. E., Rogers, W. H., Davies, A. R., Sherbourne, C. D., Goldberg, G. A., Camp, P., et al. (2006). The health insurance experiment: a classic rand study speaks to the current health care reform debate. *Santa Monica, CA: RAND Corporation*.

Brot-Goldberg, Z. C., Chandra, A., Handel, B. R., and Kolstad, J. T. (2017). What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318.

Cabral, M. (2017). Claim timing and ex post adverse selection. *Review of Economic Studies*.

Chevalier, J. and Goolsbee, A. (2009). Are durable goods consumers forward-looking? evidence from college textbooks. *The Quarterly Journal of Economics*, 124(4):1853–1884.

Cronin, C. J. (2019). Insurance-induced moral hazard: A dynamic model of within-year medical care decision making under uncertainty. *International Economic Review*.

Darmouni, O. and Zeltzer, D. (2017). Horizon effects and adverse selection in health insurance markets. *working paper*.

Devereux, K., Balesh Abadi, M., and Omran, F. (2019). Correcting for transitory effects in rcts: Application to the rand health insurance experiment.

Einav, L. and Finkelstein, A. (2018). Moral hazard in health insurance: What we know and how we know it. *Journal of the European Economic Association*, 16(4):957–982.

Einav, L., Finkelstein, A., Ryan, S. P., Schrimpf, P., and Cullen, M. R. (2013). Selection on moral hazard in health insurance. *American Economic Review*.

Einav, L., Finkelstein, A., and Schrimpf, P. (2015). The response of drug expenditure to nonlinear contract design: evidence from medicare part d. *Quarterly Journal of Economics*.

Ellis, R. P. (1986). Rational behavior in the presence of coverage ceilings and deductibles. *The RAND Journal of Economics*.

Fang, H., Keane, M. P., and Silverman, D. (2008). Sources of advantageous selection: Evidence from the medigap insurance market. *Journal of political Economy*, 116(2):303–350.

Finkelstein, A. and McGarry, K. (2006). Multiple dimensions of private information: Evidence from the long-term care insurance market. *American Economic Review*.

Fox, J. T., il Kim, K., and Yang, C. (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. *Journal of Econometrics*, 195(2):236–254.

Fox, J. T., Kim, K. I., Ryan, S. P., and Bajari, P. (2011). A simple estimator for the distribution of random coefficients. *Quantitative Economics*, 2(3):381–418.

Ghili, S., Handel, B., Hendel, I., and Whinston, M. D. (2020). Optimal long-term health insurance contracts: Characterization, computation, and welfare effects. *working paper*.

Guo, A. and Zhang, J. (2019). What to expect when you are expecting: Are health care consumers forward-looking? *Journal of Health Economics*.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320.

Ho, K. and Lee, R. S. (2021). Health insurance menu design for large employers. *Unpublished*.

Hong, L. and Mommaerts, C. (2021). Time aggregation in health insurance deductibles. *working paper*.

Hua, L.-K. and Wang, Y. (2012). *Applications of number theory to numerical analysis*. Springer Science & Business Media.

Judd, K. L. (1998). *Numerical methods in economics*. MIT press.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*.

Keast, P. (1973). Optimal parameters for multidimensional integration. *SIAM Journal on Numerical Analysis*, 10(5):831–838.

Keeler, E. B., Newhouse, J. P., and Phelps, C. E. (1977a). Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty. *Econometrica*.

Keeler, E. B., Relles, D. A., and Rolph, J. E. (1977b). The choice between family and individual deductibles in health insurance policies. *Journal of Economic Theory*.

Keeler, E. B. and Rolph, J. E. (1988). The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*.

Korobov, A. (1959). The approximate computation of multiple integrals. In *Dokl. Akad. Nauk SSSR*, volume 124, pages 1207–1210.

Kowalski, A. E. (2015). Estimating the tradeoff between risk protection and moral hazard with a nonlinear budget set model of health insurance. *International Journal of Industrial Organization*.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer.

Lin, H. and Sacks, D. W. (2019). Intertemporal substitution in health care demand: Evidence from the rand health insurance experiment. *Journal of Public Economics*, 175:29–43.

Marone, V. R. and Sabety, A. (2021). Should there be vertical choice in health insurance markets? *Unpublished*.

Morris, C. (1979). A finite selection model for experimental design of the health insurance study. *Journal of Econometrics*.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. *Econometrica*, 84(2):411–443.

Newhouse, J. P. (1999). Rand health insurance experiment [in metropolitan and non-metropolitan areas of the united states], 1974-1982. *Inter-university Consortium for Political and Social Research*.

Newhouse, J. P. and The Insurance Experiment Group (1993). *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard Univeristy Press.

Niederreiter, H. (1992). *Random number generation and quasi-Monte Carlo methods*. SIAM.

Pauly, M. V. (1968). The economics of moral hazard: comment. *The american economic review*, 58(3):531–537.

Rossi, P. E., McCulloch, R. E., and Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340.

Train, K. E. (2008). Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69.

Vera-Hernández, M. (2003). Structural estimation of a principal-agent model: moral hazard in medical insurance. *RAND Journal of Economics*.

Appendix A Experimental enrollment dates by site

The RAND HIE defined a contract year as the 12-month period following each anniversary of the enrollment date. The staggered enrollment dates provide the variation needed to separate the effect of time trends and seasonal demand shocks on the timing of health care utilization decisions.

