HHS Public Access

Author manuscript

Am Econ J Appl Econ. Author manuscript; available in PMC 2016 July 14.

Published in final edited form as:

Am Econ J Appl Econ. 2016 April ; 8(2): 195–224. doi:10.1257/app.20150131.

Beyond Statistics: The Economic Content of Risk Scores

Liran Einav,
Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305-6072 and NBER

Amy Finkelstein,
Department of Economics, MIT, 50 Memorial Drive, Cambrdige, MA 02142-1347 and NBER

Raymond Kluender, and
Department of Economics, MIT, 50 Memorial Drive, Cambrdige, MA 02142-1347

Paul Schrimpf
Department of Economics, The University of British Columbia, 997-1873 East Mall, Vancouver B.C. Canada V6T 1Z1

Liran Einav: leinav@stanford.edu; Amy Finkelstein: ank@mit.edu; Raymond Kluender: kluender@mit.edu; Paul Schrimpf: schrimpf@mail.ubc.ca

Abstract

"Big data" and statistical techniques to score potential transactions have transformed insurance and credit markets. In this paper, we observe that these widely-used statistical scores summarize a much richer heterogeneity, and may be endogenous to the context in which they get applied. We demonstrate this point empirically using data from Medicare Part D, showing that risk scores confound underlying health and endogenous spending response to insurance. We then illustrate theoretically that when individuals have heterogeneous behavioral responses to contracts, strategic incentives for cream skimming can still exist, even in the presence of "perfect" risk scoring under a given contract.

Over the last two decades, many markets have been transformed by the increased use of information technology, "big data," and statistical techniques. Credit and insurance markets are two leading examples (Edelberg 1996; Brown et al. 2014; Einav, Jenkins, and Levin 2013b). Nowadays, it is almost impossible to obtain credit or insurance without providing a long list of personalized information, which private lenders and insurance providers use to provide individually-customized prices or contracts. The government also actively uses such "risk scores" to regulate and reimburse private providers. In credit markets, for example, the government uses FICO scores–designed to predict an individual's default risk – to regulate the availability and terms of private mortgages. In the context of health insurance, the government uses health spending risk scores – designed to predict an individual's medical spending – to set Medicare reimbursement rates for private insurers. The state Health Insurance Exchanges created by the 2010 Affordable Care Act have increased interest in

---

Correspondence to: Paul Schrimpf, schrimpf@mail.ubc.ca.

how best to design and use health spending risk scores in regulating government reimbursement of private insurance offered on the exchanges.

These types of scoring algorithms predominantly rely on widely available predictive modeling techniques, which are commonly used in statistics and computer science. Typically one begins with a large individual-level data set that contains a key outcome one is trying to predict (such as medical spending or default on a loan) and a long and rich list of potential regressors; the creators of the algorithm then deploy state-of-the-art predictive models to select regressors and obtain the “best” predictive model.

Our paper is motivated by the observation that the outcomes that risk scores are designed to predict, such as loan default or medical spending, are, naturally, economic as well as statistical objects. While these outcomes may depend on certain individual characteristics that are invariant to the contract an individual chooses, they may also be affected by individual behavior. This behavior may well be endogenous to the context. Crucially, the behavioral response to the context may itself be heterogeneous across individuals.

The unidimensional risk score, however, is not designed to distinguish differences across individuals in their contract-invariant individual characteristics from differences in their behavioral response to another contract. Therefore, public reimbursement based on existing risk scores can give private providers incentives to cream-skim customers whose behavior under the contract is likely to make them lower cost than the risk score would predict. This suggests that risk scoring should be treated as a partially economic, rather than purely statistical, object, with properties that may need to be customized to a particular context and objective.

While this point is quite general, we develop and illustrate it in the particular context of the health spending risk scores that Medicare assigns to Medicare beneficiaries. These risk scores predict Medicare spending in traditional fee for service Medicare as a function of the beneficiaries’ demographics and medical diagnoses in the previous year. They are used, among other things, to set reimbursement rates to private providers of different Medicare Part D prescription drug insurance plans, and to private providers of Medicare Advantage (MA) plans, privately run managed care plans that nowadays enroll almost a third of Medicare beneficiaries.

Risk scoring is a natural way for the government to try to prevent - or at least reduce - cherry picking of low cost individuals by private firms (Newhouse 1996**).** By adjusting reimbursement based on observable individual characteristics that correlate with the individual’s cost to the private firm, the government can try to reduce these cream-skimming incentives. The key point of departure of this paper is to consider the possibility that an individual’s cost to the provider partly reflects the individual’s behavioral response to the provider’s contract, and that this behavioral response may differ across individuals – just as the standard, statistical, cost-related characteristics of the individual may differ – but will not be captured by current risk scoring practices.

We illustrate these points empirically in the specific context of the Medicare Part D prescription drug program. The introduction of prescription drug coverage in 2006, which

constituted the largest expansion of benefits in Medicare's half-century of existence, accounts for about 11 percent of total Medicare spending (Kaiser Family Foundation 2012a, 2012b). Medicare Part D enrollees can choose among different prescription drug plans offered by private insurers. Medicare reimburses private plans as a function of the "Part D risk scores" for their enrollees; these predict a beneficiary's prescription drug spending as a function of demographics and prior medical diagnoses.

We describe the data and the empirical strategy in Section I. Our research design exploits the famous "donut hole," or "gap," in Part D coverage, within which insurance becomes discontinuously much less generous at the margin. We previously used this research design, together with detailed micro data on prescription drug claims of Medicare Part D beneficiaries from 2007 to 2009, to help identify the behavioral response of drug utilization to cost-sharing (Einav, Finkelstein, and Schrimpf 2015). Here, in Section II, we use the same machinery to provide graphical evidence on two distinct, new results which are the focus of the current paper.

