Overview

- What is moral hazard and why do we care about it?
- Moral hazard in health insurance: what do we know and how do we know it?
  - Definition / Intellectual history
  - Existence - value of compelling descriptive work
  - Implications for spending under alternative health insurance contracts
Methodological themes

- Reduced form evidence as complement (not substitutes) for "structural modeling" - and vice versa
- Value (and limitations of) reduced form work
- Uses (and abuses) of structural estimation
- Importance of modeling choices: a given reduced form object can have very different out-of-sample implications depending on the model
What is moral hazard and why do we care about it?

- Moral hazard: hidden action (vs selection - hidden information)
  - Classic example: unobserved search effort to find a job
- Unemployment insurance may affect (unobserved) search effort
- Moral hazard prevents us from achieving first best of full insurance (full consumption smoothing)
  - (Constrained) optimum insurance level trades of benefits from insurance (consumption smoothing) with costs (moral hazard)
  - Will study in depth in next unit
- NB: Unlike with selection, government typically does not have a comparative advantage in addressing moral hazard
What is moral hazard in health insurance?

- In health insurance “moral hazard” typically refers to impact of health insurance on medical spending (i.e. “price elasticity of demand for medical care”)
  - Hidden information: how much medical care I would consume if I faced the full marginal cost of the medical care consumption
  - What is the hidden action?

- Intellectual history:
  - Arrow (1963 AER) (first?) use of moral hazard to mean “medical insurance increases demand for medical care”
  - Pauly (1968 AER): (first?) explicit use of term to refer to impact of insurance on medical care via its effect on reducing the price
Moral hazard in health insurance (con’t)

- Two distinct phenomena referred to as “moral hazard” in health insurance context
- “Ex ante moral hazard”: I invest less effort in my health (smoking, drinking, exercise) and so my health worsens (Ehrlich and Becker 1972)
  - Seems a compelling example of “hidden action”
- “Ex post moral hazard”: Conditional on my health, I consume more medical care (Pauly 1968)
  - Focus of empirical literature
  - What’s the hidden action?
Ex Ante Moral Hazard

“Ex ante moral hazard”: I invest less effort in my health (smoking, drinking, exercise) and so my health worsens (Ehrlich and Becker 1972)

Little empirical work / evidence

Spenkuch (JHE 2012)

- RCT in Mexico (geographic level) - Seguro Popular (King et al. 2009)
- Re-analyzes, grouping for power and finds some evidence of declines in preventive care (flu shots, mammograms etc)
- Question of interpretation: ex ante mh or congestion? [Cleaner tests?]

Note: even “full” health insurance only insures medical expenses (not health)
“Ex post” moral hazard

“Ex post moral hazard”: Conditional on my health, I consume more medical care (Pauly 1968)

What is the hidden action?

- You or your physician make less effort to shop around for a good price (Arrow 1963)
- Lower price → consume higher quantity of care (Pauly 1968).
  - How is this hidden action? Cutler and Zeckhauser (2000): Action not hidden but motivation is (re-writing hidden info problems as hidden action problems... Milgrom 1979)

From now on will focus on “ex post” moral hazard: response of health care spending ($P*Q$) or health care use to consumer price

- Only recently have we started to try to disentangle role of "price shopping" (i.e. $P$ a la Arrow) - Brot-Golberg et al. (QJE 2017) find no evidence that high deductibles encourage search for lower prices
Outline of these lectures

1. Existence of moral hazard in health insurance:
   1. RAND HIE
   2. Oregon HIE

2. Implications for spending under alternative contracts:
   1. Selection on moral hazard (Einav et al., 2013) - why we might not want random plan assignment
   2. Response to non-linear pricing - why we need models to know how to use reduced form estimates for counterfactuals
      1. Einav et al. (2015; 2017)
The moral hazard argument makes sense... only if we consume health care in the same way that we consume other consumer goods, and to economists like [John] Nyman this assumption is plainly absurd. We go to the doctor grudgingly, only because we’re sick. “Moral hazard is overblown,” the Princeton economist Uwe Reinhardt says. “You always hear that the demand for health care is unlimited. This is just not true... Do people really like to go to the doctor? Do they check into the hospital instead of playing golf?”

The dream of a free lunch

In contrast to the moral hazard theory that health insurance will increase health care spending, there is also a view out there that health insurance will *decrease* health care spending.

How?

- Reduce inappropriate and inefficient use of (expensive) emergency rooms
- Improve health and therefore reduce health care use
Ultimately this is an empirical question

- Empirical challenge: people w/ and w/o insurance differ for reasons that may be related to the outcome of interest (i.e. health care use)
  - In particular, recall adverse selection: sick select into market
- Inference issue: separating selection from treatment (moral hazard)
- Arguably no better way to convincingly test the null of no moral hazard than with a randomized experiment
  - Randomly assign insurance across individuals
  - By construction, the insured and uninsured are on average identical except for whether or not they have insurance.
- Three RCTs in US health insurance
  - RAND Health Insurance Experiment (1970s)
  - Oregon’s 2008 Medicaid lottery (www.nber.org/oregon)
  - AB Demonstration project (Michalapoulous et al. 2011)
Rand HIE: Overview

- 1974-1981: Randomize approx 2,000 families (5,800 people) into plans with different consumer cost-sharing
  - Conducted at 6 different locations across the US
  - Designed to be representative of families w/ adults under age 62
  - Assigned to experiment for 3 to 5 years
- Designed to study effects of consumer cost sharing on health care spending and health
- Pioneering – One of the earliest RCTs in the US
  - PI: Joe Newhouse
  - To date, still one of the largest (~$300 million in 2011 $)
Presenting the RAND results

- Follow Aron-Dine, Einav and Finkelstein (JEP, 2013) attempt to re-present and re-examine core findings from a “modern” perspective
  - Focusing only on the spending results (experiment also looks at health effects)
- Two key issues in design and analysis of RCTs
  - Threats to validity
  - Translating experimental treatment effects to economic objects of interest
For more on RAND

- Classic reference in economics: Manning et al. (1987) AER summary article
Rand HIE – basic design

- Main feature: randomly assign families to plans with different consumer cost-sharing
  - ranging from full coverage (free care plan) to a plan with almost no coverage for first ~$4,000 (in 2011 $) incurred during year

- 6 main plans. Coinsurance:
  - 0% (free plan); 25%; 50%; 95%
  - Mixed plans:
    - 25% except mental and dental (50%)
    - "Individual deductible plan": 95% outpatient, 0% inpatient.

- To limit participants’ financial exposure, randomly assigned (w/in coinsurance rates) to plans with different Maximum Dollar Expenditure (MDE) for years
  - Typically 5, 10 or 15% of income up to max of ~$3,000 - $4,000 in 2011 dollars
  - One average, ~1/3 of families hit stop loss within year
  - Interpretation implications: Variation not over catastrophic coverage
Not simple random assignment

Within a site and enrollment month assigned by stratified random sampling designed to achieve better balance across a set of baseline covariates than would likely be achieved by chance alone.

Data on medical use and medical spending come from claims filed by participants with the experiment.

- During the duration of the experiment the experiment acts as your insurer.
- To be reimbursed, one needs to file claims.
- Claims provide detailed information on health care use and spending.
### Plan summary statistics

**Table 1**

Plan Summary Statistics and Refusal and Attrition Rates

<table>
<thead>
<tr>
<th>Plan</th>
<th>Individuals (Families)</th>
<th>Avg. out-of-pocket share$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(1)</strong></td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Free Care</td>
<td>1,894 (626)</td>
<td>0%</td>
</tr>
<tr>
<td>25% Coinsurance</td>
<td>647 (224)</td>
<td>23%</td>
</tr>
<tr>
<td>Mixed Coinsurance$^a$</td>
<td>490 (172)</td>
<td>28%</td>
</tr>
<tr>
<td>50% Coinsurance</td>
<td>383 (130)</td>
<td>44%</td>
</tr>
<tr>
<td>Individual Deductible$^b$</td>
<td>1,276 (451)</td>
<td>59%</td>
</tr>
<tr>
<td>95% Coinsurance</td>
<td>1,121 (382)</td>
<td>76%</td>
</tr>
<tr>
<td><strong>All Plans</strong></td>
<td>5,811 (1,985)</td>
<td>34%</td>
</tr>
</tbody>
</table>

Notes:
- Plan summary statistics (see Appendix A).
- Refusal rates of newly assigned enrollees found to have statistically significant differences across plans.

$^a$ Mixed coinsurance plans include dental.

$^b$ Both individual and family deductibles.

