Essays in Applied Economics

Daniel Sacks
University of Pennsylvania, sacks.daniel@gmail.com

Follow this and additional works at: http://repository.upenn.edu/edissertations
Part of the Economics Commons

Recommended Citation
http://repository.upenn.edu/edissertations/1430

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/1430
For more information, please contact libraryrepository@pobox.upenn.edu.
Essays in Applied Economics

Abstract
Essay 1 studies physician agency problems, which arise whenever physicians fail to maximize their patients’ preferences, given available information. These agency problems are well documented, but the magnitude of their welfare consequences for patients—the losses from suboptimal treatment choice induced by agency—are unclear. I infer patient drug preference from their compliance decisions. I begin by showing that initial prescriptions respond to physician financial incentives to control costs and to pharmaceutical detailing, but compliance does not, pointing to agency problems. I then develop and estimate a model of physician-patient interactions where physician write initial prescriptions, but patients choose whether to comply. Fully eliminating agency problems increases compliance by 6.5 percentage points, and raises patient welfare by 22\% of drug spending. Contracts that better align doctor and patient preferences can improve patient welfare, but attain only half the gains from eliminating agency completely. Although physician agency problems reduce patient welfare, eliminating them is thus likely difficult.

Essay 2, co-authored with Alexander M. Gelber and Damon Jones, studies frictions in adjusting earnings to changes in the Social Security Annual Earnings Test (AET) using a panel of Social Security Administration microdata on one percent of the U.S. population from 1961 to 2006. Individuals continue to "bunch" at the convex kink the AET creates even when they are no longer subject to the AET, consistent with the existence of earnings adjustment frictions in the U.S. We develop a novel estimation framework and estimate in a baseline case that the earnings elasticity with respect to the implicit net-of-tax share is 0.23, and the fixed cost of adjustment is $152.08.

Essay 3 studies the impact of health expenditure risk on annuitization. Theoretical research suggests that such risk can have an ambiguous influence on the annuitization decisions of the elderly. I provide empirical evidence on this linkage, by estimating the impact of supplemental Medicare insurance (Medigap) coverage on the annuity demand of older Americans. Medigap coverage has a strong impact on annuitization: the extensive margin elasticity is 0.39, the overall elasticity of private annuity income with respect to Medigap coverage is 0.56. These results are robust to controls for health, wealth, and preferences, as well as other robustness tests. They suggest that medical expenditure risk has a large impact on underannuitization.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Applied Economics

First Advisor
Uli Doraszelski

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1430
Second Advisor
Katja Seim

Subject Categories
Economics

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1430
ESSAYS IN APPLIED ECONOMICS

Daniel W. Sacks

A DISSERTATION

in

Applied Economics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

Uli Doraszelski
Professor of Business Economics and Public Policy

Co-Supervisor of Dissertation

Katja Seim
Associate Professor of Business Economics and Public Policy

Graduate Group Chairperson

Eric Bradlow, Professor of Marketing Statistics and Education

Dissertation Committee

Uli Doraszelski, Professor of Business Economics and Public Policy

Katja Seim, Associate Professor of Business Economics and Public Policy

Mark Duggan, Professor of Business Economics and Public Policy

Bob Town, Associate Professor of Health Care Management
ACKNOWLEDGEMENT

I am deeply grateful to my dissertation committee, Mark Duggan, Bob Town, and especially my advisors Uli Doraszelski and Katja Seim. Their advice and comments, and their high standards, have improved the dissertation enormously. Alex Gelber and J.F. Houde were not on my committee but might as well have been; their advice, too, has been hugely valuable. Many others have provided helpful comments on one or multiple papers in the dissertation: Mike Abito Raj Chetty, Jim Cole, Anthony Defusco, Hanming Fang Jonathan Fisher, Richard Freeman, John Friedman, Bill Gale, Matt Grennan, Josh Gottlieb, Ben Handel, Hilary Hoynes, Adam Isen, Henrik Kleven, Matt Notowidigdo, Mark Pauley Emmanuel Saez, Naomi Sacks, Kent Smetters, Cindy Soo, Amanda Starc, Ashley Swanson, and Jeremy Tobacman. I am grateful to Dr. Stephen Gottlieb and Dr. Sarah Miller for providing helpful detail about prescribing anti-cholesterol drugs. I acknowledge financial support from the Wharton Center for Human Resources and the Wharton Risk and Decision Processes Center, from NIH grant #1R03 AG043039-01, from a National Science Foundation Graduate Research Fellowship, and from support from the U.S. Social Security Administration through grant #5RRC08098400-05-00 to the National Bureau of Economic Research (NBER) as part of the SSA Retirement Research Consortium. The findings and conclusions expressed are solely those of the authors and do not represent the views of SSA, any agency of the Federal Government, or the NBER. The research uses data from the Census Bureau’s Longitudinal Employer Household Dynamics Program, which was partially supported by the following National Science Foundation Grants: SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation. All errors are my own.
ESSAYS IN APPLIED ECONOMICS

Daniel W. Sacks

Uli Doraszelski

Katja Seim

Essay 1 studies physician agency problems, which arise whenever physicians fail to maximize their patients’ preferences, given available information. These agency problems are well documented, but the magnitude of their welfare consequences for patients—the losses from suboptimal treatment choice induced by agency—are unclear. I infer patient drug preference from their compliance decisions. I begin by showing that initial prescriptions respond to physician financial incentives to control costs and to pharmaceutical detailing, but compliance does not, pointing to agency problems. I then develop and estimate a model of physician-patient interactions where physician write initial prescriptions, but patients choose whether to comply. Fully eliminating agency problems increases compliance by 6.5 percentage points, and raises patient welfare by 22% of drug spending. Contracts that better align doctor and patient preferences can improve patient welfare, but attain only half the gains from eliminating agency completely. Although physician agency problems reduce patient welfare, eliminating them is thus likely difficult.

Essay 2, co-authored with Alexander M. Gelber and Damon Jones, studies frictions in adjusting earnings to changes in the Social Security Annual Earnings Test (AET) using a panel of Social Security Administration microdata on one percent of the U.S. population from 1961 to 2006. Individuals continue to ”bunch” at the convex kink the AET creates even when they are no longer subject to the AET, consistent with the existence of earnings adjustment frictions in the U.S. We develop a novel estimation framework and estimate in a baseline case that the earnings elasticity with respect to the implicit net-of-tax share is
0.23, and the fixed cost of adjustment is $152.08.