First, we show that two dimensions of heterogeneity are present and visible in the data. Unremarkably, we document heterogeneity in health; there are clear and expected relationships between annual drug spending and various individual characteristics, such as age or the presence of specific chronic conditions. More interestingly, we also document heterogeneity in the individual's utilization response to the contract. Specifically, we find that those who reduce their drug spending on the margin in response to the kink in the budget set created by the donut hole are more likely to be male, younger, and healthier, presumably reflecting their greater flexibility to forego drug purchases when the price increases.

Our second key empirical finding is that current risk scores do not capture this second dimension of heterogeneity. Risk scores increase smoothly with annual spending, but without exhibiting any noticeable pattern around the kink. This illustrates that the current risk scores do not capture differences across individuals in their behavioral response to consumer cost-sharing. This is by design; the creation of risk scores is currently treated as a statistical exercise, designed to generate the best predictor of an individual's costs under the observed environment, rather than an economic model of what their costs might be under an alternative contract.

In Section III we consider theoretically some of the potential implications of these empirical findings. In particular, we show that when individuals are heterogeneous not only in their underlying health but also in their utilization response to a health insurance contract, risk scores that are "perfect" in the statistical sense of capturing all residual heterogeneity under a given contract can still create cream-skimming incentives for private providers. We stop short of the more ambitious undertaking of estimating an equilibrium model of supply and demand for different health insurance contracts that would allow us to provide a empirical assessment of the implications of observed and alternative risk scoring for equilibrium cream-skimming. This is a natural direction for further work.

Our paper contributes to a large literature on risk adjustment in health insurance markets, which was reviewed in Van de Ven and Ellis (2000) and Ellis (2008). Much of this literature has focused on predictive (statistical) modeling. A recent focus has been on the fact that risk adjustment relies on diagnoses recorded in clinical and administrative records, which may reflect differences in diagnostic and treatment practices across insurers and providers, in addition to underlying health (Song et al. 2010). There has also been attention to the incentives for cream-skimming and "gaming" that such risk scores provide. However, the focus of the existing analysis of cream-skimming is that in the presence of *imperfect* prediction of individual risk, private insurers have an incentive to try to attract ("cream skim") individuals who, given their predicted risk, have (imperfectly priced) characteristics that (in expectation) generate lower realized risk.[1] Glazer and McGuire (2000) provide the classic theoretical framework for this type of strategic cream-skimming; they show that in the presence of imperfect risk adjustment, the relationship between reimbursement and predicted risk should be amplified in order to minimize cream-skimming incentives. Empirically, two recent papers – Brown et al. (2014) and Newhouse et al. (2012) – use a similar framework to examine providers' strategic response to imperfect risk scoring in the context of Medicare Advantage.

The key distinction between the current paper and this existing risk-adjustment literature is that the latter is focused on the problem of imperfect risk adjustment in an environment with unidimensional heterogeneity. In this setting, a "perfect" (in a statistical sense) risk prediction model would eliminate cream-skimming incentives, and the market would operate like any traditional product market. Although the assumption of imperfect risk adjustment is a natural one, the cream skimming incentives considered by the existing literature could, at least in principle, be eliminated with rich enough data and sophisticated enough statistical modeling, thus obviating the need for economic models. In contrast, our focus is on a different challenge in using risk scores, a challenge that cannot – even in principle – be solved with rich enough data and perfect scoring. Our key observation is that the outcome the risk score attempts to predict is partially determined by individuals behavioral choices, and these may vary with the contract. Therefore, even perfect prediction of the outcome under a given contract ("perfect" risk adjustment in the sense of the prior literature) would not suffice, and an economic model of behavior is needed to think about optimal reimbursement policy when coverage contracts differ.

Our paper also relates to a large "moral hazard" literature in health economics on the impact of insurance contracts on medical care use in general, and more specifically to a smaller "moral hazard" literature in the context of Medicare Part D (Duggan and Scott Morton 2010; Einav, Finkelstein, and Schrimpf 2015). In contrast to most of this literature, which has focused on average behavioral responses, our focus here is on the potential individual heterogeneity in the behavioral response and its implications (in this case, for risk scoring). In this sense, our paper relates to previous work analyzing the role of heterogeneity in the behavioral response in contributing to adverse selection in an employer-provided health insurance setting (Einav et al. 2013a; Shepard 2015).

[1] In addition, another branch of the literature notes that insurers also have an incentive to "upcode" the individual components that enter into the risk adjustment formula to increase a given individual's reimbursement (Dafny 2005; Geruso and Layton 2014).

I. Data and Empirical Strategy

The central premise behind our analysis of risk scoring is that an individual's medical spending is determined by both underlying health and economic choices, *both of which* are potentially heterogeneous across individuals. We demonstrate this simply and visually, using data from Medicare Part D, the prescription drug coverage component of Medicare that was added in 2006. As of November 2012, 32 million people (about 60 percent of Medicare beneficiaries) were enrolled in Part D, with expenditures projected to be $60 billion in 2013, or about 11 percent of total Medicare spending (Kaiser Family Foundation 2012a,b). Unlike Medicare Parts A and B for hospital and doctor coverage, which provide a uniform public insurance package for all enrollees (except those who select into the managed care option, Medicare Advantage), private insurance companies offer various Medicare Part D contracts, and are reimbursed by Medicare as a function of their enrollees' risk scores.

While the exact features of the plans offered vary, they are all based around a standard design, shown in Figure 1. The discontinuous increase in the out-of-pocket price individuals face when they cross into the "donut hole" (or "gap"; see Figure 1) provides the research design that enables us to detect the responsiveness of individuals to the out-of-pocket price. As discussed in more detail in our earlier work (Einav, Finkelstein, and Schrimpf 2015), standard price theory suggests that individuals' annual spending will "bunch" around the convex kink in the budget set created by the gap. Importantly, the extent of bunching should be greater and more noticeable for individuals who are associated with greater price sensitivity.