$^c$ Average out-of-pocket share.
$y_{i,t} = \lambda_p + \tau_t + \alpha_{l,m} + \varepsilon_{i,t}$

- Outcome $y_{i,t}$ (e.g. medical expenditures) regressed on plan, year, and location x start month fixed effects
- Key coefficients of interest are the six plan fixed effects ($\lambda_p$)
- Because plan assignment was random conditional on location and start month, include full set of interactions
  - Need to condition on anything correlated with assignment (site, start and interaction)
  - Year fe’s (to account for multiple years of study)
  - Analyze individual level data but cluster on family (level of assignment)
Experimental treatment effects: results

Table 2
Effects on Utilization

<table>
<thead>
<tr>
<th></th>
<th>Total Spending$^a$ Share with Any OLS (levels)</th>
<th>Inpatient Spending Share with Any OLS (levels)</th>
<th>Outpatient Spending Share with Any OLS (levels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant (Free Care Plan, N = 6,840)</td>
<td>0.931 (0.006)</td>
<td>2170 (78)</td>
<td>0.103 (0.004)</td>
</tr>
<tr>
<td>25% Coins (N = 2,361)</td>
<td>-0.079 (0.015)</td>
<td>-648 (152)</td>
<td>-0.022 (0.009)</td>
</tr>
<tr>
<td>Mixed Coins (N = 1,702)</td>
<td>-0.053 (0.015)</td>
<td>-377 (178)</td>
<td>-0.018 (0.009)</td>
</tr>
<tr>
<td>50% Coins (N = 1,401)</td>
<td>-0.100 (0.019)</td>
<td>-535 (283)</td>
<td>-0.031 (0.009)</td>
</tr>
<tr>
<td>Individual Deductible (N = 4,175)</td>
<td>-0.124 (0.012)</td>
<td>-473 (121)</td>
<td>-0.006 (0.007)</td>
</tr>
<tr>
<td>95% Coins (N = 3,724)</td>
<td>-0.170 (0.015)</td>
<td>-845 (119)</td>
<td>-0.024 (0.007)</td>
</tr>
<tr>
<td>p-value: all differences from Free Care = 0</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

Notes: Table reports coefficients on plan dummies; the omitted category is the free care plan (whose mean is given by the constant term that we report in the first row). The dependent variable is given in the column headings. Standard errors, clustered on family, are in parentheses below the coefficients. Because assignment to plans was random only conditional on site and start month (Newhouse et al., 1993), all regressions include site by start month dummy variables, as well as year fixed effects. All spending variables are inflation adjusted to 2011 dollars (adjusted using the CPI–U). Site by start month and year dummy variables are demeaned so that the coefficients reflect estimates for the “average” site–month–year mix.
Experimental treatment effects: results

- Total spending lower in higher cost sharing plans
  - Can reject null that total spending in positive cost-sharing plans is equal to that in free care plans
- Break out outpatient (~58% of spending) from inpatient (~42%)
  - Pattern of lower spending in plans with higher cost sharing
  - Can reject null of no differences across plans for “any inpatient” and both measures of outpatient (any and level)
  - Effect of cost sharing on level of inpatient is consistently small and generally insignificant, suggesting it may be less price sensitive
Three main threats to validity

- Was assignment random?
- Differential attrition (participation and/or refusal) by plan assignment
- Differential measurement of outcomes by plan assignment
Experimental validity I: Randomization

- Was randomization compromised? (recall: stratified random sampling – complicated design)
- Investigated balance of pre-randomization characteristics at assignment

\[ y_{i,t} = \lambda_p + \tau_t + \alpha_{l,m} + \epsilon_{i,t} \]

- Same regression framework as for treatment effects but with pre-randomization characteristics (demographics or prior utilization) on left hand side
- Unable to reject null of balance on stratifying variables (by design!)
- Joint F-test able to reject null of balance of pre-experiment covariates not used in stratification
  - Seems only to be driven only by assignment to the (small) 50% coinsurance plan.
  - Perhaps due to chance?
Experimental validity II: Differential Attrition?

Now well-known as a key potential issue in RCTs

- Similar issue in negative income tax experiments from 1970s (Ashenfelter and Plant 1990)

Key point: attrition undermines the essence of random assignment

- Particularly concerning when attrition rates vary across treatment arms
- But even if attrition rates same, have to worry that composition of participants differs
Individuals assigned to more comprehensive plan have greater incentive to participate

RAND investigators anticipate and attempt to offset with higher lump sum payment to those in less comprehensive plan

Issue: This can equalize participation incentives across plans only on average

- Unless lump sum varies with pre-experiment expected medical spending (and it did not) incremental benefit from participating in more comprehensive coverage greater for individuals who anticipate greater medical spending

Potential bias from differential participation by plan

- Fixed cost of participation $\rightarrow$ those with lower expected spending might only participate if in more comprehensive plan $\rightarrow$ bias downward treatment effects
- If high and low expected spenders equally likely to participate in free care but high expected spenders less likely to participate when assigned less comprehensive $\rightarrow$ bias upward the treatment effects
Table 1
Plan Summary Statistics and Refusal and Attrition Rates

<table>
<thead>
<tr>
<th>Plan</th>
<th>Individuals (Families)</th>
<th>Avg. out-of-pocket share</th>
<th>Share Refusing Enrollment</th>
<th>Share Attriting</th>
<th>Share Refusing or Attriting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Care</td>
<td>1,894 (626)</td>
<td>0%</td>
<td>6%</td>
<td>5%</td>
<td>12%</td>
</tr>
<tr>
<td>25% Coinsurance</td>
<td>647 (224)</td>
<td>23%</td>
<td>20%</td>
<td>6%</td>
<td>26%</td>
</tr>
<tr>
<td>Mixed Coinsurance(^a)</td>
<td>490 (172)</td>
<td>28%</td>
<td>19%</td>
<td>9%</td>
<td>26%</td>
</tr>
<tr>
<td>50% Coinsurance</td>
<td>383 (130)</td>
<td>44%</td>
<td>17%</td>
<td>4%</td>
<td>21%</td>
</tr>
<tr>
<td>Individual Deductible(^b)</td>
<td>1,276 (451)</td>
<td>59%</td>
<td>18%</td>
<td>13%</td>
<td>28%</td>
</tr>
<tr>
<td>95% Coinsurance</td>
<td>1,121 (382)</td>
<td>76%</td>
<td>24%</td>
<td>17%</td>
<td>37%</td>
</tr>
<tr>
<td>All Plans</td>
<td>5,811 (1,985)</td>
<td>34%</td>
<td>16%</td>
<td>10%</td>
<td>24%</td>
</tr>
</tbody>
</table>

p-value, all plans equal: 0.0001, 0.0001, 0.0001
p-value, Free Care vs. 95%: 0.0001, 0.0001, 0.0001
p-value, Free Care vs. 25%: 0.4100, 0.0003, 0.0136

Notes:
1. (1) (2) (3) (4) (5) (6) reflect differing plans.
2. (a) (b) indicate differing plan levels.
3. (c) indicates differing plan type.
Differential attrition con’t

- Only 76% of individuals assigned to plans participated
- Completion rates substantially and systemetically higher in more comprehensive plans
  - 88% in (most comprehensive) free care
  - 63% in (least comprehensive) 95% coinsurance plan
- Differential participation across plans cannot be attributed to sampling variation
Differential attrition con’t

- Investigated balance of pre-randomization characteristics among those completing the experiment

\[ y_{i,t} = \lambda_p + \tau_t + \alpha_{l,m} + \varepsilon_{i,t} \]

- Considered two groups of characteristics
  - Those that directly measure pre-randomization health utilization (closely related to post-randomization outcomes)
  - Other baseline demographics

- For either group (or both combined) we reject balance across plans for those completing the experiment
  - e.g. imbalances across plans in average # of doc visits year before experiments and share of participants who had a medical exam prior to the experiment
Experimental validity III: Differential measurement

- Data on spending comes from claims filed.
- Participants in more comprehensive plans have greater incentive to file.
- Newhouse and Rogers (1985) audit study of roughly 1/3 of enrollees found under-reporting of outpatient spending ranges from 4% in free care plan to 12% in 95% coinsurance plan.
- Source of upward bias in experimental treatment effects of cost sharing.
Aside: Importance of understanding where the data come from

- We think a lot about inference problems that arise because "correlation is not causation"
- But another pitfall in inference is bias that arises because of what we are measuring
  - Increasingly important as we start to work with novel, exciting new data sets becoming available
- Examples
  - Spending comes from claims filed in RAND
  - Measuring health using insurance claims data (Song et al. 2010 NEJM)
  - Abraham Wald and WWII: Where to armor the plane?
Examining robustness of treatment effects

- Differential measurement
  - Adjust for differential filing by grossing up outpatient spending with plan-specific under-reporting estimate
  - Make no inpatient adjustment because lack estimates (and may be less likely to be under-reported)

- Differential attrition
  - Adjust for selection on observables: control for pre-randomization characteristics
    - only as good as your observables
  - How to adjust for selection on unobservables?
"Addressing" differential participation on unobservables

Three main possible approaches

Approach 1: Collect data on outcomes (e.g. health care utilization) for all people in study, including non-participants - e.g. administrative data

- Would allow comparison of outcomes based on assignment, regardless of participation (intent-to-treatment)
- Could instrument for plan enrollment with plan assignment
- Unfortunately: no administrative data available here.