Essay 3 studies the impact of health expenditure risk on annuitization. Theoretical research suggests that such risk can have an ambiguous influence on the annuitization decisions of the elderly. I provide empirical evidence on this linkage, by estimating the impact of supplemental Medicare insurance (Medigap) coverage on the annuity demand of older Americans. Medigap coverage has a strong impact on annuitization: the extensive margin elasticity is 0.39, the overall elasticity of private annuity income with respect to Medigap coverage is 0.56. These results are robust to controls for health, wealth, and preferences, as well as other robustness tests. They suggest that medical expenditure risk has a large impact on underannuitization.
ACKNOWLEDGEMENT ................................................................. iii

ABSTRACT ........................................................................ iv

LIST OF TABLES ................................................................. ix

LIST OF ILLUSTRATIONS ..................................................... xii

CHAPTER 1 : Physician Agency, Compliance, and Patient Welfare: Evidence from Anti-Cholesterol Drugs ..................................................... 1
  1.1 Introduction ................................................................. 1
  1.2 Background on anti-cholesterol drugs ................................. 7
  1.3 Data ......................................................................... 8
  1.4 Initial evidence on the impact of prices and advertising .......... 13
  1.5 A model of doctor-patient interactions ............................... 22
  1.6 Estimation results ....................................................... 34
  1.7 Can better contracts reduce agency problems? ....................... 39
  1.8 Conclusion and directions for future work ............................ 43

CHAPTER 2 : Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test ..................................................... 57
  2.1 Introduction ................................................................. 57
  2.2 Policy Environment ..................................................... 61
  2.3 Initial Bunching Framework ............................................. 65
  2.4 Data ......................................................................... 67
  2.5 Earnings Response to Policy ............................................. 69
  2.6 Mechanisms ................................................................ 75
2.7 Estimating Elasticities and Adjustment Costs ........................................... 79
2.8 Conclusion ..................................................................................................... 91

CHAPTER 3: Health Expenditure Risk and Annuitzation: Evidence from Medigap Coverage ................................................................. 102
3.1 Background on Medigap insurance ................................................................. 106
3.2 Data: Weiss Ratings and the Health and Retirement Study ......................... 108
3.3 Empirical strategy and first stage results ....................................................... 111
3.4 Results: Medigap coverage increases annuitization .................................... 118
3.5 Robustness ..................................................................................................... 121
3.6 Conclusion ..................................................................................................... 124

APPENDIX .......................................................................................................... 136

BIBLIOGRAPHY ................................................................................................. 223
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Molecule summary statistics, all years</td>
<td>46</td>
</tr>
<tr>
<td>Table 2</td>
<td>Multinomial logit estimates of the impact of patient price, plan price, and advertising on initial prescription</td>
<td>47</td>
</tr>
<tr>
<td>Table 3</td>
<td>Logit estimates of the impact of patient price, plan price, and advertising on compliance</td>
<td>48</td>
</tr>
<tr>
<td>Table 4</td>
<td>Differential responses by capitation status</td>
<td>49</td>
</tr>
<tr>
<td>Table 5</td>
<td>Robustness to informational content of advertising</td>
<td>50</td>
</tr>
<tr>
<td>Table 6</td>
<td>Parameters of doctor’s and patient’s utility function</td>
<td>51</td>
</tr>
<tr>
<td>Table 7</td>
<td>Match quality distribution</td>
<td>52</td>
</tr>
<tr>
<td>Table 8</td>
<td>Goodness-of-fit</td>
<td>53</td>
</tr>
<tr>
<td>Table 9</td>
<td>Impact of agency</td>
<td>54</td>
</tr>
<tr>
<td>Table 10</td>
<td>Summary Statistics, Social Security Administration Master Earnings File</td>
<td>100</td>
</tr>
<tr>
<td>Table 11</td>
<td>Estimates of Elasticity and Adjustment Cost Using Sharp Change Method</td>
<td>101</td>
</tr>
<tr>
<td>Table 12</td>
<td>Summary statistics</td>
<td>127</td>
</tr>
<tr>
<td>Table 13</td>
<td>First stage results</td>
<td>128</td>
</tr>
<tr>
<td>Table 14</td>
<td>Correlation between instrument and regressors</td>
<td>129</td>
</tr>
<tr>
<td>Table 15</td>
<td>Impact of Medigap coverage on annuity income</td>
<td>130</td>
</tr>
<tr>
<td>Table 16</td>
<td>Impact of Medigap coverage on extensive margin of annuity demand</td>
<td>131</td>
</tr>
<tr>
<td>Table 17</td>
<td>Impact of Medigap coverage on “true” annuity demand, extensive margin</td>
<td>132</td>
</tr>
<tr>
<td>Table 18</td>
<td>Placebo test for impact on retirees with employer-sponsored health insurance</td>
<td>133</td>
</tr>
<tr>
<td>TABLE</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>19</td>
<td>Include state-level Medicare expenditures</td>
<td>134</td>
</tr>
<tr>
<td>20</td>
<td>Include Medicaid policy variables</td>
<td>135</td>
</tr>
<tr>
<td>21</td>
<td>Imputed and mode prices for two plans, 2003</td>
<td>148</td>
</tr>
<tr>
<td>22</td>
<td>Predictive power of imputed prices</td>
<td>149</td>
</tr>
<tr>
<td>23</td>
<td>Sample sizes and composition</td>
<td>150</td>
</tr>
<tr>
<td>24</td>
<td>Patient summary statistics</td>
<td>151</td>
</tr>
<tr>
<td>25</td>
<td>Robustness of initial prescription estimates to alternative samples</td>
<td>152</td>
</tr>
<tr>
<td>26</td>
<td>Robustness of compliance estimates to alternative samples</td>
<td>153</td>
</tr>
<tr>
<td>27</td>
<td>Compliance regression: robustness to alternative definitions of compliance</td>
<td>154</td>
</tr>
<tr>
<td>28</td>
<td>Relationship between compliance and actual prices</td>
<td>155</td>
</tr>
<tr>
<td>29</td>
<td>Robustness to including outside option</td>
<td>155</td>
</tr>
<tr>
<td>30</td>
<td>LDL cholesterol reduction, by drug</td>
<td>156</td>
</tr>
<tr>
<td>31</td>
<td>Results from the Monte Carlo Experiment: Utility Function Estimates</td>
<td>156</td>
</tr>
<tr>
<td>32</td>
<td>Results from the Monte Carlo Experiment: Match Quality Distribution</td>
<td>157</td>
</tr>
<tr>
<td>33</td>
<td>Robustness to CP(2) in permanent income and preferences</td>
<td>160</td>
</tr>
<tr>
<td>34</td>
<td>Control for pre-age 65 endogenous variables</td>
<td>161</td>
</tr>
<tr>
<td>35</td>
<td>Control for spouse’s preferences</td>
<td>162</td>
</tr>
<tr>
<td>36</td>
<td>Exclude Medicaid recipients</td>
<td>163</td>
</tr>
<tr>
<td>37</td>
<td>Estimates of Elasticity and Adjustment Cost Using Sharp Change Method and Disappearance of Kink at Age 70</td>
<td>219</td>
</tr>
<tr>
<td>38</td>
<td>Estimates of Elasticity and Adjustment Cost Using Sharp Change Method and 1990 Policy Change, Assuming no Pre-Period Bunching Attenuation</td>
<td>220</td>
</tr>
<tr>
<td>40</td>
<td>Estimates of Elasticity and Adjustment Cost Using Comparative Static Method and 1990 Policy Change</td>
<td>222</td>
</tr>
</tbody>
</table>
# LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGURE 1</td>
<td>Out-of-pocket and plan prices, by plan, 2008</td>
<td>55</td>
</tr>
<tr>
<td>FIGURE 2</td>
<td>Out of sample fit for Lipitor, Zocor, and Mevacor</td>
<td>55</td>
</tr>
<tr>
<td>FIGURE 3</td>
<td>Impact of alternative contracts on compliance and patient welfare</td>
<td>56</td>
</tr>
<tr>
<td>FIGURE 4</td>
<td>Key Earnings Test Rules, 1961-2009</td>
<td>94</td>
</tr>
<tr>
<td>FIGURE 5</td>
<td>Histograms of Earnings, 59-73-year-olds Claiming OASI by Age 65, 1990-1999</td>
<td>95</td>
</tr>
<tr>
<td>FIGURE 7</td>
<td>Adjustment Across Years: Histograms of Earnings and Normalized Excess Mass, 66-69 year olds Claiming OASI by Age 65, 1996-2004</td>
<td>97</td>
</tr>
<tr>
<td>FIGURE 8</td>
<td>Normalized Excess Bunching by Year, 1961-2005</td>
<td>98</td>
</tr>
<tr>
<td>FIGURE 9</td>
<td>Bunching Responses to a Convex Kink, with Fixed Adjustment Costs</td>
<td>99</td>
</tr>
<tr>
<td>FIGURE 10</td>
<td>Relationship between Medigap and ln $\text{price}$</td>
<td>125</td>
</tr>
<tr>
<td>FIGURE 11</td>
<td>Relationship between endogenous variables and ln $\text{price}$</td>
<td>126</td>
</tr>
<tr>
<td>FIGURE 12</td>
<td>Distribution of prices in two plans, 2003</td>
<td>158</td>
</tr>
<tr>
<td>FIGURE 13</td>
<td>Empirical distribution of prices in 2001</td>
<td>159</td>
</tr>
<tr>
<td>FIGURE 14</td>
<td>Bunching Response to a Convex Kink (Frictionless Case)</td>
<td>197</td>
</tr>
<tr>
<td>FIGURE 15</td>
<td>Bunching Estimation Methodology</td>
<td>198</td>
</tr>
<tr>
<td>FIGURE 16</td>
<td>Mean Percentage Change in Earnings from Age 70 to 71, by Earnings at 70, 1990-1998</td>
<td>199</td>
</tr>
</tbody>
</table>
FIGURE 35: Probability of claiming OASI in year \( t+1 \) among 61-68 year-olds in year \( t \) who are not claiming, 1990-1998.
CHAPTER 1: Physician Agency, Compliance, and Patient Welfare: Evidence from Anti-Cholesterol Drugs