A. Data

We use data on a 20 percent random sample of all Medicare part D beneficiaries over the years 2007–2009. The data include basic demographic information (such as age and gender) and detailed information on the cost-sharing characteristics of each beneficiary's prescription drug plan. We also observe detailed, claim-level information on our beneficiaries' Medicare utilization from 2006–2010. This includes both prescription drug purchases (covered under Medicare Part D), as well as inpatient, emergency room, and outpatient (non emergency) use (covered under Medicare Part A and B). Finally, we observe mortality through 2010.

We use the same sample that we used in Einav, Finkelstein, and Schrimpf (2015) with the additional restriction that beneficiaries were enrolled in Medicare in the previous year. It excludes various groups of beneficiaries for whom the empirical strategy is not applicable, such as individuals in Medicare Advantage and certain low income individuals for whom the basic benefit design we are studying does not apply. We also limit the analysis to individuals aged 65 and over. See Einav, Finkelstein, and Schrimpf (2015) for a complete discussion and details of the sample.

Our analysis sample consists of 3.7 million beneficiary-years (1.6 million unique beneficiaries) during the years 2007–2009. The average age in our sample is 76, and about two thirds of the individuals are females. Average annual, per-beneficiary drug spending is just over $1,900 dollars; on average, approximately $800 are paid out of pocket. Spending is

very right skewed: about 5 percent of beneficiaries have no annual drug spending, median spending is about $1,400, and the 90th percentile is about $4,000.

As noted, there is variation in the insurance contract design, including the extent of any coverage in the gap. On average, a beneficiary in our sample faces a 60 cent increase in out-of-pocket spending for every dollar spent, as his annual spending hits the kink. Specifically, we estimate that average out-of-pocket cost sharing in our sample is 34 cents on the dollar below the kink and 93 cents on the dollar in the gap. The exact location of the kink, as a function of total drug spending, also varies across observations in our sample depending on the year, but on average it hits at roughly the 75th percentile of the drug spending distribution.

We use the Centers of Medicare and Medicaid's Services (CMS) 2012 RxHCC risk adjustment model which is designed to predict a beneficiary's prescription drug spending in year t as a function of their inpatient and outpatient diagnosis data from year $t-1$, as well as demographic information (including gender, age, and the original reason for entitlement to Medicare). The model takes more than 14,000 disease (ICD-9) codes and aggregates them into 167 "condition categories." The model imposes a hierarchy on the condition categories in order to group them together into clinically meaningful diagnoses which predict costs. These final "hierarchical condition categories" (HCCs) are the level of diagnoses used to specify the risk score model, out of which the model selects those HCCs that are found to be most predictive of drug spending.

The final version of the risk adjustment model uses an additively separable predictive model, which relies on risk-score coefficients that are associated with 78 selected HCCs from year $t-1$, a gender dummy variable, dummy variables for each five-year age bin, and a dummy variable associated with the original reason for Medicare entitlement. Predicted year-t drug spending is then computed by simply adding up all the risk-score coefficients that are associated with those dummy variables that are "turned on" for a given beneficiary. For an individual's first year in Medicare (typically when he turns 65), when diagnosis information from the previous year is not available, a new-enrollee risk score is generated solely on the basis of the demographic information. All predictions are normalized by the prediction for a representative Part D beneficiary, who is assigned a risk score of 1.[2]

Private insurers submit annual bids to CMS for their projected costs of covering a Medicare Part D beneficiary with a risk score of 1 (excluding catastrophic coverage provided by CMS). CMS calculates the market's average bid and multiplies it by an individual's risk score to determine the direct subsidy paid to the private insurer. A similar methodology is used to reimburse private insurers providing Medicare Advantage coverage. Our sample average Part D risk score is 0.88, indicating that they are 12 percent less expensive to cover than the representative Part D beneficiary.

[2]CMS' risk adjustment models for Medicare Advantage operate in a similar way, except that they are designed to predict overall Medicare spending (not just drug spending), and include variables for Medicaid eligibility and a different selection of HCCs.

B. Empirical strategy

We use simple graphical illustrations of the average characteristics of individuals as a function of total annual drug spending to illustrate the two dimensions of heterogeneity that are our focus: heterogeneity in health and heterogeneity in the behavioral response to the contract. Monotonic patterns of individual average demographic characteristics and diagnoses as a function of total drug spending show the heterogeneity in health that is the focus of current risk scoring. Sharp deviations from these monotonic patterns around the kink in the budget set illustrate heterogeneity in the behavioral response to the contract.

Our strategy for detecting heterogeneity in the behavioral response to the contract builds on our prior work detecting the average behavioral response to the contract from the fact that individuals bunch at the kink. Figure 2 replicates this prior bunching analysis from Einav, Finkelstein, and Schrimpf (2015). Because the kink location has changed from year to year (from $2,400 in 2007, to $2,510 in 2008, and $2,700 in 2009), in all our figures we normalize annual spending by the kink location. We plot the distribution of (normalized) annual spending (in $20 bins) for individuals whose spending is within $2,000 of the kink (on either side). This constitutes 66 percent of our sample. The presence of significant "excess mass," or "bunching" of annual spending levels around the convex kink in the budget set (that is created by the gap) indicates the presence of a behavioral response to the increased consumer cost-sharing at the kink. The response to the kink is apparent: there is a noticeable spike in the distribution of annual spending around the kink. In Einav, Finkelstein, and Schrimpf (2015) we presented this result in greater detail, showing how the location of the spike moves as the kink location changes from year to year and analyzing the types of drugs that individuals appear to stop purchasing when they slow down their drug utilization and "bunch" at the kink.