- But this is a very useful / important approach in many settings (see e.g. Oregon HIE, NIT etc)
- For more on value of administrative data see J-PAL NA’s "using administrative data for randomized evaluations" (Feeney et al. 2015)
Three main possible approaches

Approach 2: Make (economic) assumptions about likely economic model of selection and use this to adjust point estimates

- Depending on economic model, might conclude existing point estimates are under- or over-estimates of true experimental treatment effects (see earlier discussion)
- Note: this is moving beyond the pure statistical nature of the RCT to impose economic modeling

Approach 3: Statistical exercise designed to find lower bound of treatment effects (under some statistical assumptions)

- Lee (2009) bounding procedure
Lee bounding procedure

- Statistical question: how bad could bias be from differential participation?
- Drop highest spenders in the lower cost sharing (more comprehensive) plan until participation rates equalized with higher cost sharing (less comprehensive) plan
  - e.g. since have 88% participation in free care but only 63% participation in 95% coinsurance plan, drop highest 28% (=\((88-63)/88\)) of spenders from free care sample to equalize participation rates
- Lee (2009) shows that this gives worst case lower bound for treatment effects under assumption that any participant who refused participation in a given plan would have also refused participation in a less comprehensive plan (monotonicity assumption)
- Note: This approach does not get you an alternative point estimate. It gets you a lower bound
Adjustment for differential under-reporting reduces estimated effects, but not by much.

Adjustments for differential participation on observables further reduces estimated treatment effects, but again not by much.

- Only reassuring in so far as believe we have rich set of observables.

At Lee lower bound, can still reject null of no utilization response to cost sharing. But...

- Can make more extreme assumptions to kill this result. e.g. Manski (1990) extreme bounds: assign non-participants the value that would minimize the treatment effect.
- Lee bounds show considerable uncertainty about magnitude (lower bounds can be ~50% lower than point estimates)
- With Lee bounds no longer able to reject null of no response of inpatient utilization to higher cost sharing (may not be price sensitive)
Summary: treatment effects

- Robust result: able to reject null of no effect of cost sharing on total medical spending
- Considerable scope for uncertainty about magnitudes of impact of cost-sharing
- Can no longer reject null of no response of inpatient spending
Oregon’s Medicaid expansion program covers those financially but not categorically eligible for Medicaid

- Low income (below 100% FPL) uninsured, able-bodied adults
- Covers doctors, hospitals, drugs, mental health etc w/ no consumer cost sharing and low or no premiums
- In 2008, Oregon had money to cover some but not all of those eligible
  - Ran a lottery (for fairness reasons)
  - 75,000 individuals signed up
  - 30,000 randomly selected to be able to apply for Medicaid
  - First stage of lottery selection on Medicaid coverage: 0.25
  - more info (including the data) at www.nber.org/oregon
Oregon: Probability of Hospitalizations

Hospital Discharge Data

- **All**: Control Mean
- **Via Emergency Department**: Control Mean plus Medicaid Effect
- **Not Via Emergency Department**: CI for Medicaid Effect

Outcomes measured over an approximately 18 month period.
Emergency Department Data

Percent with Any Visits

Any

Number

Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Outcomes measured over an approximately 18 month period.
Oregon: Types of ER visits

Emergency Department Data

Number of Visits

<table>
<thead>
<tr>
<th>Category</th>
<th>Control Mean</th>
<th>Control Mean plus Medicaid Effect</th>
<th>CI for Medicaid Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Visits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergent, Not Preventable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergent, Preventable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Care Treatable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-emergent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclassified</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Outcomes measured over an approximately 18 month period.
Oregon: Different types of health care

Mail Survey Data

- Any prescription drugs (Current)
- Any outpatient visits (Last 6 months)
- Number of Prescriptions (Current)
- Number of Outpatient visits (Last 6 months)

- Control Mean
- Control Mean plus Medicaid Effect
- CI for Medicaid Effect
Thus far:
- Definition of moral hazard
- Existence of moral hazard - Compelling evidence from RAND HIE and Oregon HIE

Modeling spending under alternative contracts:
- Why we might not want random assignment - Selection on moral hazard (Einav et al. AER 2013)
- Non-linear pricing in health insurance contracts - why we need additional modeling
  - Will start with how to construct / interpret famous RAND elasticity estimate of -0.2

Welfare consequences?
1. Random assignment shuts down selection - do we want to? (Einav et al. 2013)

2. How to translate into an economic object of interest that can be to forecast impact of (counterfactual) health insurance plans?
Random assignment of health insurance solves causal inference problem: gives impact of consumer cost sharing on medical spending.

In most settings, individuals are offered a choice of health insurance options and the policy question is how to design the choice set:
- E.g. Introduce a high deductible health savings account option
- Choice even within social insurance programs (e.g. Medicare Part D)

If choices are based in part on one’s anticipated behavioral response to the contract (i.e. one’s “moral hazard type”), then magnitude of spending effect of offering high deductible plan in option set may be very different than effect of randomly assigning high deductible plan.

Einav, Finkelstein, Ryan, Schrimpf and Cullen (AER 2013) “Selection on moral hazard in health insurance”
- General theme: Growing emphasis / awareness in applied micro on existence and substantive importance of heterogeneous treatment effects
Motivation

- Two determinants of coverage choice:
  - Expected risk (e.g., probability of seeing a doctor)
  - Preferences (e.g., risk aversion)
- This paper: when “moral hazard” (contract effect) is present, may be important to break down selection on risk:
  - Risk that’s invariant to covg choice (“traditional” selection)
  - Risk that arises b/c of coverage (“selection on moral hazard”)
- In addition to “traditional” selection based on one’s health risk (sicker individuals choose more comprehensive coverage) individual’s may also select health insurance on basis of their moral hazard type
Selection on "level" vs. "slope": all you can eat restaurants

- Who goes to all you can eat restaurants?
- People with big appetites (level)
- People who will eat a lot more than usual when the food is free on the margin (slope)
Aside: necessary (but not sufficient): heterogeneity in slope

- Found it in our paper, but (as often happens) later found more compelling evidence in a different context
  - Einav et al. (AEJ Applied 2013) Beyond Statistics: The Economic Content of Risk Scores
  - Back to the Medicare Part D donut hole!
- Prescription drug coverage introduced in 2006
  - Largest expansion of Medicare since inception
  - 32 million beneficiaries, 11% of Medicare spending
Typical plan

Figure 1: Standard benefit design (in 2008)

The figure shows the standard benefit design in 2008. "Pre-Kink coverage" refers to coverage prior to the Initial Coverage Limit (ICL) which is where there is a kink in the budget set and the gap, or donut hole, begins. As described in the text, the actual level at which the catastrophic coverage kicks in is defined in terms of out-of-pocket spending (of $4,050), which we convert to the total expenditure amount provided in the figure. Once catastrophic coverage kicks in, the actual standard coverage specifies a set of co-pays (dollar amounts) for particular types of drugs, while in the figure we use instead a 7% co-insurance rate, which is the empirical average of these co-pays in our data.
Data and sample

- 20% random sample of all Part D-covered individuals (2007 - 2009)
- Cost sharing features of each plan, basic demographics, and detailed, claim-level information on drugs purchased
Response to price: bunching at the kink

- Sharp increase in price when go into donut hole
  - On average price goes from 34 to 93 cents for every dollar
- Standard economic theory: with convex preferences smoothly distributed in population, should see bunching at the convex kink

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

This figure illustrates graphically the theoretical prediction that individuals will bunch at the convex kink point in their budget set. The solid line illustrates the budget set of the same standard benefit design as in Figure I; the standard budget set has a kink (price increase) at $2,510 in total spending. By contrast, the dashed line considers an alternative budget set with a linear budget (above the deductible) at the co-insurance arm’s cost sharing rate. The solid and dashed indifference curves represent two individuals with different healthcare needs who would have different total drug spending under the linear contract. The (healthier) individual denoted by the solid indifference curve is not affected by the introduction of this kink; his indifference curve remains tangent to the lower part of the budget set. The (sicker) individual with the dashed indifference curves consumed above the kink under the linear budget set; with the introduction of the kink her indifference curve is now exactly tangent to the upper part of the budget set at the kink. With the introduction of the kink, this latter individual would therefore decrease total spending to the level of the kink location. By extension, any individual whose indifference curve was tangent to the linear budget set at a spending level between that of the two individuals shown would likewise decrease total spending to the level of the kink location, thereby creating “bunching” at the kink.
Figure II: The Distribution of Annual Drug Expenditure for Medicare Part D Enrollees in 2008