Table A1: Enrollment dates

Dayton	Seattle	Massachusetts	South Carolina
11-01-74	01-01-76	07-01-76	11-01-76 (5 yr.)
12-01-74	02-01-76	08-01-76	12-01-76 (5 yr.)
01-01-75	03-01-76	09-01-76	12-31-76 (5 yr.)
02-01-75	04-01-76	10-01-76	01-31-77 (5 yr.)
	05-01-76		11-01-78 (3 yr.)
	06-01-76		12-01-78 (3 yr.)
	07-01-76		01-01-79 (3 yr.)
	08-01-76		02-01-79 (3 yr.)
	09-01-76		

Note: The three-year and five-year groups enrolled at the same times in the locations above, and thus exited two years apart. In South Carolina, the three-year and five-year groups enrolled two years apart and exited at the same time.
Source: Table 4, Codebook 208, Newhouse (1999).

Appendix B Beginning- and end-of-experiment effects

Families enrolled in the RAND HIE were informed before they agree to participate that the experiment would end after either 3 or 5 years (randomly assigned before the start of the experiment). This might induce an increase in utilization during the last year of the experiment. Once the experiment is over, enrolled families would most probably return to the insurance plans they had before. Families assigned to full insurance during the experiment would be weakly worse off after the experiment ends, since their experimental plan was probably the most generous plan available in the market. For those enrolled in cost-sharing plans, there is some probability that their

before-the-experiment plans were less generous compared to their experimental plan, so they would potentially be worse off once the experiment ends. Similar reasoning applies to the first year of the experiment.

The presence of transitory effects on the demand of health care is not unique to RCTs. Using data from a large self-insured firm, Brot-Goldberg et al. (2017) find that consumer health care utilization ramps up at the end of the year after which a required plan shift from full insurance to a less generous plan took place. They use the term *anticipatory spending* to describe the extra health care utilization by consumers before the required plan switch actually occurred, when health care was cheaper. In the context of the RAND HIE, the presence of transitory effects has been overlooked until recently. To my knowledge, Lin and Sacks (2019) is the first to document graphically that health care utilization in the RAND HIE free-care plan ramps up over the last months of the experiment. In a concurrent work, Devereux et al. (2019) use the term *deadline effect* to describe a spike in health care utilization in the final year of the RAND HIE.

In my context, identifying the presence of beginning- and end-of-experiment effects is important for obtaining a true impact of dynamic moral hazard in health insurance contracts. Without recognizing the presence of an end-of-experiment effect, for example, my estimates of how weekly utilization changes as the end of the contract nears could potentially be positive, contradicting the implication from my model of forward-looking families, on average and ceteris paribus. To uncover the presence of these transitory effects, I introduce first and last contract year fixed effects to the empirical specification of Aron-Dine et al. (2013).

As a baseline, first consider their empirical framework:

$$y_{iq} = \lambda_p + \tau_t + \alpha_{lm} + \epsilon_{iq}, \tag{26}$$

which estimates a set of HIE plan p effects given by λ_p on several measures of health care utilization y for individual i in contract year q. Because plan assignment was only random conditional on location and enrollment month, the specification includes a full set of location l by start month m interactions, α_{lm}. Calendar year fixed effects, τ_t, account for any underlying time trend in the cost of health care. Because plans were assigned at the family rather than individual level, all regression results cluster the standard errors on the family.

Now define by $Term_i$ the enrollment term for individual i, with $Term_i \in \{3, 5\}$. I introduce two dummy variables to capture the beginning- and end-of-experiment effects in the first and last contract year, respectively, in the following equation:

$$y_{iq} = \lambda_p + \iota \times \mathbb{1}(q = 1) + \delta \times \mathbb{1}(q = Term_i) + \tau_t + \alpha_{lm} + \epsilon_{iq}, \tag{27}$$

where the parameters ι and δ capture the additional health care utilization in the first and last year of the experiment, respectively, compared to middle years. I limit the sample to non-attriters and exclude infants born during the experiment, who enter the experiment as part of the family unit. This ensures a balanced panel. I also exclude the first contract year from Dayton, Ohio, because some health services were treated differently.

Table A2 reports the results based on estimating equation (27) for various measures of health care utilization. In column 1, the dependent variable is the amount of annual health care utilization (in 2011 dollars). I fail to detect any transitory effect in health care utilization when I aggregate all health care categories. This is consistent with some early technical reports by the original RAND investigators. I find that it is necessary to disaggregate utilization by inpatient versus outpatient health care services in order to uncover a significant end-of-experiment effect. Column 3 shows that individuals increase outpatient health care utilization by $203 in their last year of the experiment, relative to middle years, on average. For this reason, I exclude from the estimation sample the last contract year.

Table A2: Beginning- and end-of-experiment effects

	(1) Total utilization in 2011 $	(2) Inpatient utilization in 2011 $	(3) Outpatient utilization in 2011 $
First Contract Year	-71.14	-159.12	87.03
	(150.60)	(127.80)	(69.68)
Last Contract Year	189.70	-13.38	203.27***
	(130.87)	(108.17)	(55.59)
Adjusted R^2	0.011	0.004	0.030
Site x enrol.	Y	Y	Y
Cal. years fe	Y	Y	Y
Family fe	N	N	N
Clustered se	Y	Y	Y
Families	2076	2076	2076
N	12535	12535	12535

Standard errors in parentheses

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

Notes: Table A2 reports selected least-squares coefficients estimates from equation (27). Standard errors, clustered on family, are in parentheses below the coefficients. All spending variables are inflation adjusted to 2011 dollars (adjusted using the CPI-U). Site by start month and calendar year dummy variables are de-meaned so that the coefficients reflect estimates for the "average" site-month-year mix.

Appendix C Model properties of the shadow price of care

Appendix D Econometric Details

D.1 Step 1: Solving the Model

As I describe in Section 5 of the main text, in the first step of the estimation algorithm, I solve the dynamic problem for a large number of types, once for each type, and store the optimal policy.