In this paper, we focus instead on heterogeneity in the responsiveness across different groups of individuals, interpreting greater bunching around the kink for different populations as reflecting greater demand sensitivity to out-of-pocket price. We identify heterogeneity in this behavioral response by documenting sharp changes in the presence of specific individual characteristics around the kink. An individual characteristic (such as being male or having a particular health condition) that is over-represented among individuals around the kink indicates that individuals with this characteristic have a greater behavioral response to the kink (and are therefore over-represented around the kink). Conversely, a characteristic which is under-represented among individuals whose spending is around the kink suggests that individuals with this characteristic are less responsive to the contract.

II. Results

A. Evidence of two-dimensional heterogeneity

In Figure 3 we present several summary statistics on the beneficiaries, by their spending bin. Summary statistics are mostly monotone in annual spending in expected ways: individuals who spend more are older and sicker. This illustrates the heterogeneity in underlying health that current risk scoring is designed to capture.

The novel observation in Figure 3, however, is not the monotone pattern, but rather the noticeable non-monotone pattern around the kink for some of the individual attributes. Recall that beneficiaries bunch around the kink (see Figure 2). Therefore, the distinct demographics around the kink location capture the distinct demographics of those beneficiaries who are more likely to bunch around the kink, or in other words, the more price sensitive individuals.

Figure 3(a) shows the patterns of various demographics: age (top panel) and gender (bottom panel). Average age is generally monotonically increasing in annual spending, but there is a sharp dip in average age at the kink. Likewise, there is a sharp dip in the probability of being female right around the kink. That is, we find that younger males are more likely to bunch around the kink, which we interpret as evidence that they are more price elastic.

Figure 3(b) examines the frequency of a handful of selected health conditions (HCCs) that enter the risk adjustment formula. The frequency of each condition is generally increasing monotonically in annual spending, reflecting the fact that individuals with a given condition spend, on average, more. However, for some of the conditions there appear to be some noticeable non-monotone patterns around the kink. In particular, the probability of depression and congestive heart failure appear to dip slightly around the kink, suggesting that these conditions are associated with a lower drug use response to price. By contrast, some other health conditions – such as coronary artery disease or chronic obstructive pulmonary disease (COPD) and asthma – are not associated with any noticeable pattern around the kink, suggesting that these conditions are not associated with a price response.

Finally, Figure 3(c) examines mortality and non-drug healthcare utilization in the subsequent calendar year (year $t + 1$) as a function of annual drug spending in the current year (year t). Specifically we look at mortality for the full year ($t + 1$) and emergency room (ER) visits, inpatient admissions, and (non-ER) outpatient visits during January to June of $t + 1$. Again, there is a natural monotone pattern: individuals who spend more on drugs in year t are presumably sicker, and are therefore associated with greater non-drug healthcare utilization and greater mortality in the subsequent year. However, once again, there are distinct non-monotonicities around the kink. The probability of death in year $t + 1$ drops sharply for those who are around the kink. The figure also shows some evidence that individuals who are approaching the kink in year t are less likely to use other medical care (emergency room, non-emergency outpatient care, or inpatient care) in the first six months of year $t + 1$. The effect on the use of other medical care is weaker, as it is not based on a non-monotone pattern around the kink, but only relies on the local change in slope around the kink.

The interpretation of Figure 3(c) is a little more subtle. We interpret it as additional evidence that the individuals who are more price sensitive and therefore bunch at the kink are also healthier, as measured by their subsequent (non-drug) healthcare use and mortality rate.[3] Of course, since subsequent health and healthcare use are potentially directly affected by current drug utilization decisions, it is possible that these results reflect a causal treatment

[3]Interpreting these patterns as reflecting heterogeneity in underlying health (rather than an effect of drug spending on subsequent health) is also consistent with a related finding by Joyce, Zissimopoulos, and Goldman (2013), that the decline in drug purchases for diabetics who entered the gap is not associated with increased use of medical services.

effect of drug utilization (which varies across individuals depending on their price sensitivity) on health.

B. Risk scores do not capture both dimensions

Figure 4 illustrates the other key point of the paper: the current risk scores do not capture heterogeneity across individuals in their behavioral response to the contract. Figure 4(a) presents a similar analysis to those shown in Figure 3, except that we now summarize the risk scores that Medicare Part D assigns these individuals.

It shows an overall smooth, monotone pattern of average Part D risk score, reflecting (by design) that individuals with higher average spending have higher risk scores. Strikingly, however, the individuals around the kink (i.e. those who are more likely to be "bunchers") appear to follow the increasing pattern of health spending risk scores, *without any visible pattern around the kink.* That is, the health spending risk score predicts well spending under the observed contract – as it is designed to do – without capturing (by design) the fact that some of this spending reflects a price response, which is endogenous to the coverage contract.

There are two different possible ways to reconcile the evidence in Figure 3 that healthier individuals are more likely to bunch at the kink, with the evidence in Figure 4(a) that the Part D risk scores do not reflect any lower predicted spending for individuals at the kink. One is that the demographics that change sharply around the kink in Figure 3 are not quantitatively important in generating risk scores, and thus do not affect much the average risk scores in Figure 4. The other is that there are other components of the risk score that move in the opposite direction around the kink, thus offsetting the patterns presented in Figure 3. The interpretation does not affect our main point, which is that the current risk scores do not capture differences in spending that arise from differences in the behavioral response to the contract.