The figure displays the distribution of total annual prescription drug spending in 2008 for our baseline sample. Each bar represents the set of people that spent up to $100 above the value that is on the x-axis, so that the first bar represents individuals who spent less than $100 during the year, the second bar represents $100-200 spending, and so on. For visual clarity, we omit from the graph the 3% of the sample whose spending exceeds $6,500. The kink location (in 2008) is at $2,510. N = 1,251,984.
Bunching at kink III: All years, normalized around standard kink

Figure IV: The Magnitude of The Excess Mass of Annual Drug Expenditure around The Kink

Total annual prescription drug spending on the x-axis is reported relative to the (year-specific) location of the kink, which is normalized to zero. Sample uses beneficiary-years in our 2007-2009 baseline sample whose annual spending is within $2,000 of the (year-specific) kink location. The points in the figure display the distribution of annual spending; each point represents the set of people that spent up to $20 above the value that is on the x-axis, so that the first point represents individuals who spent between -$2,000 and -$1,980 from the kink, the second point represents individuals between -$1,980 and -$1,960, and so on. We normalize the frequencies so that they add up to one for the range of annual spending shown. The dashed line presents the counterfactual distribution of spending in the absence of a kink. This is calculated by fitting a cubic CDF function – that is, for each $20 bin of spending \((x; y)\) we fit \(F(y) = F(x)\), where \(F(z) = a + bz + cz^2 + dz^3\) – using only individuals with annual spending (relative to the kink location) between -$2,000 and -$200, and subject to the integration constraints that \(F(2000) = 0\) and \(F(+2000) = 1\). N = 2,589,458.

Estimate excess mass of 29.1% (standard error = 0.003) statistically significant excess mass rejects null of no behavioral response to price

Amy Finkelstein ()
Spring 2019 52 / 127
Detecting heterogeneity in moral hazard

- Greater bunching around the kink for a given population (e.g. males) indicates greater demand sensitivity to out-of-pocket price.
- Empirically, we identify heterogeneity in the behavioral response by documenting sharp changes in the presence of specific individual characteristics around the kink.
- Idea: an individual characteristic (being male, having a particular diagnosis) that is over-represented among individuals around the kink indicates that individuals with this characteristic have a greater behavioral response to kink.
  - Likewise an individual characteristic that is under-represented at the kink is less responsive.
"Bunchers" are younger

Panel A. Average age

Total annual drug expenditure (relative to the kink location)
"Bunchers" are more likely to be male
"Bunchers" have better subsequent outcomes

So does decreased drug use improve health?
Why does this matter?

- Beyond Statistics: The Economic Content of Risk Scores
- Risk scores - health, credit etc
  - Health risk scores (designed to predict individual’s healthcare spending), used to set government reimbursement rates for private insurers in Medicare Advantage, and in health insurance exchanges.

What we found:
- Heterogeneity in utilization response to price: younger, healthier, and male more likely to reduce drug consumption in response to price increase
- However, risk score (predicted spending based on demographics) is smooth through the donut hole (Figure 4a)

Risk score created as a statistical exercise: predicting individuals’ spending under current environment
- Does not capture behavioral responses that might affect spending under an alternative contract

Thus even with "perfect" statistical predictions under a given contract, could still have cream skimming incentives.
Empirical challenge:
- thus far have identified risk type and preferences from claims and choices, assuming no moral hazard
- now want to allow for moral hazard so need another data element - variation in menu of options

Setting: employer-provided health insurance (Alcoa, Inc.)
- Now want to identify not only risk type and risk aversion from insurance choice and insurance claims but also moral hazard
- Therefore need some (hopefully) exogenous variation in choice set to move people across contracts

Find substantial heterogeneity in moral hazard (spending) effects of cost-sharing (necessary but not sufficient)

Find substantial selection on moral hazard:
- More behaviorally responsive individuals more likely to choose more comprehensive coverage
- For determining plan choice, selection on moral hazard roughly as important as "traditional" selection on health risk, and considerably more important than selection on risk aversion
Implications (example)

- For reducing health care spending / combatting mh through higher consumer cost sharing:
  - Selection on mh implies non-random selection into plans - consider high deductible plans, selected by low "moral hazard" types
  - Abstracting from selection on moral hazard could lead to substantial over-estimation of spending reduction associated with offering a high deductible plan
Selection on “level” vs. “slope”: implications

- IO vs. Labor: How useful is a "good" LATE vs. a "bad" OLS?
- Imagine an RCT that randomizes a subsidy to encourage take up of a program
  - Imagine that very few people take it up
  - The RCT solves the "selection on levels" problem but if there is substantial heterogeneity in treatment and selection on "slope" how useful is the estimate?
- Imagine an OLS regression comparing people on vs off the program, attempting to control for "stuff"
  - Now don’t have to worry about selection on slope (have population) but selection on levels a concern
Ideal experiment?

- [If you can do it]
- Proposal
  - (Endogenous) choice of two linear coverage (constant coinsurance) plans (high and low)
  - Within each (endogeneously chosen plan): randomly assign a new (constant coinsurance) plan $\rightarrow$ estimate behavioral response of those with each old plan
  - Do those who chose higher coverage endogeneously have different estimated moral hazard effect?
- Advantages
  - Estimate treatment effects “cleanly” (via random assignment) and purged of selection
  - Linear coverage makes “treatment” easy to define
- But even still need a model
  - Counterfactuals - extrapolate to other (somewhat different) contracts
  - Welfare (vs spending) analysis?
Ideal experiment?

- [Did Tom Price Give Us a Gift?]
- Medicare created a 5 year (2016-20202) (mandatory participation) bundled payment intervention randomized across MSAs
  - Then-representative Price objected to "experimenting with the health of the american public"
  - In October 2017, HHS Secretary Price converted intervention to voluntary in half the MSAs (starting in 2018).
- Who selected into intervention and how is this correlated with level and slope?
  - In progress (with Liran Einav, Yunan Ji and Neale Mahoney).
Growing Awareness of Importance of Potential "Selection on Treatment"

- Selection on moral hazard is a specific (economic) application of a more general (econometric) point - "selection on gains"
- Heterogeneity in treatment effects + selection into treatment based on anticipated treatment effects
  - Heckman, Urzua and Vytlacil (2006) discuss properties of IV in this setting ("essential heterogeneity")
  - Tracing out marginal treatment effects (MTEs) as exciting, active area of work (will see again when get to unit on disability insurance...)

Amy Finkelstein ()

Spring 2019 63 / 127
Recap

1. Existence of moral hazard in health insurance - evidence from experiments [Done]

2. Implications for spending under alternative contracts:
   1. Selection on moral hazard (Einav et al., 2013) [Done]
   2. Response to non-linear pricing - how to use reduced form estimates for counterfactuals [Up next]
How to interpret / use RAND treatment effects

- Testing: can reject null of no impact of cost sharing on health care spending
- Quantifying: How do we learn from these treatment effects beyond rejection of the null?
  - How to transform plan fixed effects into an economically interpretable construct that can be applied out of sample (to impact of plan designs not observed in the experiment)
  - RAND investigators made one such attempt and concluded: price elasticity of demand for medical care = -0.2 (FAMOUS. Treated with reverence by profession)
  - Key point: This famous result derives from the experimental data plus a large number of (economic and statistical) assumptions.
    - True of any out-of-sample extrapolation of experimental treatment effects more generally.
    - To learn the most from an experiment (or any reduced form estimate) we often have to layer on additional assumptions.
How to translate to a price elasticity of demand?

- Challenge (in real world and in RAND experiment): health insurance contracts are highly-non-linear
- Price faced by consumer falls as total medical spending cumulates during the year
- In general:
  - 100% during deductible
  - 10-20% in coinsurance rate
  - 0% once out of pocket spending limit has been reached
- With a non-linear budget set, what is “the price” of medical care (or the elasticity w.r.t “the price”)

Amy Finkelstein ()
Spring 2019 66 / 127
What price?