1.1. Introduction

Imperfect physician agency arises whenever physicians fail to maximize their patients’ preferences, given available information. Because doctors face myriad financial incentives, explicit and implicit, to provide some treatments but not others, there are many opportunities for imperfect agency, with broad consequences. Physician incentives affect the quantity and type of care provided (Gruber and Owings, 1996; Yip, 1998; Stern and Trajtenberg, 1998; Hellerstein, 1998; Coscelli, 2000; Iizuka, 2007; Liu et al., 2009; Limbrock, 2011; Epstein and Ketcham, 2012; Clemens and Gottlieb, 2013; Engelberg et al., 2013), the amount spent by private and public insurers (Iizuka, 2012; Dickstein, 2012), and even patient health (Johnson and Rehavi, 2013; Jacobson et al., 2013).\(^1\)

Despite the many studies documenting physician agency problems, there exists little work quantifying its welfare losses for patients or exploring whether alternative physician contracts could correct it. The welfare losses include not only worse health, but also more expensive or less desirable treatment. Quantifying the welfare losses from agency therefore requires learning patient preferences for different treatments. This is difficult since patients have few ways to reveal their preferences over medical treatment. Indeed, the existing literature on the demand for healthcare does not distinguish between doctor and patient preferences, and the resulting demand estimates recover some combination of the two.

To avoid this problem, I follow a suggestion by Ellickson et al. (2001), and focus on a unique aspect of the prescription drug market: although patients cannot purchase drugs without a prescription from a doctor, once they have a prescription they are free to comply with it or not. My premise in this paper is that compliance decisions reveal patient preferences. By separating initial prescriptions and compliance decisions, I learn doctors’ and patients’

\(^1\)For recent surveys of the large literature on physician agency, see McGuire (2000), McClellan (2011), and Chandra et al. (2012).
preferences from two separate demand curves. I first provide descriptive evidence on agency problems, then develop and estimate a structural model of doctor-patient interactions, and use the model to quantify the welfare consequences of eliminating agency entirely or developing contracts to better align doctor and patient preferences.

I study the market for anti-cholesterol drugs. Anti-cholesterol drugs, with $20 billion in sales, are a large segment of the prescription drug market, itself a $320 billion sector (IMS, 2011). These drugs are important for health, reducing mortality, stroke and heart disease (Baigent et al., 2005). Attaining these benefits requires actually taking these drugs, so compliance is a medically important outcome in its own right.

There are several potential sources of agency problems in this market. Pharmaceutical companies spent nearly $650 million in 2004 on promotional activity targeting physicians—called detailing—that is intended to persuade them to prescribe new, expensive drugs. While some of this advertising is no doubt informative, critics of detailing allege that it skews prescriptions away from the patient’s ideal. Moreover, insurance companies provide incentives, implicit or explicit, for doctors to control costs, in part by prescribing drugs that are cheap for the insurance plan to procure. A final source of agency problems is legitimate disagreement between doctor and patient about which drugs are medically appropriate; in such cases, the doctor’s views may trump the patient’s wishes.

Using a large insurance claims database which features rich information on health and healthcare utilization, I study how initial prescriptions and compliance respond to detailing expenditures, out-of-pocket prices, and plan prices, i.e. the cost to the insurance plan of procuring the drug. To identify patient and doctor price sensitivity, I exploit the considerable variation across insurance plans in out-of-pocket prices and plan prices charged for a given drug in a given year. To avoid the problem that patients may select plans with generous coverage for their drug, I limit my sample to patients new to taking these drugs.