Our analysis suggests that the monotone pattern of risk scores through the kink in Figure 4(a) in fact reflects offsetting effects: the characteristics that exhibit greater propensity around the kink have a noticeable effect on risk scores, but they are offset by other characteristics that display the opposite pattern at the kink. To determine this, we generated a prediction of the value of each component of the risk score around the kink, using its values below the kink. That is, for each component of the risk score (age category, gender, and each specific HCC), we ran a linear regression based on the relationship between spending and that component of the risk score in the (−$2,000,−$200) range and then, using the estimated regression, generated predictions for that component in the (−$200,+$200) range. We then split the individual components into those that exhibited excess bunching around the kink (that is, those whose actual values in the (−$200,+$200) range was on average higher than the corresponding prediction in this range) and those that exhibited a dip around the kink (that is, those whose actual values in the (−$200,+$200) range was on average lower than the corresponding prediction in this range). We then produced two different versions of "predicted" overall risk scores. In one, we used the predicted values for those components that exhibit bunching around the kink and the actual values for the rest. In the other, we used the predicted values for those components that exhibit dips around the kink, and the actual

values for the rest. If the components that exhibit bunching and dipping around the kink do not do so in a manner that is quantitatively important for the risk score, we would expect these two different versions of the predicted risk scores to lie very close to each other (and to the actual risk score) around the kink. Figure 4(b) shows that, in fact, the two different versions of the predicted risk scores lie apart from each other on either side of the actual risk score. This suggests that the patterns for individual components around the kink are quantitatively important, but offset each other. Table 1 shows the underlying components that are most important in affecting the positive and negative shifts in risk scores around the kink.

C. A quantitative assessment of heterogeneity in the behavioral response

These findings document that there is heterogeneity in the behavioral response to cost-sharing that is not captured by the risk score. A natural question is whether this has quantitatively important implications, not only at the kink (which is the focus of our research design) but more generally throughout the non-linear budget set created by the contract. To answer this, one needs to develop and estimate a behavioral model of healthcare spending under different contract designs, and investigate the extent to which an individual's ranking in the spending distribution is the same under alternative contracts.

As we discuss in the next section, given that insurers could apply any non-linear transformation to a given set of risk scores, the key role of a risk scoring system is in its ordinal ranking of individuals in term of their expected risk. Thus, one way to assess the quantitative importance of the heterogeneity in the behavioral response to the contract design is to quantify the extent to which individuals' position in the population's expected risk distribution (that is, in the contract-specific risk score distribution) gets reshuffled as they move across contracts. If heterogeneity in the behavioral response to the contract is not quantitatively important, individuals ranking would remain relatively stable across contracts.

The research design we have used thus far is not sufficient for such an exercise, as it doesn't attempt to model health utilization behavior under alternative contract designs. However, we can shed some light on this question by using the model of healthcare utilization that we developed and estimated in our earlier, related work (Einav, Finkelstein, and Schrimpf 2015). There we develop a complete, dynamic behavioral model of the individual's drug purchasing decisions under non-linear coverage contract, allowing for heterogeneity across individuals in both health risk and in the spending response to coverage. In our prior paper, we estimated the model's parameters using the same data set as in the current paper; the "bunching at the kink" we have examined here is one of the elements used for identification in estimating that model. In Appendix Table A1, we use the model estimates to predict spending under the standard contract shown in Figure 1, and then predict spending (for the same set of individuals and associated sequences of health shocks) for two alternative, counterfactual contracts. One is a "filled gap" contract that eliminates the gap by providing pre-gap cost sharing up to the catastrophic coverage limit; the Affordable Care Act aims to make this type of contract become the standard contract by 2020. A second contract is actuarially equivalent to the standard contract shown in Figure 1, but it eliminates the

deductible in the standard contract, and instead offers higher cost-sharing (of 38.9 cents for each dollar, instead of 25 cents) for spending below the gap.

Appendix Table A1 shows the extent to which individuals' ranking in the spending distribution changes under alternative contracts, relative to their spending percentile under the standard contract presented in Figure 1. We split individuals into ventiles of spending under the standard contract, and report (for each ventile) the share of the individuals who are expected to stay within the same ventile, and the share that moves to other spending ventiles (up or down) under the alternative contract. Of course, expected spending is primarily driven by expected health, so the vast majority of individuals remain within the same spending ventile. Yet, as Appendix Table A1 shows, a non-negligible share of individuals get reshuffled in their ranking, especially in the region where the price changes. For example, Panel A shows that "filling" the coverage gap leads to a fair amount of "reshuffling" in the expected spending of high spenders, who are those who are most affected by the change in coverage in the gap. Panel B shows that eliminating the deductible leads to a fair amount of "reshuffling" in the expected spending of low spenders, who are those are likely to be affected the most by the deductible.

This exercise illustrates the perils of using predicted spending under one contract to generate predicted spending (i.e. risk scores) under alternative contracts. The results in Appendix Table A1 show that if one generated a risk score based on spending under the standard contract and used it to predict spending under alternative contracts, the generated risk scores would be highly imprecise for those regions of spending that are most affected by the alternative contract.

A natural follow-up question is how important this imprecision of risk scores would be for equilibrium cream skimming and market outcomes. Answering this question empirically would require not only a model of demand (that is, of health care utilization), but also a model of competition and pricing, which is beyond the scope of the current paper. Instead, in the next section we briefly explore theoretically some potential implications for cream skimming incentives and optimal risk adjustment.

III. Implications

The evidence in the preceding section established that Medicare's risk scores reflect expected medical spending under the *existing* benefit design, and that this one-dimensional score hides a richer heterogeneity that determines medical spending. The multi-dimensional heterogeneity that determines medical spending reflects heterogenous price sensitivity as well as heterogeneous health. In this final section, we illustrate theoretically how reimbursement based on the (unidimensional) risk score can create incentives for private providers to cream-skim customers whose behavior under their private contract is likely to make them lower cost than the risk score would predict (as it is based on behavior under an alternative contract). Importantly, this incentive for cream-skimming cannot be combatted by richer statistical modeling of utilization behavior under a given contract.