Figure 2: Non-Linear Health Insurance Coverage

- High Deductible Plan
- Low Deductible Plan
- Constant Coinsurance Plan

Out-of-pocket Expenditure ($US)

Total Medical Expenditure ($US)
Analytical decisions

- How to analyze medical expenditures that occur at different times, under potentially different cost sharing rules but stem from same underlying health event?
- What price does individual respond to in making medical spending decision?
  - Current “spot” price of care (fully myopic)
  - Expected end-of-year price (fully forward looking + model of expectation formation)
  - Realized end-of-year price (health care consumption happens on the margin)
  - Weighted average of the prices paid over year (ad hoc)

- NB: These modeling challenges are inherent in problem of extrapolating from any study of impact of a particular health insurance policy on spending.
  - Not unique to RAND.
  - Will come up again in course (and perhaps in your research...)
Briefly review the assumptions that allowed them to transform the experimental treatment effects into their famous estimate of “the price” elasticity of demand for medical care of -0.2

Original source / more detail in Keeler and Rolph (1988 JHE)

For our pedagogical purposes, the details per se are not important (although I will summarize some to give you a flavor)

What is important is that you see how this estimate is derived from the experiment + assumptions

- Look inside the sausage factory (always a good idea)
Group claims into “episodes” of care

- An episode is an unbreakable and perfectly forecastable “bundle” of individual claims
- Precise grouping relies on detailed clinical input
  - e.g. each hospitalization is a separate episode; routine spending on diabetes care over the year is considered a single episode but “flare-ups” are not. Each cold or accident is a separate episode, but these can run concurrently

Regress average costs per episode on plan dummies and various controls and conclude that cost sharing has no affect on costs per episode conditional on having an episode

- Ignores possible compositional differences in episodes of care arising from different coinsurance
- Focus analysis going forward on extensive margin only (i.e. occurrence rate of episodes)
What price do individuals respond to?

To ascertain: Compare occurrence rate of episodes for people who are closer vs further from the MDE and for people in cost-sharing plans who have exceeded their MDE vs free care plan

- Trying to get at whether “future” price matters
- Concern: families with higher underlying spending propensities more likely to come closer to hitting MDE. This selection problem addressed via various modeling assumptions.

Find no evidence of higher episode rates among individuals who are closer to hitting MDE

Conclude that extensive margin response is completely myopic - i.e. based on “spot” price of care (not forward looking)
Final step: limit the sample to individuals in periods of the year with >$400 from hitting MDE and compute arc elasticities of spending wrt experimentally assigned coinsurance rate

- Use spot price bc have found myopia on extensive margin

Compute arc-elasticities for particular categories (hospital, acute outpatient etc) and particular pairs of plans

Average (eyeball) various arc-elasticities across pairs of plans and categories of spending: conclude: about =0.2
Table 11
Arc price elasticities of medical spending.

<table>
<thead>
<tr>
<th>Range</th>
<th>Outpatient</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acute</td>
<td>Chronic</td>
<td>Well</td>
<td>Total</td>
<td>Hospital</td>
<td>Total</td>
<td>Dental</td>
</tr>
<tr>
<td>0-25</td>
<td>-0.16</td>
<td>-0.20</td>
<td>-0.14</td>
<td>-0.17</td>
<td>-0.17</td>
<td>-0.17</td>
<td>-0.12</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>25-95</td>
<td>-0.32</td>
<td>-0.23</td>
<td>-0.43</td>
<td>-0.31</td>
<td>-0.14</td>
<td>-0.22</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.10)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
</tbody>
</table>

Note: Standard errors are given in parentheses. Arc elasticities come from the formula \((q(1) - q(2))/((p(1) - p(2)) \times (p(1) + p(2)))/2)/(q(1) + q(2))/2\). Table 8 shows that errors in \(q(1)\) and \(q(2)\) are small relative to \(q(1) + q(2)\), but large relative to \(q(1) - q(2)\). We assume that the error in estimating \(q(1) + q(2)\) is negligible. We also assume the error in estimating \(q(25) - q(95)\) is the square root of \(1/2\) the sum of the squared errors of estimating \(q(25)\) and \(q(95)\). The \(1/2\) comes from an assumption that half of the error in \(q(25)\) and \(q(95)\) that are estimated relative to the same free-plan experience comes from variation in the free-plan and drops out when one subtracts.

Source: Keeler and Rolph (1988, JHE,).
Using the RAND elasticity

- Applying the -0.2 estimate in a manner consistent with how it is generated is non-trivial
  - Elasticity estimate assumes that in deciding whether to consume medical care, individuals fully anticipate spending within an “episode” but make their decisions myopically with respect to the potential for other episodes

- Researcher needs to obtain micro data on medical claims, group into episodes, calculate spot price for each episode
  - Many subsequent researchers have instead applied in a simpler fashion, summarizing a plan with a single price and applying the RAND single elasticity to that single price
  - Different assumptions made about what “the price” is (average out of pocket, realized end of year price, expected end of year price)
What price is used can be important

- Example: What would be spending effect of replacing 28% constant coinsurance plan with RAND’s 25% coinsurance plan up to the MDE
- Some ways to summarize RAND 25% coinsurance plan with a single price
  - Dollar-weighted average price (10%)
  - Person-weighted average price (17%)
  - person-weighted average end of year price (13%)
- Applying -0.2 estimate to changing from each of these prices to 28 cents yields predicted reduction in health spending of 18%, 9%, and 14% respectively
- In this example, decision of how to apply the price leads to differences in the predicted reduction of spending that vary by a factor of 2
Dangerous to use a single price for a non-linear contract

- In general no “right” way to summarize a non-linear budget set with a single price
- And we just saw how reasonable yet ad hoc “fixes” can have very different implications
- Suggests advantages of applying non-linear budget set estimation techniques (see 471)
- But still face issues original RAND investigators grappled with:
  - Must model distribution of medical shocks throughout the year and evolution of individuals’ beliefs about those shocks
  - How forward looking are individuals? i.e. do they take account of entire non-linear budget set or respond only to spot price or something in between?
  - Will investigate this next
- Modeling the response to the non-linear budget set induced by health insurance (or other!) contracts is an open / active area of research
Rand HIE: Summary

- Rejection of null of no price sensitivity of demand for health care
  - Uncertainty regarding magnitude of spending differences across plans with different cost-sharing due to differential participation and measurement
- Translation of experimental estimates to economic objects of interest requires assumptions beyond “raw” experimental treatment estimates
  - Famous elasticity of -0.2 does not come directly from the experiment
  - More work needed on economic modeling of spending response to non-linear budget sets
Non-linear budget set: which price?

- Consider impact on spending of introducing high deductible plan (previously no deductible)
  - Completely myopic individual: reacts to “price” increase to 100%
  - Fully forward looking individuals with annual expenditures typically above the new deductible might not change his behavior much
- Which plan provides more incentives to economize on medical spending:
  - Plan A (10% coinsurance and $5,000 out of pocket max) vs. Plan B (50% coinsurance and $5,000 out of pocket max)
  - Naïve answer: B is less generous and will lead to lower medical utilization
  - But depends on distribution of medical spending w/o insurance and how ff looking people are: if have $10,000 surgery early in year (or expect to have it during year), for rest of year spot price lower under plan B
Basic empirical question: do individuals take dynamic incentives into account in their medical consumption decisions?
  i.e. do they respond to "future price" of medical care?

Why you might be affected by the current ("spot") price:
  Myopic (completely discount future)
  Liquidity constrained
  Misunderstanding of price schedule

Empirically challenging to test null that individuals respond only to spot price
  Challenge: how to separately identify effect of spot and future price?

Spot price and future price often vary jointly
  Low spending individual faces both a higher spot price and a higher future price
  Variation in insurance contracts (e.g. changes in deductibles, coinsurance etc) will change both spot price and future price
Ideal experiment

- Randomize people across plans with same spot, different future price
  - Can test whether respond to future price
- RAND actually did this!!
  - Randomized MDE within coinsurance amount
  - Sadly, though issues of low power (and also low MDEs $\Rightarrow$ can affect spot price almost immediately)
Approximating the ideal experiment

- Aron-Dine et al. (ReStat 2015)

Idea: typical health insurance contracts offer coverage for a fixed duration, and reset on pre-specified dates

- Generate identifying variation in future price when they get applied to individuals whose initial coverage period is shorter

Example: Annual health insurance contract (January 1- December 31) with annual deductible

- When people join mid-year, deductible remains at annual level but applies only till end of calendar year
- Individuals who join the plan later in the year face:
  - higher future price (have fewer months to spend past the deductible)
  - same spot price
- Examine initial care utilization across individuals who join in different months
Two applications

- Health insurance contracts are annual, with open enrollment periods (typically Oct and Nov) to change coverage for following calendar year
  - How do we find variation in contract length / when join plan?
- Application #1: Employer provided health insurance
  - New hire: Individuals who join the firm mid-year
- Application #2: Medicare Part D prescription drug coverage
  - Newly eligibles: Can join in the birth-month you turn 65
Graphical evidence: employer-provided health insurance