For a plan, k, and family type, h, I solve the finite-horizon dynamic program recursively. To do so, I discretize the C_t state to a grid of 1,000 points with spacing of size Δc dollars. Time is naturally discrete ($t = 1, 2, \ldots, 52$ over a contract year with $T = 52$ weeks) for my weekly data. These discretizations leave ν_t as the only continuous state variable. Because the family does not know ν_t prior to period t, I can integrate it out and the solution to the dynamic programming problem for each type of family can be characterized by the expected value functions, $E[V_{hkt}(C_{t-1})]$, and policy functions, $E[c^{\star}_{hkt}(C_{t-1})]$. To perform the numerical integration over the bounded support of ν_t, $[0, \overline{\nu}]$, I use adaptive Simpson quadrature.

Having solved the dynamic program for a family of type h, I generate the transition process for the state vector implied by the solution. The transition probabilities between the 52,000 possible states (1000 x 52) are implicitly defined by threshold values for ν_t. For example, consider a family of type h on plan k, that has consumed C_{t-1} prior to period t. The threshold, $\nu_t(z)$, is defined as the value of ν_t that makes a family indifferent between consuming z units of size Δc dollars and $z+1$ units, such that the marginal utility (net of any out-of-pocket expenditures) of an additional unit of consumption

$$u_h\big((z+1)\Delta c, y_t, \nu_t(z); k\big) - u_h\big(z\Delta c, y_t, \nu_t(z); k\big)$$

is equated to the loss in the net present value of future utility

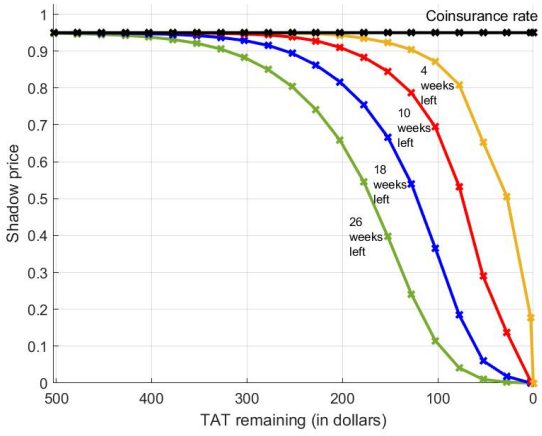
$$E\big[V_{hk(t+1)}\big(C_{t-1} + (z+1)\Delta c\big)\big] - E\big[V_{hk(t+1)}\big(C_{t-1} + z\Delta c\big)\big].$$

These thresholds, along with all families' initial condition ($C_0 = 0$), define the transition process between states. For each family type h and plan k, I characterize this transition process by the CDF of cumulative health care consumption that it generates,

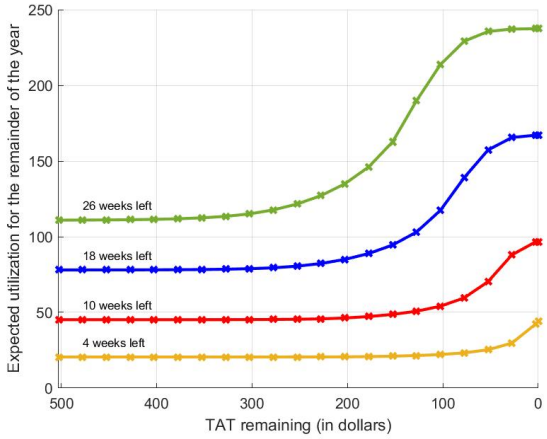
$$\Gamma_{hkt}(C) = \text{Prob}(C_{t-1} < C),$$

the proportion of families that have consumed less than C through period t of the contract year.

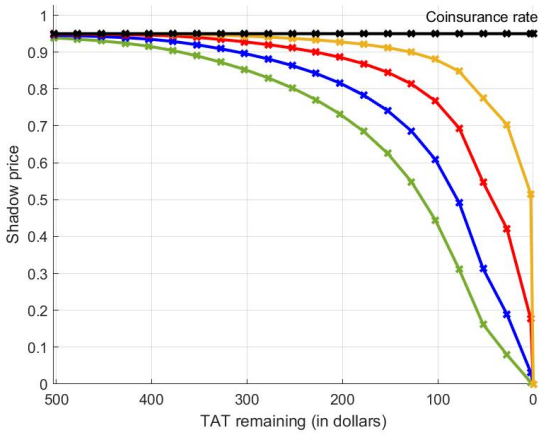
Figure 14: Model properties of the shadow price of care



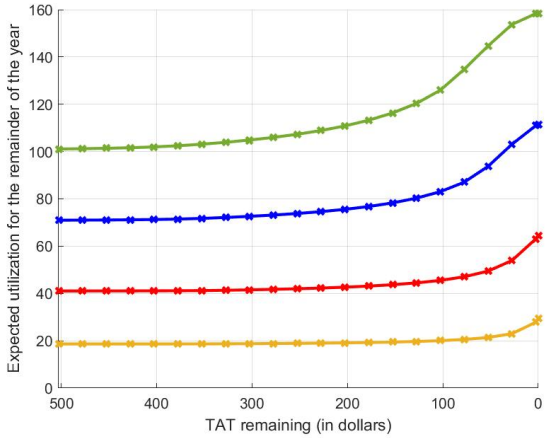
(a) Shadow price, low variance of F_ν, low ω



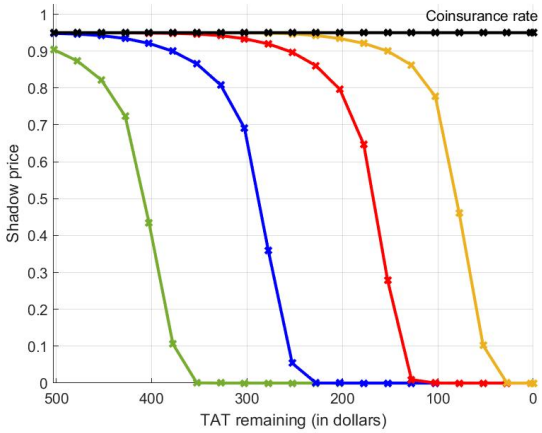
(b) Expected future demand, low variance of F_ν, low ω