Cream-skimming by providers of individuals who are lower cost than their risk score would suggest is the classic problem analyzed by theoretical and empirical work on risk scoring (Glazer and McGuire 2000; Newhouse et al. 2012; Brown et al. 2014). In these existing analyses, if the risk scoring is "perfect" in a statistical sense (i.e. conditional on the risk score, there are no residual characteristics of the individual that predict spending under a given contract) the cream-skimming problem goes away.[4]

However, once we enrich the model to allow individuals to have heterogeneous behavioral responses to the coverage contract, strategic incentives for cream skimming can still exist, even in the presence of "perfect" risk scoring under a given contract. This is because individuals of the same risk score (and hence same predicted medical spending in one particular contract) may have different predicted medical spending under a different contract, due to their differential behavioral responses. Providers therefore can have an incentive to try to design contracts to attract those whose behavioral response to an alternative contract makes them lower expected cost than their risk score would predict.

A. A stylized framework

We start with a stylized model of healthcare utilization that emphasizes two forms of individual heterogeneity. The model is drawn from our earlier work (Einav et al. 2013a), which used a similar framework to examine a related question in a different setting.

An individual in the model is defined by a two-dimensional type, (γ, ω). In this definition, $\gamma \geq 0$ denotes the individual's underlying health and $\omega \geq 0$ denotes his price sensitivity of demand for medical care, or how responsive healthcare utilization choices are to insurance coverage. We focus on these two different dimensions that determine healthcare utilization.[5] We assume, in the spirit of the empirical results in the last section, that they cannot be separately distinguished by a unidimensional risk score.

For illustrative purposes, we consider individuals with a linear insurance coverage with a price of healthcare of $c \in [0; 1]$. That is, for every dollar of spending on healthcare, the individual pays c and the insurance provider pays $1 - c$.

Individuals make their healthcare utilization decision to maximize a tradeoff between health and money (residual income). Health depends on one's underlying health λ but is increasing in his monetized healthcare use (or medical spending) given by $m \geq 0$: Residual income $y(m)$ is decreasing in m at a rate that depends on the health insurance contract's c. More specifically, individual utility is given by

$$u(m;\lambda,\omega)=\left[(m-\lambda)-\frac{1}{2\omega}(m-\lambda)^2\right]+(y-c\cdot m). \quad (1)$$

[4]Interestingly, Brown et al. (2014) have recently highlighted that improvements in risk scoring that do not make the score "perfect" may, perversely, exacerbate cream-skimming.

[5]For concreteness, we model heterogeneity in the behavioral response to price, since this is what we document in the empirical results. In principle, one could derive similar analyses with behavioral heterogeneity in the response to other aspects of the contract, such as coverage of "star" hospitals, as in Shepard (2015).

The first component (in square brackets) captures the individual's health, which can be improved by greater utilization m. The second component captures residual income, which is given by the individual's income y net of his out-of-pocket spending $c \cdot m$.

Optimal medical spending m^* is chosen to maximize utility, that is by solving $\max_{m \geq 0} u(m; \lambda; \omega)$. This yields the first order condition

$$m^*(\lambda, \omega) = \lambda + \omega (1 - c). \quad (2)$$

Optimal medical spending depends on the individual's underlying health (λ), the out-of-pocket price of medical care (c), and the responsiveness of spending to that price (ω). Individual utility, given optimal medical spending, is then given by

$$u^*(\lambda, \omega) = u(m^*(\lambda, \omega); \lambda, \omega) = y - c \cdot \lambda + \frac{1}{2} (1 - c)^2 \omega. \quad (3)$$

To facilitate intuition of the model, consider the case of full coverage ($c = 0$) and no insurance ($c = 1$). In these cases, equation (2) indicates that the individual would spend $m^*_{c=1} = \lambda$ with no insurance and $m^*_{c=0} = \lambda + \omega$ with full insurance. Thus, individual medical spending depends on both a "level" term λ and a "slope" term ω. The individual has a level spending λ no matter what coverage he faces, but he then spends an additional ω when he has full coverage and does not need to pay for this additional utilization out of pocket. It is natural to view λ as related to the individual health, reflecting health conditions that need to get treated regardless of insurance coverage.

This ω term is typically referred to as "moral hazard" in the health economics literature (Pauly 1968). The structural interpretation of ω is not obvious. It likely reflects a combination of individual preferences over health and income as well as the nature of his health conditions and the extent to which treatment or type of treatment is optional or discretionary. Fortunately, the exact interpretation of ω is not crucial for the main point we try to advance in this paper, although our empirical work shed some light on the individual characteristics that correlate with ω. Rather, the key point is that two different economic objects – health λ and behavioral response to insurance contract ω – determine medical spending m.

B. Relation to empirical work

The empirical results shown in Figure 3 provided a simple illustration of one of the two key points of the paper: a one-dimensional summary measure is unlikely to be sufficient in describing individual types. The combination of generally monotone patterns in average individual characteristics as a function of annual drug spending and systematic non-monotonicity around the kink suggests that individuals vary not only in the health (λ) but also in their responsiveness to contract features like price (ω). Our results also indicate which types of individuals exhibit greater price sensitivity: those who "bunch" at the kink

are younger, more likely to be male, and appear healthier on many – but not all – measures of health conditions. These individuals appear to have greater exibility regarding prescription filling. The results therefore suggest that in our setting, at least for individuals around the kink, underlying health λ and price sensitivity ω are negatively correlated. The fact that the greater price responsiveness is more pronounced for some health measures but not for others underscores the richness of the potential underlying heterogeneity; our summary health measure λ itself likely encodes a richer heterogeneity, although in the context of our simple model a two-dimensional description of individuals would be sufficient.