Initial claims

Initial spending

All utilization measures refer to utilization by the employee and any covered dependents. Initial claims (left panels) are defined as any claims within the first three months, and initial spending (right panels) is defined as the sum of all claim amounts of the claims that were made within the first three months. Expected end-of-year price is computed for the deductible plan only and corresponds to the end-of-year prices reported in Table 2. Sample sizes by plan and join quarter are reported in Table 2.
Graphical evidence: Medicare Part D

Figure 2: Probability of initial claim and expected end-of-year-price by enrollment month

Figure graphs the pattern of expected end-of-year price and of any initial drug claim by enrollment month for individuals in Medicare Part D during their first year of eligibility (once they turn 65). We graph results separately for individuals in deductible plans and no deductible plans. We calculate the expected end-of-year price separately for each individual based on his plan and birth month, using all other individuals who enrolled in the same plan that month. The fraction with initial claim is measured as the share of individuals (by plan type and enrollment month) who had at least one claim over the first three months. See Appendix B for more details on the construction of variables used in this figure. $N=137,536$ ($N=108,577$ for no deductible plans, and $N=28,959$ for deductible plans).
Beyond testing: quantifying response to dynamics

How to quantify spending response to non-linear budget set?

- Reduced form results suggest don’t want to summarize budget set with a single "price"

Quantifying requires additional economic and statistical modeling assumptions

- Two related papers: Einav, Finkelstein and Schrimpf (2015 QJE; 2017 JPubEc)
Broad Motivation

- "Credibility revolution" in economics
  - (Rightly) emphasize value of research design that produces compelling (often visual) evidence of a behavioral response

- "Structural" models
  - (Rightly) emphasize defining and estimating economic objects that can be used to predict behavior in counterfactual environments

- "Sufficient statistics" (Chetty, 2009)
  - Use simple models to directly and transparently map reduced form parameters into economic objects of interest

- Simple (not novel) point: choice of model can be consequential
  - Will show how two "reasonable" models can match the reduced form facts but produce very different counterfactual predictions
Specific application: Bunching estimators

- Increased analysis in public economics of "bunching" at kink points in convex budget sets (Kleven 2016 Annual Reviews)
  - Existence of bunching (or "excess mass") can provide compelling, visual evidence against null of no behavioral response to incentives
  - Magnitude of excess mass often used to infer relevant elasticities

- Many applications with non-linear schedules: income taxes, home sale taxes, pensions, electricity, fuel economy, mortgages, cell phones, ....

- Two factors behind recent popularity:
  - Detecting bunching: Increased availability of rich, large administrative data
  - Interpreting bunching: Saez (2010) seminal paper
    - Illustrates how to translate observed bunching into a "structural" behavioral elasticity parameter
Specific context: highly non-linear nature of Part D contracts

The figure shows the standard benefit design in 2008. "Pre-Kink coverage" refers to coverage prior to the Initial Coverage Limit (ICL) which is where there is a kink in the budget set and the gap, or donut hole, begins. As described in the text, the actual level at which the catastrophic coverage kicks in is defined in terms of out-of-pocket spending (of $4,050), which we convert to the total expenditure amount provided in the figure. Once catastrophic coverage kicks in, the actual standard coverage specifies a set of co-pays (dollar amounts) for particular types of drugs, while in the figure we use instead a 7% co-insurance rate, which is the empirical average of these co-pays in our data.
Medicare Part D

- Prescription drug coverage introduced in 2006
  - Largest expansion of Medicare since inception
  - 32 million beneficiaries, 11% of Medicare spending
  - Typical coverage highly non-linear
    - Government sets standard plan, but actual plans often provide different coverage
- Many planned and potential changes to contract design
  - E.g., under ACA, “donut hole” will be “filled” by 2020
- Objective: Develop a model that allows us to assess responses of drug spending to non-linear contracts ("moral hazard")
  - Conceptual: e.g. anticipatory behavior
  - Quantitative: Impact of changes to contracts on drug spending
Data and sample

- 20% random sample of all Part D-covered individuals (2007 - 2009)
- Cost sharing features of each plan, basic demographics, and detailed, claim-level information on drugs purchased
Response to price: bunching at the kink

- Sharp increase in price when go into donut hole
  - On average price goes from 34 to 93 cents for every dollar
- Standard economic theory: with convex preferences smoothly distributed in population, should see bunching at the convex kink

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

This figure illustrates graphically the theoretical prediction that individuals will bunch at the convex kink point in their budget set. The solid line illustrates the budget set of the same standard benefit design as in Figure I; the standard budget set has a kink (price increase) at $2,510 in total spending. By contrast, the dashed line considers an alternative budget set with a linear budget (above the deductible) at the co-insurance arm’s cost sharing rate. The solid and dashed indifference curves represent two individuals with different healthcare needs who would have different total drug spending under the linear contract. The (healthier) individual denoted by the solid indifference curve is not affected by the introduction of this kink; his indifference curve remains tangent to the lower part of the budget set. The (sicker) individual with the dashed indifference curves consumed above the kink under the linear budget set; with the introduction of the kink her indifference curve is now exactly tangent to the upper part of the budget set at the kink. With the introduction of the kink, this latter individual would therefore decrease total spending to the level of the kink location. By extension, any individual whose indifference curve was tangent to the linear budget set at a spending level between that of the two individuals shown would likewise decrease total spending to the level of the kink location, thereby creating “bunching” at the kink.
Figure II: The Distribution of Annual Drug Expenditure for Medicare Part D Enrollees in 2008

Figure displays the distribution of total annual prescription drug spending in 2008 for our baseline sample. Each bar represents the set of people that spent up to $100 above the value that is on the x-axis, so that the first bar represents individuals who spent less than $100 during the year, the second bar represents $100-200 spending, and so on. For visual clarity, we omit from the graph the 3% of the sample whose spending exceeds $6,500. The kink location (in 2008) is at $2,510. N = 1,251,984.
Bunching at kink II: Year-to-year movement in standard kink

Figure III: The Distribution of Annual Drug Expenditure around The Kink, for each Year

Figure displays the distribution of total annual prescription drug spending, separately by year, for individuals in our baseline sample whose annual spending in a given year was between $1,500 and $3,500 (N=1,332,748 overall; by year it is 447,012 (2007), 443,323 (2008), and 442,413 (2009)). Each point in the graph represents the set of people that spent up to $20 above the value that is on the x-axis, so that the first point represents individuals who spent between $1,500 and $1,520, the second bar represents $1,520-1,540 spending, and so on. We normalize the frequencies so that they add up to one for each series (year) shown.
Bunching at kink III: All years, normalized around standard kink

Total annual prescription drug spending on the x-axis is reported relative to the (year-specific) location of the kink, which is normalized to zero. Sample uses beneficiary-years in our 2007-2009 baseline sample whose annual spending is within $2,000 of the (year-specific) kink location. The points in the figure display the distribution of annual spending; each point represents the set of people that spent up to $20 above the value that is on the x-axis, so that the first point represents individuals who spent between -$2,000 and -$1,980 from the kink, the second point represents individuals between -$1,980 and -$1,960, and so on. We normalize the frequencies so that they add up to one for the range of annual spending shown. The dashed line presents the counterfactual distribution of spending in the absence of a kink. This is calculated by fitting a cubic CDF function – that is, for each $20 bin of spending $(x; y)$ we fit $F(y) = F(x)$, where $F(z) = a + bz + cz^2 + dz^3$ – using only individuals with annual spending (relative to the kink location) between -$2,000 and -$200, and subject to the integration constraints that $F(2000) = 0$ and $F(+2000) = 1$. $N = 2,589,458$.

Estimate excess mass of 29.1% (standard error = 0.003) statistically significant excess mass rejects null of no behavioral response to price.

Amy Finkelstein ()
Spring 2019 94 / 127
Timing of purchases (December)

[Graph showing fraction of people with at least one drug purchase during December against total annual expenditure relative to the kink location.]
Timing of purchases (Sept - Dec)
Where do we go from here?