---

²The advocacy group No Free Lunch believes that “drug companies, by means of samples, gifts, and food, exert significant influence on provider behavior,” and their “promotional materials and presentations are often biased and non-informative.” See nofreelunch.org.
Thus the price variation I exploit for identification is plausibly exogenous to patient drug preferences.

I show that the drug the doctor initially prescribes depends on out-of-pocket prices, on detailing expenditures, and on plan prices. Of these variables, however, only out-of-pocket prices affect the compliance decision. These results, which survive extensive robustness tests, provide clear evidence of agency problems. A perfect agent would only respond to variables that affect a patient’s utility, but because detailing and plan prices do not directly affect compliance, they likely do not directly affect patient welfare.

To quantify the consumer welfare consequences of these agency problems, I must address two challenges. First, the doctor’s initial prescription decision depends on her preferences and her patient’s, and these two factors need to be separated. Second, the initial prescription also likely depends on patient-drug match quality, observed by doctors but unobserved by the econometrician. The doctor only prescribes an expensive drug if her patient has a high match quality for it—for example, a patient with a more severe cholesterol disorder will likely be matched to a more powerful drug. This selection problem may bias estimates of patient preferences. My next step, therefore, is to implement a structural model that separately identifies doctor and patient preferences, allowing for this heterogeneous, unobserved match quality.

The model formalizes the notion that doctors write initial prescriptions but patients choose whether to comply, and these separate decisions can be used to separately identify doctor and patient preferences. In the model, the patient derives utility from drug consumption. Drug utility depends on drug and patient characteristics as well as unobserved patient-drug match quality. The patient cannot choose his preferred drug; instead, the doctor chooses an initial prescription for him. The doctor cares about the patient’s utility but her choice may also depend on detailing and financial incentives, as well as differential evaluation of drug quality, and idiosyncratic factors. Physician agency is imperfect to the extent that the doctor cares about more than just the patient’s utility. After filling the initial prescription,
the patient is free to quit treatment, to comply with the prescribed drug, or to return to the doctor to demand a different prescription.

Since doctors make discrete choices for the initial prescription, the model resembles the discrete choice demand models used in a growing literature that explores the industrial organization of prescription drug markets (Stern, 1996; Ellison et al., 1997; Cleanthous, 2002; Branstetter et al., 2011; Bokhari and Fournier, 2012; Arcidiacono et al., 2012; Dunn, 2012). This literature uses standard demand estimation techniques to recover preferences and uses the estimated demand systems to study counterfactual policies. With physician agency problems, however, market demand data do not necessarily reveal either patient or doctor preferences, and so these estimated parameters cannot be interpreted as preferences. As Dunn (2012) notes, to the extent that agency problems affect choice, “the model [he estimates] will only be an approximation to individual utility, and may be more appropriately viewed as a market demand function” (p. 173). This concern applies to estimates of patient utility in the prescription drug context, so the model I develop may be useful more broadly by providing a way to separate doctor and patient preferences.

I estimate the model via maximum likelihood. Estimation requires integrating over the high-dimensional match quality distribution, so I use sparse grid integration to make the problem computationally tractable (Heiss and Winschel, 2008). To solve the identification problem that the prescribed drug’s match quality is correlated with the drug’s characteristics, I rely on exclusion restrictions. The characteristics of other drugs serve as natural excluded variables, since they affect the probability of receiving a prescription for a given drug, but not the utility from filling it. To see how excluded variables help identification, consider increasing the price of one drug (say, Lipitor). This increases the share of people prescribed another drug (say, Zocor) but also changes their composition: in particular, people who switch to Zocor because of Lipitor’s price increase have a relatively low match quality for Zocor (or else they would have been prescribed it to begin with), and so average compliance for Zocor falls. The rate at which Zocor’s compliance falls as Lipitor’s price rises
identifies the density of Zocor’s match quality distribution. Once this density is known, it is straightforward to parametrically correct for the selection bias.

The point estimates, consistent with the descriptive results, indicate that doctors value not only patient utility but also plan prices and detailing. For example, doctors are roughly indifferent between a $1 increase in the monetized utility patients receive from a given drug, and a $2.41 decrease in plan prices for that drug. Doctors also evaluate the health benefits of drugs differently than do patients. The model therefore clearly points to imperfect physician agency. Accounting for match quality is important; ignoring it biases patient price sensitivity downward, by nearly half. The model fits the data well, both in sample and in a random hold out sample.

I use the estimated model to quantify the welfare losses from agency problems. Shutting down all agency problems, by requiring doctors to prescribe the utility-maximizing prescription, increases patient utility by about 22% of total anti-cholesterol drug spending. Compliance also rises, from 53.1% to 59.6%. Plan spending changes only slightly, and so the overall impact on efficiency of eliminating agency is large and positive. Physician agency therefore has important consequences for patient welfare, compliance, and overall efficiency. Eliminating individual aspects of agency, however, has markedly different effects. Removing doctor’s cost-containment incentives, for example, increases plan spending by $48 per patient per year but has only a small impact on patient welfare, so overall it reduces efficiency. This result implies that insurance companies are setting incentives fairly effectively: they are able to control costs without overly harming patient welfare.

Fully eliminating agency problems is likely difficult, because prescribing the patient’s utility maximizing drug requires observing match quality. Instead, I consider easily implementable alternative contracts to better align doctor and patient preferences. First I consider paying

---

3In these calculations, I assume that the patient’s revealed preferences are relevant for welfare calculations. As I discuss in more detail below, however, it could be that non-compliance reflects mistake by patients due to misperception of drug benefits, myopia, or other factors (Baicker et al., 2013). In that case, the impact of agency on compliance itself may be more relevant for welfare.
doctors a bonus for prescribing the drug with the highest average quality net of price. This contract encourages doctors to prescribe drugs that work well for the average patient. One danger of such a contract is that it may encourage doctors to prescribe drugs that they know to be a poor match, simply because they are attractive to the average patient. To avoid this problem, I also consider contracts that pay doctors a bonus based on the \textit{ex post} compliance of their patients. Since doctors use unobserved match quality to forecast compliance, these contracts take advantage of doctors’ private information.

Both contracts increase compliance and patient utility. Paying for compliance, by taking advantage of the doctor’s private information, produces bigger gains for patients than paying for average quality; with strong enough incentives, it can increase compliance to over 58\%, close to the agency-free level. The contracts attain only about half of the gain in consumer surplus of moving to the agency-free level, however. Realizing the full gains requires changing the prescriptions of inframarginal patients, who would comply with their current drug but would nonetheless prefer a different one. Such patients are hard to identify and so it is difficult to design contracts to help them. Overall, therefore, I find that agency problems have a substantial impact on patient welfare and compliance, but fully solving these problems is difficult with feasible contracts.