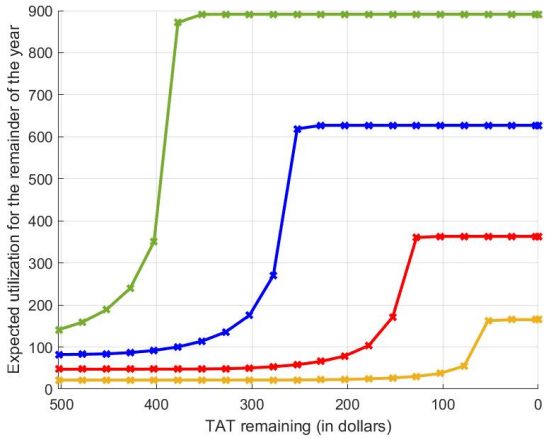
(c) Shadow price, high variance of F_ν



(d) Expected future demand, high variance of F_ν



(e) Shadow price, high ω



(f) Expected future demand, high ω

Due to the discretized state space, $\Gamma_{hkt}(C)$ is a step function.

D.2 Step 2: Estimation

The second step of my estimation approach matches empirical moments I recover from the data to those predicted by my model by choosing weights for each family type.

As I describe in Section 5 in the main text, my estimates of the weights are chosen to maximize the objective function. I set the weighting matrix $\hat{V}^{-1}$ equal to the identity matrix.

To recover the cumulative distribution of C_{t-1} for each contract week t and plan k, I use a smooth version of a simple Kaplan-Meier estimator,

$$\widehat{\Gamma}_{kt}(C) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{1}[C_{i(t-1)} < C],$$

where N_k denotes the number of family-years in plan k and C represents each of the points of the discretized state C_{t-1}. I estimate these moments for each k and t, considering values of C such that $\widehat{\Gamma}_{kt}(C) \in [0, 1]$, ensuring that I fit the tails of the annual health care utilization distribution.

I recover the moments of health care utilization at each state by estimating a smooth surface using a nearest-neighbor approach. Consider a point in the state space, (C_{t-1}, t). A neighbor is an observation in the data for which the family is t weeks into the contract year and cumulative health care utilization up until contract week t is within five percent of C_{t-1}. Denote the number of neighbors by $N_{kt}(C_{t-1})$. Then, I estimate the conditional (on reaching the state) mean at (C_{t-1}, t) using

$$\widehat{E}[c^{\star}_{kt}(C_{t-1})] = \frac{1}{N_{kt}(C_{t-1})} \sum_{i=1}^{N_{kt}(C_{t-1})} c_i,$$

where $i \in \{1, \ldots, N_{kt}(C_{t-1})\}$ indexes the set of nearest neighbors. If $N_{kt}(C_{t-1}) > 500$, I use those 500 neighbors nearest to C_{t-1}. Note that this gives me the average expenditure conditional on a family arriving at the state. To recover the unconditional mean, I multiply $\widehat{E}[c^{\star}_{kt}(C_{t-1})]$ by the probability of observing a family at state (C_{t-1}, t), recovered from the estimated CDF of cumulative expenditure.

I estimate both moments at the same set of state space points used when numerically solving the dynamic programming problem for each family type. This results in 104,000 moments for each plan of the 10 plans, or 10 x 104,000 = 1,040,000 moments in total.

D.3 The choice of moments

Following Nevo et al. (2016), I recover the first set of moments at each state by estimating a smooth surface using a nearest-neighbor approach. Consider a point in the state space, (C_{t-1}, t). A neighbor is an observation in the data for which the family is t weeks into the contract year and cumulative health care consumption up until week t is within five percent of C_{t-1}. Denote the number of neighbors by $NN_{kt}(C_{t-1})$. Then, I estimate the unconditional (on reaching the state) mean at (C_{t-1}, t) using

$$\widehat{m}^{\text{dat}}_{k,1}(C,t) = \frac{\text{Prob}\big(C_{hk(t-1)} = C, t\big)}{N\big(C^{\text{dat}}_{k(t-1)} = C, t\big)} \sum_{i=1}^{N\left(C^{\text{dat}}_{k(t-1)}=C,t\right)} c^{\text{dat}}_{ikt} \,.$$

D.4 Bootstrap Procedure for Inference

As described in Lahiri (2003), the basic idea behind the bootstrap method is to recreate the relation between the population and the sample using the sample itself. For dependent data, the most common approach to this problem is to resample "blocks" of observations instead of single observations, which preserves the dependence structure of the underlying process *within* the resampled blocks.

My block-bootstrap estimator proceeds with the following repeated procedure:

1. I first draw an alternative dataset sampling with replacement at the family-year level. Specifically, I sample the data by family-year with replacement, keeping all 52 weeks for each family-year drawn. The new dataset has the same number of family-years as the original data.

2. I then use this new dataset to recalculate the proportion of family-years in each plan. These plan weights will be useful for recovering the overall distribution of heterogeneity.

3. For each plan $k = 1, \ldots, 10$ separately, I recalculate the inputs to the moments, $\boldsymbol{m}^{\text{dat}}_k$, and then re-estimate the structural parameters of my model, i.e., the plan-specific weights $\widehat{\boldsymbol{\theta}}_k$.

4. Finally, I calculate the overall type distribution by weighting each $\widehat{\boldsymbol{\theta}}_k$ with the corresponding proportion of family-years in plan k and summing across plans.