- Goal: want to make quantitative inferences about behavior under counterfactual contracts
- EFS (2017) consider two models
  - Static, frictionless model (adaptation of Saez 2010)
  - Dynamic model (EFS 2015)
- Punchline: Both models match (by construction) the bunching / excess mass, but produce very different out-of-sample predictions
Static model of drug spending, Saez-style

- **Saez (2010)**
  - Static, frictionless model of labor supply
  - Key insight: in this model, can translate observed bunching in annual earnings around convex kinks in income tax schedule into an estimate of labor supply elasticities

- We adapt it to Part D context, sticking as closely as possible to Saez’s original model
Static model of drug use

- Assume individual $i$ has quasi-linear utility over total drug spending $m$ and residual income $y$
- Parametric assumptions:

$$u_i(m, y) = \left[ 2m - \frac{\zeta_i}{1 + \frac{1}{\alpha}} \left( \frac{m}{\zeta_i} \right)^{1 + \frac{1}{\alpha}} \right] + \left[ l_i - C(m) \right]$$

where $C(m)$ maps total spending $m$ into out of pocket spending
- $g_i(m)$ is chosen to obtain a tractable, constant elasticity form of the spending function similar to Saez
Static mode (con’t)l

- Parametric assumptions:

\[ u_i(m, y) = \left[ 2m - \frac{\zeta_i}{1 + \frac{1}{\alpha}} \left( \frac{m}{\zeta_i} \right)^{1+\frac{1}{\alpha}} \right] + \left[ l_i - C(m) \right] \]

\[ g_i(m) \]

- With linear coverage \((C(m) = c \cdot m, c \in [0, 1])\) optimal drug expenditure is

\[ m = \zeta_i (2 - c)^\alpha. \]

- Specification implies a constant elasticity \(\alpha\) of spending with respect to \((2 - c)\).
  
  - Very similar to Saez: constant elasticity with respect to \((1 - t)\) where \(t\) is marginal tax rate on income.
  - Rest of our derivation follows his closely; derives mapping between empirical "bunching" and elasticity \(\alpha\).
Derive (a la Saez) expression relating elasticity ($\alpha$) to a bunching estimate $B$:

$$B = m^* \left[ \left( \frac{2 - c_0}{2 - c_1} \right)^\alpha - 1 \right] \frac{h(m^*)_- + h(m^*)_+}{2} \left( \frac{2-c_0}{2-c_1} \right)^\alpha$$

- $B = N_{actual} - N_{counter}$; number of people empirically around kink over and above number we (counterfactually) estimate would be in this area if kink did not exist.
- $c_1 \gg c_0$ are marginal price of drugs after and before gap, respectively.
- $m^*$ location of kink.
Approximate counterfactual distribution of spending near kink by fitting a polynomial approximation to spending below the kink, subject to integration constraint

- Use counterfactual to project into $200 window around kink to estimate $B$
- Explore sensitivity to polynomial choice, spending size bin, exclusion window

- Use model to map estimates of $B$ to $\alpha$
### Elasticity estimates from the static model

<table>
<thead>
<tr>
<th>Counterfactual distribution</th>
<th>Exclusion window$^a$</th>
<th>Bin size$^b$</th>
<th>Excess mass$^c$</th>
<th>Elasticity$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>200</td>
<td>40</td>
<td>0.401</td>
<td>-0.047</td>
</tr>
<tr>
<td>Cubic</td>
<td>200</td>
<td>40</td>
<td>0.314</td>
<td>-0.037</td>
</tr>
<tr>
<td>Linear</td>
<td>200</td>
<td>60</td>
<td>0.418</td>
<td>-0.049</td>
</tr>
<tr>
<td>Linear</td>
<td>100</td>
<td>40</td>
<td>0.586</td>
<td>-0.034</td>
</tr>
</tbody>
</table>

$^a$ Exclusion window refers to the distance from the kink location within which we calculate the response to the kink.

$^b$ Bin size refers to the spending size of bins, which is used to fit the pre-kink spending distribution.

$^c$ Excess mass: \[ \frac{B}{N_{\text{counter}}} = \frac{N_{\text{actual}} - N_{\text{counter}}}{N_{\text{counter}}} \].

$^d$ Elasticity of spending calculated wrt end-of-year cost-sharing rate $C$ of each individual and estimate of $\alpha$. We report the average estimated elasticity across individuals.
Dynamic model of drug use

- EFS (2015 QJE)
- Risk-neutral fwd-looking individual faces uncertain health shocks
- Prescriptions are defined by \((\theta, \omega)\), where \(\theta > 0\) is the prescription’s (total) cost and \(\omega > 0\) is the monetized cost of not taking the drug
  - Arrive at weekly rate \(\lambda\), drawn from \(G(\theta, \omega) = G_2(\omega|\theta)G_1(\theta)\)
- Insurance specifies (discrete) covg length \(T\) and defines \(c(\theta, x)\) – the out-of-pocket cost associated with a prescription that costs \(\theta\) when total spending so far is \(x\).
- When a shock arrives, individuals make binary choice (fill prescription or not)
- Flow utility given by

\[
\begin{align*}
  u(\theta, \omega; x) &= \begin{cases} 
  -c(\theta, x) & \text{if filled} \\
  -\omega & \text{if not filled}
  \end{cases}
\end{align*}
\]
Individual choice: fill prescription or not

Optimal behavior characterized by simple finite horizon dynamic problem

Value function given by solution to following Bellman equation:

$$\nu(x, t) = (1 - \lambda) \delta \nu(x, t - 1) +$$

$$\lambda \int \max \left\{ \begin{array}{l} -c(\theta, x) + \delta \nu(x + \theta, t - 1), \\ -\omega + \delta \nu(x, t - 1) \end{array} \right\} dG(\theta, \omega)$$

with terminal condition $\nu(x, 0) = 0$ for all $x$
Three key economic objects

- Statistical description of distribution of health shocks: $\lambda$ and $G_1(\theta)$
- “Primitive” price elasticity capturing substitution between health and income: $G_2(\omega|\theta)$
  - If $\omega \geq \theta$, fill even if have to pay full cost
  - If $\omega < \theta$, fill only if some portion of cost (effectively) paid by insurer
  - Convenient to think about the ratio $\omega/\theta$
- Extent to which individuals understand and respond to dynamic incentives in non-linear contract: $\delta \in [0, 1]$
  - “Full” myopia ($\delta = 0$): don’t fill if $\omega < c(\theta, x)$
  - Dynamic response ($\delta > 0$): utilization depends on both spot and future price
  - $\delta$ is context specific! ... Captures salience, discounting, and perhaps liquidity constraints
Estimation

- We parameterize the model with distributional and functional form assumptions, and model heterogeneity using a discrete type space
- Estimate using simulated minimum distance
- Moments:
  - Distribution of annual spending: average, standard deviation, pct zero, etc..
  - **Bunching**: Histogram of total spending around the kink (+/- $500)
  - Claim timing pattern around kink
  - Covariance in spending between first half and second half of year
Model Fit: Spending around the kink

Figure VI: Actual and Predicted Distributions of Annual Drug Expenditure

Figure shows the distribution of observed and predicted total annual drug spending. The top panel shows the results for the whole distribution, where each bar represents a $100 spending bin above the value on the x-axis (except for the last bar, which includes all spending above $5,900). The bottom panel "zooms in" on spending within $1,000 of the (year-specific) kink (which is normalized to 0) and shows observed and predicted spending in $20 bins, where each point represents individuals who spend within $20 above the value on the x-axis. Frequencies in the bottom panel are normalized to sum to 1 across the displayed range. We note that the figure is based on the estimation sample rather than the baseline sample (see footnote 23), so the summary statistics do not perfectly match those presented in Table I.

Amy Finkelstein ()
Main counterfactual exercise considers “filling the gap” as specified by ACA by 2020:

- Coinsurance rate in standard contract will remain at its pre-gap level (of 25%) until out of pocket spending puts individual at CCL

Consider spending effect of “filling the gap” in the 2008 standard benefit design

- On average, increases total spending by $204 (12%)
- Insurer spending increases by $358; out of pocket declines by $154
Heterogeneity in who is affected

Figure VIII: Predicted Changes in Drug Expenditure as a Result of "Filling The Gap"

Figure shows the change in spending from "filling the gap" (i.e. providing 25% cost-sharing in the gap) for the 2008 standard benefit (which provides no coverage in the gap). In the top panel, the x-axis shows predicted spending under the 2008 standard contract. The solid black line shows the mean change in spending for individuals whose predicted spending under the 2008 standard contract is on the x-axis. The dashed lines show the 10th, 25th, 50th, 75th, and 90th percentile changes in spending. In the bottom panel, we show the average predicted weekly spending, by calendar week, for the 2008 standard benefit (gray) and for the "filled gap" contract (black).
Some subtle implications of non-linear contracts

- Change in spending by people far from gap / endogeneity of people at risk of bunching
  - Arises due to dynamic considerations / forward looking behavior
  - Estimate that about 25% of average $204 / person increase in annual spending comes from "anticipatory" response by people more than $200 below kink location
- “Filling” donut hole causes some people to decrease spending
  - Catastrophic coverage limit held constant with respect to out of pocket (vs. total) spending, so for some people marginal price actually rises
  - General point: with non-linear contracts, a given contract change can provide more coverage on margin to some individuals but less coverage to others
Elasticity estimates from the dynamic model

- Consider uniform % price reduction on all arms of standard plan
- Simulate spending for each individual under original coverage plan and modified plan and use these to compute elasticities

<table>
<thead>
<tr>
<th>(Uniform) Price Reduction(^a)</th>
<th>Average Annual Spending</th>
<th>Implied &quot;Elasticity&quot;(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% (Baseline)</td>
<td>1,838</td>
<td>-0.22</td>
</tr>
<tr>
<td>1.0%</td>
<td>1,842</td>
<td>-0.22</td>
</tr>
<tr>
<td>5.0%</td>
<td>1,860</td>
<td>-0.24</td>
</tr>
<tr>
<td>10.0%</td>
<td>1,883</td>
<td>-0.24</td>
</tr>
<tr>
<td>15.0%</td>
<td>1,906</td>
<td>-0.25</td>
</tr>
<tr>
<td>25.0%</td>
<td>1,958</td>
<td>-0.26</td>
</tr>
</tbody>
</table>

\(^a\) “Uniform price reduction” achieved by reducing price in every arm of each plan by the percent shown in the table.