The remainder of the paper is structured as followed. Section 1.2 provides background on the market for anti-cholesterol drugs, and Section 1.3 describes the data sources and the creation of key variables, and contains summary statistics. Section 1.4 shows the basic evidence for physician agency. Section 1.5 introduces the model and estimation procedure, and section 1.6 presents the point estimates, quality of fit, and impact of eliminating agency. Section 1.7 explores the impact of alternative contracts, and Section 1.8 concludes and discusses possible directions for future work.
1.2. Background on anti-cholesterol drugs

Anti-cholesterol drugs help regulate the level of cholesterol in the body. The most popular class of anti-cholesterol drugs is statins, which are both effective in reducing cholesterol and widely tolerated. The statin class contains six different molecules during my sample period. In addition to statins, several other classes of drugs reduce cholesterol, although non-statins are not as effective and often have worse side effects than statins. Drugs from these classes are indicated when patients are statin-intolerant or have particular cholesterol disorders for which statins are ineffective. Some classes can also be prescribed in combination, and combination products may be more powerful in reducing cholesterol or targeting multiple cholesterol problems than statins alone.

The National Cholesterol Education Program offers clear guidelines on when to prescribe anti-cholesterol drug (Gundry et al., 2001): if a patient has high cholesterol given his risk factors and if lifestyle intervention fails. The risk factors include cigarette smoking, hypertension, low HDL cholesterol, family history of chronic heart disease, age (greater than 45 for men and 55 for women), and diabetes and heart disease. The more risk factors a patient has, the lower the cholesterol threshold at which to prescribe. Once drug therapy begins, the report recommend that it continue indefinitely; anti-cholesterol drugs do not permanently cure cholesterol disorders. Cholesterol drugs are therefore purely maintenance medications: once they are prescribed, patients must keep taking them in order to realize their benefits. This fact is critical for my interpretation of non-compliance: failing to comply with the medication—ceasing treatment—actually indicates dissatisfaction, and not a cure.

Despite these guidelines, the report offers doctors less guidance about which drug to prescribe. Talbert (2008) suggests statins as a first line therapy, but does not recommend a particular statin. Doctors apparently have a great deal of flexibility on this margin, especially among statins.

---

3This section is based on Gundry et al. (2001), the report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, and Talbert (2008)’s textbook treatment.
1.3. Data

I study a sample drawn from the Thompson-Reuters Marketscan databases. These databases contain inpatient, outpatient, and prescription drug insurance claims from about 100 large, self-insured companies. These claims include all drugs purchased and procedures performed, as well as the prices paid, by the patient and by the insurer. The data are limited, however: they contain very little demographic information and no information on income. Although I observe which insurance plan each patient belongs to, I have no information on the set of insurance plans available to him, making it impossible to model plan choice. A further limitation is that Marketscan defines a “plan” by the benefits offered, so that if a plan changes its benefits from year to year, the plan’s identifier changes; this rules out exploiting within-plan, over-time variation in benefits.

I supplement the Marketscan databases with drug-by-year information on total expenditures on direct to consumer advertising (DTCA) and detailing, from IMS health. These data are only available for 2001-2005 and 2008-2009 so I limit the Marketscan data to these years as well. In this section, I provide a description of the key aspects of the data. Appendix A.1 contains a more detailed description of the data, including definitions of all variables and products.

1.3.1. Drug price imputation

From the claims data, I observe two prices in each transaction: the out of pocket price paid by the patient, and the total amount paid to the pharmacy by the insurer and the patient, less the out of pocket payment. I call the first price the out-of-pocket price and the second price the plan price. The out-of-pocket price represents the actual cost to the patient of obtaining the drug. The plan price is related to but not exactly the cost to the plan of procuring the drug; the true cost accounts for the rebates that many insurance companies have negotiated with pharmaceutical companies. As in all empirical work on drugs, I do not observe these rebates and so I assume that the plan price is the true price. Nonetheless,
plan prices’ impact on prescribing behavior is likely mitigated by rebates, and this affects the interpretation of my results.\footnote{Arcidiacono et al. (2012), however, infer rebates from pricing and demand data as well as institutional details of the drug market. Their procedure relies on an equilibrium pricing model for pharmaceutical companies; incorporating the drug supply side with physician agency, while an interesting topic for future research, is substantially beyond the scope of this paper.}

I only observe prices paid, but demand estimation requires the full menu of prices faced, by the patient and by the plan. I therefore impute out-of-pocket and plan prices as the average price per days supplied at the plan-year-molecule level. Some small plan-years have zero sales for a given molecule; for these plan-year-molecules, I impute the price as the maximum of the other prices, and in all specifications I control for an indicator for this kind of imputation.

This imputation aggregates over three sources of price heterogeneity. First, some plans offer discounts for bulk purchases of 90 days supplied; the imputation reflects these discounts, yielding lower prices for plans with more generous discounts.\footnote{In principle this procedure generates a mechanical relationship between compliance and out-of-pocket price, since patients who purchase 90 days supplied are much more likely to comply and also face lower prices. I experimented with imputing prices for small cells with fewer than 500 claims, where this problem is likely to be most important, and found the results were extremely similar.} Second, identical molecules in a given plan can transact at different prices because of differences in dosage and in branding. I average over the differences in dosage. In practice the differences in branding are not important because when a molecule faces generic competition, branded sales are a trivial fraction of all sales (for example, two percent for simvastatin/Zocor; see Appendix A.1.1 for further discussion). Third, for plans which have a coinsurance or deductible for drug coverage, the out-of-pocket price will depend on local pharmacy prices and on spending throughout the year, and the average price will be a poor proxy for the price a given patient faces.

The procedure therefore works best in insurance plans with formularies, in which only a small number of out-of-pocket prices are charged. For example, a plan might charge $10 for generic drugs, $20 for preferred branded drugs, and $40 for all other drugs. Although the
Marketscan databases do not contain any indicators for formulary plans, it is easy to identify such plans from the distribution of prices paid: in formulary plans this distribution will have very limited support, while non-formulary plans will have a dispersed price distribution. Appendix Figure A.5 shows the empirical distribution of prices for the 24 largest plans in 2001. As the figure shows, some plans have an empirically obvious formulary: essentially all of the claims occur at three prices. Other plans have much more dispersed prices. I therefore define “formulary plans” as plans in which at least 80% of claims are at the top four modal prices (since many plans have a small mass of claims with zero price), and I limit the sample to formulary plans only. Appendix Table A.5 shows that the imputation procedure is much more accurate for formulary than non-formulary plans.