For a given type h, the bootstrap confidence interval with $1-\alpha$ coverage can be constructed as

$$CI_{(1-\alpha)\%} = \left[\widehat{\theta}_h - q^{\star}_h\left(1-\frac{\alpha}{2}\right), \widehat{\theta}_h - q^{\star}_h\left(\frac{\alpha}{2}\right)\right], \tag{28}$$

where $q_h^\star$ is the quantile function of $\widehat{\theta}_h^\star - \widehat{\theta}_h$.

I report results from 1,000 bootstrap draws.

Appendix E Details about the construction of the grid of family types

I use the *method of good lattice points* to generate a finite collection of points in the five-dimensional space of type heterogeneity. This method was proposed by Korobov (1959) for numerical evaluation of multivariate integrals. The basic idea of a quasi-Monte Carlo method is to replace random samples in a Monte Carlo method by well-chosen deterministic points. The criterion for the choice of deterministic points depends on the numerical problem at hand. For the important problem of numerical integration, the selection criterion is easy to find and leads to the concepts of uniformly distributed sequence and discrepancy. The discrepancy can be viewed as a quantitative measure for the deviation from uniform distribution.

E.1 Discrepancy

The concept of *discrepancy* provides a measure of how dispersed a collection of points is. Let $x_j \in I \equiv [0,1]$, $j = 1, \ldots, N$, be a sequence of scalars. If $S \subset I$, define the cardinality of a set X in a set S, $\text{card}(S \cap X)$, to be the number of elements of X which are also in S. Following Judd (1998), I define a notion of discrepancy for finite sets.

Definition E.1 *(Niederreiter, 1992) The* ***star discrepancy*** $D_N^\star$ *of the set* $X \equiv \{x_1, x_2, \ldots, x_N\} \subset [0,1]$ *is*

$$D_N^\star(X) = \sup_{0 \le t \le 1} \left| \frac{card([0,t) \cap X)}{N} - t \right|.$$

Note that $0 \le D_N^\star(X) \le 1$ always. This definition allow us to measure the deviations from uniformity of sets. Even though a continuum of intervals is used in the definition, we need only to check open intervals of the form $(0, x_l)$, $1 \le l \le N$. In the one-dimensional case, a simple explicit formula for $D_N^\star(X)$ can be given.

Theorem E.1 *(Niederreiter, 1992) If* $0 \le x_1 \le x_2 \le \cdots \le x_N \le 1$, *then*

$$D_N^\star(X) = \frac{1}{2N} + \max_{1 \le n \le N} \left| x_n - \frac{2n-1}{2N} \right|.$$

Proof. See Theorem 2.6 in Niederreiter (1992).

It follows from the theorem that we always have $D_N^\star(X) \ge 1/(2N)$, and equality holds if $x_n = (2n-1)/(2N)$ for $1 \le n \le N$. This implies that in the one-dimensional case, the minimum of

the star discrepancy $D_N^\star(X)$ is $1/(2N)$ and the classical N-panel midpoint rule for the interval $[0,1]$ achieves this bound. Thus, for low-discrepancy sets in the one-dimensional case, quasi-Monte Carlo methods are not so important. Below I consider the concept of star discrepancy in the multidimensional case.

Definition E.2 *(Niederreiter, 1992) The* ***star discrepancy*** $D_N^\star$ *of the set* $X \equiv \{x_1, x_2, \ldots, x_N\} \subset I^d$ *is*

$$D_N^\star(X) = \sup_{0 \le t_1, \ldots, t_d \le 1} \left| \frac{card([0,t_1) \times \cdots \times [0,t_d) \cap X)}{N} - \prod_{j=1}^{d} t_j \right|,$$

where I^d *is the closed d-dimensional unit cube.*

A small discrepancy says that the set evenly fills up the hypercube I^d. A set of points X consisting of N elements of I^d is called a low-discrepancy set if $D_N^\star(X)$ is small.

E.2 The method of good lattice points

The good lattice point method was proposed by Korobov (1959) for numerical evaluation of multivariate integrals. The method of good lattice points begins with an integer N and a vector of good lattice points $g \in \{0, 1, \ldots, N-1\}^d$, forms the finite collection of points

$$x_l = \left\{ \frac{l}{N} g \right\}, \quad l = 1, \ldots, N, \tag{29}$$

and computes the quasi-Monte Carlo approximation

$$\int_{I^d} f(x)dx \doteq \frac{1}{N} \sum_{l=1}^{N} f(x_l), \tag{30}$$

where the expression $\{z\}$ denotes the *fractional part of* z.[61] The task is to find combinations of N and g such that the approximation in equation (30) is good. In other words, we want to minimize the error in the approximation of equation (30), which can be written as

$$R = \left| \int_{I^d} f(x)dx - \frac{1}{N} \sum_{l=1}^{N} f(x_l) \right|$$

The value of R is closely related to the star discrepancy $D_N^\star(X)$, if $F(x)$ satisfies certain conditions. The error analysis for quasi-Monte Carlo integration in Niederreiter (1992) has demonstrated that small errors are guaranteed if sets with small star discrepancy are used. From the view point of

[61] The fractional part of z is formally defined by $\{z\} \equiv z - \max\{k \in \mathbb{Z} | k \le z\}$.

numerical analysis, we demand not only the star discrepancy of X should be low but also the set of points X should be convenient for computation.