\(^b\) Implied “elasticity” calculated as ratio of percent change in spending (relative to the baseline) to percent change in price (relative to the baseline).
Comparing static and dynamic models

- **Key point:**
  - Both models match bunching estimates
  - Deliver different elasticity estimates: dynamic model elasticity about five times larger than static model (-0.25 vs -0.05)

- Models are not vertically rankable
  - Saez model
    - Simple and transparent mapping from descriptive fact to economic object of interest
    - Relatedly, can be implemented quickly and easily
  - Dynamic model:
    - More computationally challenging and time consuming to implement
    - More "black box" relationship between underlying data objects and economic objects of interest
    - Allows us to account for potentially important economic forces that Saez-style model abstracts from (e.g. anticipatory responses)
Models are conceptually different (and non-nested)

- Saez model is frictionless
  - Implementation allows for some frictions since bunching is measured with some bandwidth (vs kink)
    - Will miss any behavioral response outside the bandwidth used to measure bunching
  - Dynamic model allows lumpiness by modeling a discrete series of (weekly) health shocks and purchase decisions

- Saez model is static
  - All uncertainty realized prior to any spending decision
  - Dynamic model: individuals make sequential purchase decisions throughout the year as information is revealed
    - Potential anticipatory behavior - set of people "at risk" of bunching may be endogeneously affected by presence of kink
    - Previous work suggests existence and importance of anticipatory behavior (Aron-Dine et al. 2016, EFS 2015)
Explored what features may be quantitatively important

- Considered two "restricted" versions of the "full" dynamic model
  - "No dynamics model": assume no discounting or uncertainty; continue to allow for frictions in the form of lumpy spending
  - "No discounting model": allows for lumpiness in spending and also uncertainty in timing and nature of shocks throughout year but imposes $\delta = 1$
    - All dynamic behavior due to uncertainty about future, rather than to time preferences
  - Estimate each model, again fitting bunching patterns
    - Note: distinct from "comparative statics" (without re-estimation)
  - Elasticity results suggest allowing for lumpiness and uncertainty important; discounting less so
    - Full dynamics: -0.25 (vs Saez -0.05)
    - "No dynamics": -0.13
    - "No discounting": -0.22
Challenges and opportunities

- Current "frontier" of research
  - Focus on compelling evidence of behavioral response
  - Map the reduced form / compelling evidence to an economic object of interest

- Key point: mapping choice can be consequential
  - Illustrated here in context of bunching estimators, but point is more general
  - e.g. RAND HIE and recovering "the price elasticity" (Aron-Dine et al. 2013 JEP)

- Path forward?
  - Find the right model?
  - Find the right question?
    - Are there underlying primitives to recover?
Aside: two comments on paper writing

- A fun (to write and read) paper structure
  - "Reduced form" results (something to hang your hat on)
    - Evidence of forward looking behavior (Aron-Dine et al. 2015)
    - Bunching behavior at kinks (Einav et al. 2015)
  - Quantification / Interpretation / Counterfactual Analysis - Usually requires additional modeling assumptions ("structural").
    - Arguably more speculative but more interesting

- Re-using designs for different questions - more bang for the buck (greater ratio of thinking to doing!). Examples:
  - Einav-Finkelstein-et-al.: 4 donut hole papers (and counting?).
  - Handel et al. (AER 2013) and Handel-Hendal-Whinston (EMA 2015)
  - Doyle et al. (JPE 2015) and Hull (JMP 2017)
Moral hazard in HI: Summary

- Some abuse / confusion about the term (ex ante, ex post etc)
- Does moral hazard exist?
  - RAND and Oregon experiment
- Nature of moral hazard
  - Selection on moral hazard ("selection on gains")
  - Response to non-linear contract
Some areas ripe for work

- Other aspects of nature of moral hazard
  - Inter-temporal dimension / multi-year contracts
  - "Source" of moral hazard - provider vs patient
  - Price shopping (Brot-Goldberg et al. QJE 2017)
  - Ex ante moral hazard (Spenkuch JHE 2012)

- Producer vs. consumer incentives
  - For reducing health care spending, have we been looking under the wrong lampost?

- Welfare implications of MH - tricky but important
Recall classic textbook welfare analysis of moral hazard

Pay $100 per visit: No consumption smoothing
Pay $0 per visit: lots of moral hazard (why not consume infinite doctor visits)?
Optimal insurance is a tradeoff: balancing consumption smoothing and moral hazard → partial insurance

Source: Gruber textbook
Departures from this framework I: Dynamics

- Framework is static
- Health insurance may also induce development and adoption of new technologies
  - e.g. Finkelstein (QJE 2007) on Medicare Introduction
  - More of this covered in 473
  - But FYI: missing welfare analysis of induced innovation...
“Overconsumption” of medical care: WTP for marginal unit of care is less than its social cost

- But what insurance does is subsidize the price of medical care. So that only maps immediately to welfare if health care is priced at its social marginal cost.

- In many settings, $p_{hc} \gg smc_{hc}$

Example I: Prescription drugs: SMC of drug production $\sim 0$

- Drug price distorted above MC due to patents
- Classic patent analysis: tradeoff of dynamic vs static efficiency
- Insurance as two part pricing undoes that inefficiency and promotes socially beneficial increase in consumption? (Lakadwalla and Sood JPubEc 2009)
In many settings, $p_{hc} \gg smc_{hc}$

Example II: Use of ER

- Popular view that lack of insurance $\rightarrow$ “inefficient” use of expensive ER vs lower priced doctor’s offices / clinics
- (Empirical evidence? see Oregon HIE...)

But SMC of doc time for non emergency treatment may be close to 0 (despite high price)

- Have to staff ER in case of emergency; what else are they doing at 3am?
- Of course also consider opportunity cost of uninsured’s time...
We’re not focused on the “right” behavioral response for welfare.

The (socially) “costly” impact of insurance on behavior may be:

- via impact on premiums and hence insurance demand (so price elasticity of demand for insurance is relevant)
- via effect on expected market size / innovation (is this welfare increasing or welfare decreasing?)

Not yet formalized or analyzed...
Departures from this framework III: “Behavioral”

- Individuals may not consume the privately optimal level of care absent insurance
  - May under-consume care (particularly preventive care) because of myopia, lack of understanding of long run benefits etc.

- In this second best world, subsidizing price of care through insurance and inducing increased utilization may be welfare increasing
  - Baicker, Mullainathan, Schwartzstein (2015 QJE) “Behavioral hazard in health insurance”
Some methodological points: Summary

- Experiments great for testing nulls
  - Issues in experimental design

- Limitations of (some) experiments
  - Endogeneity eliminated by an experiment may be important economically (selection on mh)
  - Recovering an economic object of interest - economic models an important complement (e.g. non linear contracts - RAND HIE)

- (Some) comments on modeling
  - Choice of model is consequential ("sufficient" statistic is "sufficient" conditional on a model)
  - Sensible counterfactuals (don’t go too far out of sample)
Complementarities between RF and “structural” work

- Credibility / transparency in (good) RF estimates and presentation
  - Always good to see the existence of phenomenon of interest “as close to the data” as possible
- RF work can help guide (consequential) modeling choices - e.g. static vs. dynamic behavioral model
- Values of modeling
  - Sometimes can’t run an experiment (e.g. merger analysis; GE effects of health insurance...)
  - Counterfactual analysis (can’t run an experiment for every permutation of a question)
    - But important to be sensible / don’t go “too far” out of sample
- Welfare analysis - need utility model