1.3.2. Sample creation

Starting with the full dataset, I limit the sample in several ways, intended to improve data quality and reduce endogeneity concerns. Appendix Table A.5 shows how these restrictions affect the size and composition of the sample. To avoid off-label use, my analysis sample consists of people with at least one chronic heart disease risk factor (as discussed in Section 1.2), who fill one or more prescriptions for a cholesterol drug, with their first fill at least six months after entering the data. These criteria follow Dunn (2012). I further exclude people with heart disease risk factors who fill no prescriptions for anti-cholesterol drugs because it is impossible for me to learn about their drug preferences, since I never observe them making a compliance decision, and this group is likely to be different in unobservable ways from people who fill at least one prescription (since the large majority of people with no fills are people who never received a prescription to begin with).

I limit the sample in three further ways. First, I require that people be continuously enrolled in the Marketscan data (but possibly with different plans) for at least 12 months after their first prescription, so if a patient’s first prescription occurs in February, I require that he be enrolled through the following January.7 By focusing on patients with new prescriptions

7I observe enrollment directly; I do not infer it from claims; so attrition from the sample is due to changing
and following them for a year only, I avoid the concern that people are selecting their drug insurance plan based on their medication mix. Second, I require that people belong to formulary-based insurance plans, as defined in Section 1.3.1. This restriction improves the accuracy of imputed prices. Third, I exclude people older than 65, because in later years these people have access to drug insurance through Medicare Part D, so the ones who retain employer provided insurance are likely to be very different from the rest of my sample.

The final dataset consists of 296,760 people in 383 plans. Appendix Table A.5 offers a snapshot of the sample at the time of the first prescription fill. The sample is relatively old and male, all reflecting the characteristics of people at risk for heart disease. About two-thirds of the sample have a claim with a diagnosis for a cholesterol disorder; the other risk factors—diabetes, heart disease, and hypertension—are all prevalent, but not overwhelmingly so.

On average people in my sample spend about $140 on cholesterol drugs per year, but $425 on other drugs, and almost $600 on inpatient and outpatient procedures. These figures provide further evidence that people are unlikely to choose their insurance plan based on the price of particular anti-cholesterol drugs (conditional on plan generosity), since cholesterol drug spending represents a small share of out-of-pocket spending.

People belong to large plans, with over 92,000 claims per plan; the median plan has 4,482 claims. For the most part, prices paid are close to the modal prices defined in Section 1.3.1. 98% of claims in these plans are at one of the top four modes, suggesting that prices are accurately imputed. About 25% of patients belong to capitated plans, comparable than the national average of about 22% during this time period (Kaiser Family Foundation and Health Research and Educational Trust, 2012). Statins are the clear majority of drugs prescribed, with a market share of about 80%.

Patients comply with their initial prescription—the first drug the doctor prescribes—about 52% of the time. Here compliance is defined as refilling at least 180 days supplied of the jobs, not non-compliance.
initial prescription in the first 330 days, with no fills of another anti-cholesterol drug. If the patient fills a prescription for another drug, but still has at least 180 days supplied, I say that he has switched; on average patients switch 12% of the time.8

1.3.3. Molecule-level summary statistics

Table 1.8 shows molecule-level summary statistics, pooling all years of data. The table reveals large differences across molecules in their initial prescription probabilities. The blockbuster Lipitor captures nearly 40% of the market, and Zocor takes almost a quarter; no other molecule has even 10% market share. Despite its popularity, Lipitor is far from the most expensive drug, to plans or patients, ranking behind Zocor, Pravachol, and the Bile Acid Resins (BARs). The non-statins, Fibrate, Niacin, and BARs, have fairly low market shares, individually and collectively. Zetia/Vytorin, a relatively new product meant to complement statins, also has a fairly low market share.

It is natural to think that the most heavily prescribed drugs are the best drugs, as viewed by patients and as reflected in compliance. If so, then Lipitor and Zocor should have the highest compliance rates. While compliance for these molecules is high, they rank behind Mevacor, the third most popular drug by initial prescription rate. In general there is much more compression in the compliance rates than in the initial prescription rates, and only Mevacor stands out as having especially high compliance. Patients are also unlikely to switch away from Mevacor, although Zocor and Lipitor’s switch rates are not far behind.

The next three columns of the table provide a hint about Lipitor and Zocor’s high initial prescription rates but unexceptional compliance rates. Lipitor was under patent protection for the entirety of my sample period, and Zocor for the first half (until 2006); the patent-holders Pfizer and Merck spent heavily on promotion for these products. But their patent protection yielded high prices, both to plans and especially to patients. Mevacor’s patent,

8I define compliance as 180 days supplied out of 330 because I follow patients for a year after they first fill a prescription for an anti-cholesterol drug. All patients, by construction, have at least 30 days supplied, so I look at the next 11 months and require for compliance that patients fill at least half the possible days supplied. The precise definition of compliance is not important for the results, as Appendix Table A.5 shows.
also held by Merck, expired in 2001, so Mevacor’s price, to patients and to plans, was much lower throughout the sample period.

These descriptive statistics therefore suggest that doctors face a trade-off between drugs patients like (including low cost drugs), and drugs that are heavily detailed. Working against this effect is the fact that heavily detailed drugs are expensive for insurance plans to procure, and plans may provide incentives for doctors to prescribe cheaper drugs.

In the rest of the paper, I move beyond these descriptive statistics in two ways. First, I use a regression framework to tease apart the impact of patient prices, plan prices, and advertising on initial prescriptions and refill rates. Because prices and advertising likely depend on drug quality, these regressions include molecule-specific means that account for most of the differences in quality across drugs. Second, I develop a structural model of doctors’ prescription decisions and patients’ compliance, switch, and quit decisions. The model accounts for the fact that doctors do not match patients to drugs randomly. In particular, although patients on Mevacor are more compliant than patients on Lipitor, it does not necessarily follow that switching patients from Lipitor to Mevacor would increase compliance; perhaps Lipitor is the best match available for patients who are prescribed it.

1.4. Initial evidence on the impact of prices and advertising

In this section, I present empirical evidence testing a specific hypothesis about imperfect agency: that doctors but not patients value detailing and plan prices. Under this hypothesis, these variables will affect initial prescription decisions but not compliance. To investigate these hypotheses, I estimate the following models of the initial prescription and compliance
decisions:

\[
Pr(d \text{ prescribed}) = \frac{\exp(\alpha_l p_{td}^{oop} + A_{td} \gamma_l + \theta_{itd}^I + X_i \beta_I + \mu_d^I)}{\sum_j \exp(\alpha_j p_{tdj}^{oop} + A_{tdj} \gamma_j + \theta_{itd}^j + X_i \beta_j + \mu_d^j)},
\]

\[
Pr(Comply_{td} | d \text{ prescribed}) = \frac{\exp(\alpha_C p_{td}^{oop} + A_{td} \gamma_C + \theta_{itd}^C + X_i \beta_C + \mu_d^C)}{1 + \exp(\alpha_C p_{td}^{oop} + A_{td} \gamma_C + \theta_{itd}^C + X_i \beta_C + \mu_d^C)},
\]

where \( p_{td}^{oop} \) is the out of pocket price to individual \( i \) in year \( t \) of molecule \( d \), \( A_{td} \) a vector of molecule-year level advertising characteristics, and \( p_{td}^{plan} \) the price to the plan. The additional controls, \( X_i \), are person-level characteristics including health status and plan generosity, and \( \mu_d \) is a drug-specific mean. Note that the person-level characteristics \( X_i \) drop out of the initial prescription regressions.