Good choices of N and g are difficult to compute (Judd, 1998). A strategy pursued by Korobov and others is to examine lattice points that are simply generated and evaluate their performance in integrating certain test functions with known integrals. One test function that is particularly valuable is

$$F(x) = \prod_{j=1}^{d} \left(1 - \frac{\pi^2}{6} + \frac{\pi^2}{2}(1 - 2\{x_j\})^2\right),$$

which is defined on I^d and integrate to 1. Note that F is the function in the class of functions having Fourier series in I^d whose Fourier coefficients converge at the slowest possible rate. Korobov (1959) proposes an algorithm for finding lattice points by minimizing R for the function $F(x)$. Keast (1973) extends this algorithm and proves that the lattice points obtained are optimal in Korobov's sense. He first chooses J distinct primes, p_j, $j = 1, \ldots, J$, and lets their product p be the sample size N in equation (29). He then chooses a sequence of integers a_j, $j = 1, \ldots, J$. First, a_1 is chosen to minimize

$$H_1(a) \equiv \frac{3^d}{p_1} \sum_{k=1}^{p_1} \prod_{j=1}^{d} \left(1 - 2\left\{k\frac{a^{j-1}}{p}\right\}\right)^2$$

over $a \in \{1, \ldots, p_1 - 1\}$. More generally, for $l = 1, \ldots, J$, a_l minimizes

$$H_l(a) \equiv \frac{3^d}{p_1 \ldots p_l} \sum_{k=1}^{p_1 \ldots p_l} \prod_{j=1}^{d} \left(1 - 2\left\{k\left(\frac{a_1^{j-1}}{p_1} + \cdots + \frac{a_{l-1}^{j-1}}{p_{l-1}} + \frac{a^{j-1}}{p}\right)\right\}\right)^2$$

for $a \in \{1, \ldots, p_l - 1\}$. The *Keast good lattice point* g is then defined to be

$$g_j = \sum_{l=1}^{J} \frac{p}{p_l} a_l^{j-1}, \quad j = 1, \ldots, d.$$

As pointed out in Judd (1998), there is no assurance that the approximations are monotonically better as we increase p, the number of points. Therefore, in constructing a sequence of lattice formulas, one should keep only those formulas that do better in integrating $F(x)$ than formulas with fewer points. $H_J(g)$ serves as a performance index to rank various lattice point rules. Fortunately, there exist tables of good lattice points, g, for specific sample sizes N and dimensions d. The good lattice points have the form $g^d = (1, g_2^d, \ldots, g_d^d) \in \mathbb{R}^d$.[62]

[62] See e.g., Table 9.4 in Judd (1998), Tables 1 and 2 in Bourdeau and Pitre (1985), and Hua and Wang (2012).

E.3 Implementation

I use the method of good lattice points to generate a finite collection of points in the five-dimensional space of type heterogeneity. Following Table 9.4 in Judd (1998) which was made according to Keast's method, I fix $N = 1,069$ and $d = 5$, and find the vector of good lattice points $g = (1, 63, 762, 970, 177)$. Using equation (29), I then construct the grid of 1,069 points within the $[0, 1)$ hypercube of dimension five. The last step is to redefine the bounds of the $[0, 1)$ hypercube to capture the support of each dimension of heterogeneity.

Appendix F *glp* method versus *tpp* method: Monte Carlo Evidence

Up until now I have only highlighted one of the main advantages of the *glp* method: better coverage of the parameter space. However, there is another benefit of using *glp* versus *tpp*: computational efficiency. Next, I perform a small Monte Carlo exercise to illustrate the gains in computational time, without any loss in precision. To do that, I compare the performance of the *good-lattice-points* grid versus the *tensor-product-points* grid within the context of this paper. I fix the number of family-years to 300, and use $M = 100$ replications. I generate data using two alternative distributions $F(\boldsymbol{\beta})$ for the random coefficients. In the first design, the true distribution has the first three characteristics (i.e., μ, σ and κ) heterogeneous across types but correlated within type, and the remaining two components of the type-space are homogeneous across all types. In the second design, all five characteristics of the true distribution are heterogeneous across types and uncorrelated within type. In both designs the underlying true CDF has continuous support.

I use health care consumption data at the family-week-year level, where the true data generating process is the dynamic stochastic model in Section 4. For each fake data set, I compute the moments in Section 5.2 and estimate a type distribution $\widehat{F}(\boldsymbol{\beta})$ by matching the moments I recover from the data to the weighted average of the behavior predicted by the model. For each run, after I compute the estimate $\widehat{F}(\boldsymbol{\beta})$, I evaluate its squared difference from the true distribution function $F_0(\boldsymbol{\beta})$ at $S = 10,000$ points uniformly spaced. I use root mean integrated squared error (RMISE) to assess performance of both estimators. My definition of RMISE for an estimator $\widehat{F}$ is

$$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{S}\sum_{s=1}^{S}\left(\widehat{F}_m(\boldsymbol{\beta}_s) - F_0(\boldsymbol{\beta}_s)\right)^2\right]}, \tag{31}$$

where I use $M = 100$ replications, each with a new fake data set. I also report the integrated absolute error (IAE), which for a given replication m is

$$\frac{1}{S}\sum_{s=1}^{S}\left|\widehat{F}_m(\boldsymbol{\beta}_s) - F_0(\boldsymbol{\beta}_s)\right|. \tag{32}$$

This is a measure of the mean absolute value of the estimation error, taken across the points of evaluation for a given replication. I compute the mean, minimum, and maximum IAE's across the M replications.

The results are given in Table A3. The first column reports the sample size N, which refers to the number of family-years used for the simulation. The second column describes the method used to populate the grid of types. Whenever the *tpp* method is used, column 2 also provides details about how many points in each dimension were used. The third column reports the number H of types (or basis points) used in the estimation. The next column reports the RMISE of the estimated distribution functions. The following three columns report the mean, minimum, and maximum of the IAE. The final three columns report the mean, minimum, and maximum of the number of types that have positive weight.[63]

While performance of the *tpp* method generally increases with the number of types, it is worth noting that the fit can decrease with increases in H, as the *tpp* grids do not necessarily nest each other for marginal increases in H. In the case where one *tpp* grid is nested inside another one, the RMISE measure should decrease with the number of types R. One example of this can be noted in the first design (i.e., 3D correlated), where the *tpp* grid with 5 points per dimension is nested inside the *tpp* grid with 9 points per dimension.