This visual evidence of multi-dimensional heterogeneity complements our previous work where we estimated multi-dimensional heterogeneity in the context of a specific structural model of insurance demand, and explored its implications for consumer selection of insurance coverage with different levels of cost-sharing (Einav et al. 2013a). Here, the empirical evidence of heterogeneity along two dimensions – moral hazard type as well as health type – is relatively model-free (and arguably more compelling), coming directly from the data and the research design provided by the kink in the budget set. Our substantive focus here is also different. We examine whether this multi-dimensional heterogeneity is captured by current risk scoring models, and the resultant implications.

Figure 4 illustrated the other key empirical point in the paper: current risk score methods do not capture the behavioral responsiveness (ω) dimension of individual heterogeneity. This is by design, not only in the Medicare context but in most other risk adjustment models around the world (Ellis 2008). The Medicare risk scores attempt to predict m under a particular contract; they are constructed by employing a statistical predictive approach that attempts to find the best predictor of observed cost under Medicare Fee for Service. They therefore do not attempt to model how costs might vary across individuals under some *other* insurance contract in which individual behavior might differ from what is observed under Medicare Fee for Service, and which there might be heterogeneity across individuals in this behavioral response. Without an economic model of how costs under one contract may differ from those under another due to individual choices (and the potential heterogeneity in this difference across individuals), or a separate observed outcome that would allow the risk adjustment to observe or proxy for this second dimension of heterogeneity, it would be difficult to capture a second dimension of heterogeneity.

C. Cream-skimming incentives

We brie y explore some of the theoretical implications of the fact that current risk scores do not attempt to capture cost heterogeneity arising from heterogeneity in behavioral responses to a contract. The appendix provides a highly stylized theoretical example that illustrates how cream-skimming incentives can still exist in the presence of a “perfect” risk score under a given contract when individuals are heterogeneous in their behavioral responses to contracts. In the context of our model, a statistically “perfect” risk score means that there are no residual characteristics that predict an individual’s $\lambda_i + \omega_i$ conditional on their risk score. We briefly summarize the example and findings here.

We assume that the government offers a default contract, and consider a private (monopolist) insurer who offers a contract that competes to attract beneficiaries from the default contract.[6] We assume the default public coverage provides full insurance (i.e. $c = 0$), while the private plan has a technology to completely eliminate ω-related medical spending. Thus, in our stylized framework – see especially equations (2) and (3) – beneficiary i chooses medical spending level $\lambda_i + \omega_i$ under the public option, but only spends λ_i if enrolled by the private plan. The government reimburses the private insurer based on the risk scores of the beneficiaries it attracts. Because the government can only observe medical spending under its own, public contract, it can only set risk scores for beneficiaries and reimburse the private provider based on enrollees' medical spending under the public contract ($\lambda_i + \omega_i$). As Figures 3 and 4 illustrated empirically, this risk score does not distinguish between beneficiary costs arising from λ or from ω.

Under these assumptions, the socially efficient allocation is for everyone to be covered by the private plan, which eliminates inefficient, ω-related medical utilization. However, enrollees obtain greater utility in the less restrictive, public coverage, forcing the government to provide subsidies (potentially as a function of the risk score) to the private plan in order for it to have incentives to attract enrollees through lower premiums. This creates a tradeoff for government policy: greater subsidies create a more efficient allocation, but at the cost of higher public expenditures, and thus a greater social cost of public funds.

We analyze equilibrium selection into the private plan for a given government subsidy policy; a subsidy policy defines the government subsidy amount provided to the private plan for enrolling an individual with a given risk score. For a given subsidy policy, there are two conflicting selection pressures. On the one hand, higher-ω individuals are the most pro table for the private insurer to enroll and therefore the private insurer has an incentive to try to attract these individuals. On the other hand, higher-ω individuals are also the ones with the greatest incentive to remain under the public coverage.

The appendix presents a standard mechanism design solution to this conflict of incentives. It shows that, in equilibrium, the highest-ω individuals enroll in the private plan. These are the individuals for whom the efficiency benefits of the private plan are highest. However, the socially efficient outcome of having everyone enrolled in the private plan may not be the constrained optimum given the social cost of the public funds required to achieve it.

We can in fact solve for the optimal subsidy by the government as a function of the equilibrium solution to a given subsidy level. The optimal subsidy problem resembles a standard optimal pricing problem. Our discussion in the appendix highlights some of the key economic objects that determine the optimal subsidy, and which would need to be estimated in any particular application designed to analyze optimal risk adjustment in this environment.

[6] One loose, real-world analog might be the Medicare Advantage plans offered by private insurers who compete to attract beneficiaries from traditional fee-for-service-Medicare (Newhouse et al. 2012). Of course, for simplicity we have considered a monopolist competing against a (passive) public option, whereas oligopoly is presumably a more sensible assumption for the real-world Medicare Advantage plans.

IV. Conclusions

Our objective in this paper was to highlight the fact that risk scores that are commonly used in credit and insurance markets are not merely statistical objects, as they are generated by economic behavior. We illustrated this point empirically in the specific context of Medicare Part D, the public prescription drug insurance program that covers over 30 million individuals, and explored their implications theoretically. We exploited the famous “donut hole” where insurance becomes discontinuously much less generous at the margin.

Using this research design, we empirically illustrated two conceptual points. First, analyzing the average demographic and health characteristics of individuals as a function of annual drug spending, we showed that spending differences across individuals reflect not only heterogeneity in underlying health but also heterogeneity in the underlying behavioral response to the insurance contract. Second, we show that the current (statistical) risk scores – which are designed to predict spending under a given contract – do not capture this second dimension of heterogeneity.