These equations do not account for the interdependence between compliance and initial prescriptions, but they offer a transparent look at the major patterns in the data, i.e. the relationship between compliance probabilities or initial prescriptions, and drug prices and advertising. Ultimately these relationships will be critical for identifying the full model. A strong and robust relationship among prices, compliance, advertising, and prescriptions indicates that features of the data, and not primarily the modelling choices, drive the model.

1.4.1. Sources of price and advertising variation

Equations (1.1) and (1.2) are essentially demand equations, and the usual concerns about identification of demand systems apply here. In particular, pricing and marketing variables may be correlated with product quality. In the insurance context, this is less of a problem, especially for out-of-pocket prices, which are not set to maximize profits from drug sales. Indeed, there is considerable variation in both out-of-pocket prices and plan prices for a given drug and year, as Figure 1.8 shows. The figure shows out-of-pocket and plan prices
for two generic drugs (Zocor and Mevacor) and two on-patent drugs (Lipitor and Crestor) for each plan in 2008. Drug prices differ widely, as do relative prices and even the relative rankings of prices in plans (indicated by prices being on both sides of the 45-degree line). The plan price variation represents idiosyncratic differences in pharmacy prices and in deals negotiated between pharmacies and insurance companies. This out-of-pocket price variation reflects differences in formulary placement—some plans place Lipitor on the middle tier and others on the top tier. Although patients may select plans with generous coverage for their preferred drugs, this is unlikely in my sample of patients who are new to taking anti-cholesterol drugs.

A second source of out-of-pocket price variation represents a threat to identification: overall plan generosity. Patients who are more likely to comply with their medication may select into more generous plans, which would bias my estimated price sensitivities. In my preferred specifications, I therefore control for plan generosity, defined as the average price of branded anti-cholesterol drugs in the plan. Alternatively I control for plan fixed effects, which absorb all aspects of plan generosity.\(^9\)

Although there is considerable price variation in the data, and this price variation may be uncorrelated with patient or doctor preferences, the advertising variation is much more limited, only at the molecule-year level. Much of the advertising variation is related to patent status: Zocor, Mevacor, and Pravachol all experience a decline in advertising as they go off patent. Crestor and Zetia’s advertising, however, likely reflects perceived product quality: initial clinical trials showed that these drugs had a large impact on cholesterol, but later research suggested that these cholesterol gains did not translate into reductions in mortality, heart attack, or stroke. Advertising expenditures on these drugs initially rises and then falls, following this pattern of clinical results. Overall, I cannot separate easily separate these sources of variation—driven by quality or by market structure—and so the advertising coefficients should be interpreted with caution.

\(^9\)Since plan generosity varies across people but not across drugs for a given person, it drops out of the initial prescription specifications.
A separate identification issue arises in the compliance regressions. Because doctors may match patients to drugs on the basis of (unobserved) match quality, patients who are prescribed a relatively high price drug are likely to have a relatively high match quality (or else the doctor would have prescribed a cheaper drug). Thus by conditioning on drug selection, I end up with a biased sample. The model developed in Section 1.5 below explicitly addresses this selection problem. Despite the bias, the reduced form estimates here are still informative, since the bias is towards zero. The point estimates reported here can be viewed as lower bounds (in absolute value).

1.4.2. Results for initial prescription choice

The results for the initial prescription regressions are in Table 1.8. The main specification, presented in column (1), shows that patient prices, plan prices, and detailing all affect the initial prescription decision, even conditional on molecule fixed effects. The standard errors, here and throughout the paper, are heteroskedasticity robust and clustered on insurance plan, the level of variation for most of the independent variables. The coefficient on patient price is about twice as large as on plan price, but because patient prices are typically much less than half plan prices, the elasticity of initial prescriptions with respect to plan prices, -1.189, is about fifty percent larger than the elasticity with respect to out-of-pocket price, -0.757. Nonetheless the results clearly indicate that both prices affect the initial prescription decision.

Both DTCA and detailing have a clear impact on initial prescriptions, although the impact of detailing is about twice as large as DTCA’s, comparable to the impact of plan price. The coefficient on detailing, 0.761, gives the impact on drug utility of a one standard-deviation increase, about $110 million. Pharmacies dispensed 52.7 million prescriptions for Lipitor in 2009, implying that increasing advertising by about $1 per prescription has a similar

---

10This clustering is somewhat conservative. Individual prices vary across drugs, within plan-year, and menus of drug prices vary across years, within plans, for the handful of plans that span multiple years. Nonetheless because the entire menu of prices determines choices in the discrete choice context, and the menu varies much less within plan over time than across plans, I cluster at the plan level.
impact on prescription probabilities to reducing price by $0.61 per prescription.

These results rely on the identification assumption that prices and advertising are uncorrelated with quality. Molecule fixed effects are intended to control for quality, so in column (2) I gauge their importance by removing them from the specification. The results are strikingly different. The coefficient on plan price falls to -0.057, a statistically insignificant tenth of its original level. The coefficients on DTCA is much larger, and the coefficient on detailing slightly smaller. These changes are consistent with the endogeneity of prices and advertising: higher quality drugs are more expensive to plans and more heavily advertised. Perhaps surprisingly, however, the coefficient on out of pocket price becomes more negative, not less, and about twice as large in absolute value. But out of pocket prices are set by insurance companies, not drug companies, so they need not be positively correlated with quality. The regression results suggest that out-of-pocket prices are negatively correlated with average drug quality, consistent with the possibility that insurance companies place “best-in-class” branded drugs such as Lipitor (which has a relatively low price) in a preferred position on their formulary, while placing less effective drugs on higher tiers.

The results in column (2) suggest that the molecule fixed effects control for important aspects of drug quality, but to the extent that drug quality is changing over time, they may be inadequate. In column (3) I therefore include molecule-by-year fixed effect, which control for all within-year aspects of drug quality. Because advertising is measured at the molecule-year level, it is collinear with these fixed effects, so this specification drops the advertising variables. While adding molecule fixed effects has a very large impact on out-of-pocket and plan price sensitivity, the impact of the molecule-year fixed effects is small, suggesting that the molecule fixed effects account for most of the endogeneity problems. In the context of prescription drugs, this is perhaps not surprising, since drug characteristics essentially do not change over time, although clinical knowledge evolves.