In the context of correlated random coefficients, the *glp* method provides more flexibility to pick the grid points, as opposed to the *tpp* method. This appealing feature should help in capturing the true underlying distribution more accurately. The results in Panel A of Table A3 suggest that the *glp* grid exhibits much better performance than the *tpp* grid, even with 10 times less points in the grid. By and large, RMISE and IAE are lower in the *glp* design than in the *tpp* designs. The RMISE of the *glp* grid with 101 points is 0.1066, while the RMISE of the *tpp* grid with 1000 points is 22 percent higher. With only 42.8 grid points with positive mass (on average), the *glp* grid does an excellent job compared to the 170.87 grid points with positive mass (on average) of the *tpp* grid with 10 points per dimension.

[63] A type is considered to have a positive weight if the estimated weight is greater than or equal to 0.01 percent.

Table A3: Monte Carlo results: 3D correlated and 5D uncorrelated

				Integrated Absolute Error			No. of Positive Weights		
N	Method	R	RMISE	Mean	Min	Max	Mean	Min	Max
				Panel A: 3D correlated					
	glp	101	0.1066	0.0538	0.0354	0.0681	42.80	33	53
	tpp	5^3 =125	0.1414	0.0723	0.0586	0.0877	47.11	17	66
	tpp	6^3 =216	0.1292	0.0657	0.0487	0.0868	71.90	21	98
	tpp	7^3 =343	0.1277	0.0655	0.0427	0.0880	89.22	18	147
	tpp	8^3 =512	0.1347	0.0694	0.0409	0.0888	121.83	21	208
	tpp	9^3 =729	0.1338	0.0676	0.0427	0.0954	135.91	21	276
	tpp	10^3 =1000	0.1299	0.0661	0.0441	0.0948	170.87	20	352
	tpp	11^3 =1331	0.1301	0.0655	0.0416	0.0950	203.20	20	455
	tpp	12^3 =1728	0.1303	0.0661	0.0421	0.0910	240.53	21	582
	tpp	13^3 =2197	0.1307	0.0652	0.0423	0.0937	283.48	21	729
300	tpp	14^3 =2744	0.1311	0.0653	0.0469	0.0899	309.30	19	649
	tpp	15^3 =3375	0.1264	0.0635	0.0385	0.0939	409.96	21	785
	tpp	16^3 =4096	0.1253	0.0638	0.0461	0.0895	454.66	21	849
	tpp	17^3 =4913	0.1236	0.0629	0.0433	0.0901	560.74	23	1093
	tpp	18^3 =5832	0.1164	0.0587	0.0393	0.0890	601.30	22	1264
	tpp	19^3 =6859	0.1225	0.0628	0.0383	0.0896	725.76	18	1496
	tpp	20^3 =8000	0.1186	0.0605	0.0439	0.0884	871.84	26	1765
	tpp	21^3 =9261	0.1201	0.0611	0.0400	0.0897	898.19	23	1855
	tpp	22^3 =10648	0.1193	0.0607	0.0400	0.0898	1001.6	21	2246
	tpp	23^3 =12167	0.1260	0.0640	0.0422	0.0916	1101.8	25	2434
	tpp	24^3 =13824	0.1206	0.0621	0.0439	0.0867	1031.1	25	2670
				Panel B: 5D uncorrelated					
	glp	1069	0.0866	0.0400	0.0285	0.0651	119.77	16	323
300	tpp	5^5 =3125	0.0919	0.0484	0.0351	0.0877	329.74	35	567
	tpp	6^5 =7776	0.0849	0.0442	0.0273	0.0761	715.57	41	1227
	tpp	7^5 =16807	0.0873	0.0462	0.0313	0.0915	1143.70	69	2472

Appendix G Decomposition of Relative Social Surplus

The certainty equivalent to a contract j at premium $premium_j$ for a type-h family with initial income Y is given by $e_{hj}(\boldsymbol{\beta}_h)$, as defined in equation (21) and repeated here:

$$e_{hj}(\boldsymbol{\beta}_h) \equiv -\frac{1}{\psi} \ln \left[\int \exp(-\psi\, \tilde{u}^{\star}(\boldsymbol{\nu}, \boldsymbol{\beta}_h, j)) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) \right] + (Y - premium_j).$$

The certainty equivalent can also be expressed as

$$\begin{aligned} e_{hj}(\boldsymbol{\beta}_h) &= EV(j,h) - \frac{1}{\psi}\ln\left[\int \exp(-\psi\,\tilde{u}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j))dF_{\boldsymbol{\nu}}(\boldsymbol{\nu})\right] + (Y - premium_j) - EV(j,h) \\ &= EV(j,h) + Y - premium_j - RP(j,h), \end{aligned}$$

where $EV(j,h)+Y-premium_j$ is the expected payoff and $RP(j,h)$ is the risk premium associated with the lottery. In particular,

$$\begin{aligned} EV(j,h) &= \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T_j}((c_t^{\star}-\nu_t) - \frac{1}{2\,\omega}(c_t^{\star}-\nu_t)^2 - \mathcal{O}(c_t^{\star}, C_{t-1}; j))\right], \quad \text{and} \\ RP(j,h) &= EV(j,h) + \frac{1}{\psi}\ln\left[\int \exp(-\psi\,\tilde{u}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j))dF_{\boldsymbol{\nu}}(\boldsymbol{\nu})\right] \end{aligned} \tag{33}$$

The corresponding insurance provider's welfare for a type-h family enrolled in contract j is given by his expected profits, as defined in equation (22) and repeated here:

$$\pi_{hj}(\boldsymbol{\beta}_h) \equiv premium_j - \mathbb{E}_{\boldsymbol{\nu}}\Big[k_j\big(C_{T_j}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big)\Big],$$

where $k_j(.)$ is the function that maps family's total health care utilization to the portion covered by the provider under the price scheme of contract j.