In the second part of the paper, we use a highly stylized theoretical example to explore some of the potential implications of these findings for the standard use of risk scores, which is to predict outcomes out of sample under other contracts and use these predictions to set reimbursement rates. We showed that standard risk scoring can create incentives for private insurers to cream-skim individuals whose (unpriced) behavioral response to the contract they offer will make them lower cost than what is predicted by the risk score that was generated under a different contract. A key point is that, when there is heterogeneity in the behavioral response to the contract, these cream-skimming incentives can still exist even in the presence of “perfect” risk scoring under a given contract. While we thus illustrated, in the context of a specific theoretical example, the possibility of equilibrium selection on the behavioral response to different contracts, we did not establish its empirical existence or importance in a specific context. This would be a natural area for future work.

One potential response to the multi-dimensional heterogeneity we document is to move beyond a one-dimensional risk score and customize the risk score formula to the specific contracts to which it is applied. Risk scoring is currently conducted as a statistical prediction exercise of behavior under a given contract without any such adjustment, while our paper suggests the need to consider economic as well as statistical forces in designing risk scoring that is applied to other contracts. In practice, to do so would require empirical estimates of the heterogeneity in the behavioral response to alternative counterfactual contracts – perhaps of the flavor of those shown in Section II. Given the increased reliance on various models of risk scoring in many important markets, we view such analysis of optimal risk scoring in the presence of multi-dimensional heterogeneity in specific credit and insurance contexts to be an interesting – and potentially important – area for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Randy Ellis, Jonathan Gruber, Nathan Hendren, Ilyana Kuziemko, Robin Lee, Tom McGuire, Adam Sacarny, Julie Shi, Jonathan Skinner, and three anonymous referees for helpful comments. We gratefully acknowledge support from the National Institute on Aging (R01 AG032449).

References

Brown, Jason; Duggan, Mark; Kuziemko, Ilyana; Woolston, William. How does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program. American Economic Review. 2014; 104(10):3335–64.

Dafny, Leemore. How Do Hospitals Respond to Price Changes? American Economic Review. 2005; 95(5):1525–47.

Duggan, Mark; Morton, Fiona Scott. The Effect of Medicare Part D on Pharmaceutical Prices and Utilization. American Economic Review. 2010; 100(1):590–607.

Edelberg, Wendy. Risk-Based Pricing of Interest Rates for Consumer Loans. Journal of Monetary Economics. 2006; 53(8):2283–98.

Einav, Liran; Finkelstein, Amy; Ryan, Stephen; Schrimpf, Paul; Cullen, Mark. Selection on Moral Hazard in Health Insurance. American Economic Review. 2013a; 103(1):178–219. [PubMed: 24748682]

Einav, Liran; Finkelstein, Amy; Schrimpf, Paul. The Response of Drug Expenditure to Non-Linear Contract Design: Evidence from Medicare Part D. Quarterly Journal of Economics. 2015; 130(2): 841–99. [PubMed: 26769984]

Einav, Liran; Jenkins, Mark; Levin, Jonathan. The Impact of Credit Scoring on Consumer Lending. RAND Journal of Economics. 2013b; 44(2):249–74.

Ellis, Randall P. Risk Adjustment in Health Care Markets: Concepts and Applications. In: Lu, M.; Jonsson, E., editors. Financing Health Care: New Ideas for a Changing Society. Vol. Chapter 8. Wiley-VCH Verlag GmbH & Co; Weinheim, Germany: 2008.

Geruso, Michael; Layton, Timothy. Risk Selection, Risk adjustment, and Manipulable Medical Coding: Evidence from Medicare. Mimeo: UT Austin; 2014.

Glazer, Jacob; McGuire, Thomas G. Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care. American Economic Review. 2000; 90(4):1055–71.

Joyce, Geoffrey F.; Zissimopoulos, Julie; Goldman, Dana P. Digesting the Doughnut Hole. Journal of Health Economics. 2013; 32(6):1345–55. [PubMed: 24308883]

Kaiser Family Foundation. 2012a. http://www.kff.org/medicare/upload/7044-13.pdf

Kaiser Family Foundation. 2012b. http://www.kff.org/medicare/upload/1066-15.pdf

Medicare Payment Advisory Commission. Report to the Congress: Medicare Payment Policy. 2009.

Newhouse, Joseph P. Policy Watch: Medicare. Journal of Economic Perspectives. 1996; 10(3):159–67. [PubMed: 10165959]

Newhouse, Joseph P.; Price, Mary; Huang, Jie; Michael McWilliams, J.; Hsu, John. Steps To Reduce Favorable Risk Selection In Medicare Advantage Largely Succeeded, Boding Well For Health Insurance Exchanges. Health Affairs. 2012; 31(12):2618–28. [PubMed: 23213145]

Pauly, Mark. The Economics of Moral Hazard: Comment. American Economic Review. 1968; 58(3): 531–37.

Poterba, James. Government Intervention in The Markets for Education and Health Care: How and Why?. In: Fuchs, Victor, editor. Individual and Social Responsibility. University of Chicago Press; 1996.

Shepard, Mark. Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange. Mimeo: Harvard University; 2015.

Song, Yunjie; Skinner, Jonathan; Bynum, Julie; Sutherland, Jason; Wennberg, John; Fisher, Elliott. Regional Variations in Diagnostic Practices. New England Journal of Medicine. 2010 Jul.1:45–53. [PubMed: 20463332]

Van de Ven, Wynand PMM.; Ellis, Randall. Risk Adjustment in Competitive Health Plan Markets. In: Culyer, A.; Newhouse, J., editors. Handbook of Health Economics. Vol. 1. Elsevier; 2000.