The social surplus generated by allocating a type-h family to contract j is given by

$$SS_{hj}(\boldsymbol{\beta}_h) = e_{hj}(\boldsymbol{\beta}_h) + \pi_{hj}(\boldsymbol{\beta}_h)$$

Finally, the *relative* social surplus generated by allocating a type-h family to contract j (relative to allocating the same family to the free-care contract) is given by RSS_{hj}, as defined in equation (23) and repeated here:

$$RSS_{hj}(\boldsymbol{\beta}_h) \equiv SS_{hj}(\boldsymbol{\beta}_h) - SS_{h,\text{free}}(\boldsymbol{\beta}_h)$$

The relative social surplus can also be expressed as

$$\begin{aligned} RSS_{hj}(\boldsymbol{\beta}_h) &= \tilde{e}_{hj}(\boldsymbol{\beta}_h) - \tilde{e}_{h,\text{free}}(\boldsymbol{\beta}_h) - \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C_{T_j}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big) - C_{T_j}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,\text{free})] \\ &= [EV(j,h) - RP(j,h)] - [EV(\text{free},h) - RP(\text{free},h)] - \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C_{T_j}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big) - C_{T_j}^{\star}(\boldsymbol{\nu},\boldsymbol{\beta}_h,\text{free})] \ldots \\ &\quad + \big[EV^{\text{myopic}}(j,h) - \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C_{T_j}^{\star,\text{myopic}}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big)\big] \ldots \\ &\quad - \big[EV^{\text{myopic}}(j,h) - \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C_{T_j}^{\star,\text{myopic}}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big)\big]\big] \\ &= \underbrace{\Psi(j,h)}_{\substack{\text{Relative value of} \\ \text{risk protection}}} - \Big[\underbrace{\Sigma(j,h)}_{\substack{\text{Relative social cost of} \\ \text{static moral hazard}}} + \underbrace{\Delta(j,h)}_{\substack{\text{Relative social cost of} \\ \text{dynamic moral hazard}}}\Big], \end{aligned}$$

where

$$\begin{aligned}
\Psi(j,h) &= RP(\text{free},h) - RP(j,h) && \leq 0,\\
\Sigma(j,h) &= \big[-EV^{\text{myopic}}(j,h) + \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C^{\star,\text{myopic}}_{T_j}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big)]\big] + \big[EV(\text{free},h) - \mathbb{E}_{\boldsymbol{\nu}}[C^{\star}_{T_j}(\boldsymbol{\nu},\boldsymbol{\beta}_h,\text{free})]\big] && \leq 0,\\
\Delta(j,h) &= \big[-EV(j,h) + \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C^{\star}_{T_j}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big)]\big] + \big[EV^{\text{myopic}}(j,h) - \mathbb{E}_{\boldsymbol{\nu}}[k_j\big(C^{\star,\text{myopic}}_{T_j}(\boldsymbol{\nu},\boldsymbol{\beta}_h,j)\big)]\big] && \geq 0.
\end{aligned}$$

The (relative) value of risk protection, $\Psi(j,h)$, is non-positive because any contract j provides a weakly riskier distribution of payoffs than the free-care contract.

Appendix H Details about the RAND HIE design and the construction of the analysis sample

Approximately 2,500 nonelderly families (or 7,700 individuals) were assigned to one of 14 fee-for-service (FFS) insurance plans or to a prepaid group practice. The fee-for-service plans varied along two principal dimensions: the coinsurance rate (the fraction of billed charges paid by the participant) and the maximum dollar expenditure (MDE), a cap on family out-of-pocket expenditures during a 12-month accounting period. The design used four coinsurance percentages (0, 25, 50, and 95) and three levels of MDE (5, 10, or 15 percent of family income, up to a maximum of $1,000). In one exceptional plan the MDE was set at $150 per person or $450 per family. These various coinsurance and MDE rates were combined as follows:

- **FFS plan 1:** one plan with zero coinsurance (free care).
- **FFS plans 2 to 4:** three plans with 25 percent coinsurance and MDEs of 5, 10, or 15 percent of family income or $1,000, whichever was less.
- **FFS plans 5 to 7:** three plans with 50 percent coinsurance and MDEs of 5, 10, or 15 percent of family income or $1,000, whichever was less.
- **FFS plans 8 to 10:** three plans with 50 percent coinsurance and MDEs of 5, 10, or 15 percent of family income or $1,000, whichever was less.
- **FFS mixed plans 11 to 13:** three plans with 25 percent coinsurance for all services except outpatient mental health and dental, which were subject to 50 percent coinsurance; and MDEs of 5, 10, or 15 percent of family income or $1,000, whichever was less.
- **FFS mixed plan 14:** one plan with 95 percent coinsurance for outpatient services and 0 percent coinsurance (free care) for inpatient services and a MDE of $150 per person, subject to a maximum of $450 per family.

- **Prepaid group practice plan 15:** one plan with 0 percent coinsurance (free care) if care was received at a Seattle Health Maintainance Organization (HMO), Group Health Cooperative of Puget Sound; 95 percent coinsurance if care was received outside the HMO.

I make four restrictions to create my baseline sample.

1. My model does not distinguish between providers of service (e.g., physician versus dentist) or whether the provider belongs to the prepaid group network.
2. Dental and mental health services were treated differently in the first year of the experiment in Dayton, Ohio. Dental services for adults were covered only on the free-care plan (dental services for children were covered on all plans). Outpatient mental services were not covered.
3. For any family-year observation in the free-care plan with missing MDE, I imputed a MDE equal to zero.

The number of families at enrollment does not necessarily coincide with the number of families that completed the experiment, even in the absence of attrition (see Footnote 11). Indeed, absent attrition, the number of families at enrollment is the lower bound for the number of families that completed the experiment.