The final columns further test the robustness of the relationship between initial prescription choice and patient prices, plan prices, and advertising. The multinomial logit specification
implicitly controls for all factors that vary across people but not across drugs within person, so it is not necessary to control, for example, for type of insurance plan, health status, or plan generosity. Nonetheless if patients in different insurance plans have different tastes for drugs, and face different drug prices, then the coefficients on the price variables could be picking up this taste heterogeneity. In column (4) I include a full set of molecule-by-health status fixed effects (i.e. all the interactions between molecule fixed effects and sex, age risk, cholesterol, diabetes, hypertension, and heart disease) as well as a full set of interactions between molecule fixed effects and plan generosity, allowing people in more generous plans to have differential tastes for each drug. These additional fixed effects do not meaningfully change the point estimates.

1.4.3. Results for the compliance decision

The estimates in Table 1.8 are difficult to interpret because a positive coefficient on a variable could arise for two reasons: doctors have a direct taste for prescribing cheaper or more heavily advertised drugs, or patients prefer such drugs, and doctors take patient preferences into account in prescribing drugs. To help distinguish between these hypotheses, Table 1.8 shows the analogous results for patients’ compliance decision, estimated via logits. Since it is unlikely that patients comply with their prescriptions out of regard for their doctor’s utility, I interpret these specifications as reflecting patient preferences over drug characteristics.

The main specification in column (1) includes the price and advertising variables, as well as molecule fixed effects, plan generosity, and an indicator for imputed price. The results show clear and statistically significant out-of-pocket price sensitivity. The coefficient, -0.498, implies an elasticity of -0.141. Aside from out-of-pocket price, however, the coefficients look qualitatively unlike the initial prescription results: plan prices and detailing have essentially no impact on patients’ compliance decisions, and even DCTA has a small and insignificant effect. These results therefore suggest that plan prices and detailing affect doctor’s prescription directly through doctors’ taste for them, not indirectly through patients' preferences.
It is especially interesting that plan price is uncorrelated with the compliance decision. If plan prices were highly correlated with quality and did not otherwise affect compliance, then we would expect a positive coefficient on them. Instead, the zero coefficient suggests that these prices are not terribly endogenous.

The identifying assumption behind these regressions is that the price and advertising variation is uncorrelated with patients’ underlying compliance propensities. Two main threats to identification are that patients with higher utilization tendencies select more generous plans, with lower prices, and that better drugs have higher prices and more advertising. Both of these threats would bias the price coefficient towards positive values. I attempt to address them by controlling for molecule fixed effects and average plan generosity. In column (2) I remove the controls for plan generosity, and in column (3) I remove the molecule fixed effects. As expected, plan generosity is correlated with compliance propensity, so removing the plan generosity controls increases the coefficient on out of pocket price. Removing the molecule fixed effects, however, has a dramatic effect. It nearly doubles price sensitivity, and makes the advertising variables change sign. It also makes the plan price coefficient significantly negative. These results suggest that while advertising is highly correlated with quality (as patients perceive it), out of pocket prices are negatively correlated with it. This negative correlation could arise if insurance companies put the consensus best drug (e.g. Lipitor and Zocor) on the first or middle tier of the formulary, but put lower quality branded drugs on the top tiers.

In column (4) I add richer controls for plan generosity by controlling for plan fixed effects.\textsuperscript{11} Controlling for plan fixed effects increases out-of-pocket price sensitivity considerably (and also increases the coefficient on detailing, although it remains insignificant). These results suggest that out-of-pocket prices are correlated with plan compliance propensity, in ways not captured by my generosity measure. Somewhat surprisingly, the results here imply that people in plans with low prices have \textit{low} compliance tendencies. The results continue to

\textsuperscript{11} This necessitates dropping 60 people in 39 plans in which every person either complied or did not comply.
shiw the basic story, however, that compliance responds to out-of-pocket prices but not to plan prices or detailing. Adding molecule×year fixed effects in column (5) or a full set of interactions between molecule fixed effects and health and molecule fixed effects and plan generosity, in column (6) does not alter this conclusion.

1.4.4. Further robustness

The results in Tables 1.8 and 1.8 show robustness to alternative and increasingly detailed controls. It is possible, however, that my results are driven by my sample selection procedure, my price imputation procedure, or the definition of compliance (requiring at least 180 days supplied). In Appendix A.2, I present additional robustness tests to address these issues. The basic patterns are robust to looking at people in non-formulary plans; to looking at people in plans in which all available drugs have at least 10, 25, or 50 claims; to using actual rather than imputed prices in the compliance decision; and to defining compliance as requiring 90, 120, 150, 210, 240, 270, 300, or 330 days supplied. Finally, the initial prescription equation estimates are similar if I include an outside option—i.e., no prescription—in the choice set.

1.4.5. Differential effects by capitation status

The results suggest that doctors but not patients respond to plan price. I do not observe any details about doctors’ contracts with insurance companies, so it is difficult for me to explore the mechanisms that might generate this response. I do, however, observe plan capitation status. Doctors of patients in capitated plans are the residual claimants for medical expenditures, so they likely face true financial incentives to control costs. I therefore reestimate equations (1.1) and (1.2), but looking separately at doctors in capitated and uncapitated plans.

The results in Table 1.8 confirm that capitated doctors respond to plan prices, providing suggestive evidence that plan prices reflect doctor’s incentives to control costs. These results should be interpreted with caution, however, as capitated and uncapitated plans are clearly
different: out-of-pocket price sensitivity is much higher, and the detailing response different between these plans.

1.4.6. Agency or Information?

My preferred interpretation of the results so far is that they point to two physician agency problems. Doctors are more likely to prescribe heavily detailed drugs, but patients are no more likely to comply with prescriptions for these drugs. Doctors also shy away from prescribing drugs that are expensive to procure (even conditional on the price to the patient), but patients have no particular aversion to such drugs. These distortions imply that doctors do not prescribe the compliance maximizing or patient utility maximizing drug.

This view of the facts suggests that preferences change between the initial prescription and the compliance decision. An alternative interpretation, however, is that information has changed: at the time of the initial prescription, it is unclear how the patient will react to a certain drug, and plan prices and advertising contain information useful for forecasting that reaction. After the patient fills an initial prescription, though, there is no longer useful information in advertising or prices, and so they do not predict choice.

While I cannot rule out this interpretation entirely, two considerations make it unlikely. First, as Figure 1.8 shows, the plan prices vary at a very fine level—within drug and year, across plans. If they contained meaningful information about match quality, then this information would also have to vary across plans. This could happen if plans worked out lower prices for drugs that they thought their patients would be especially likely to use. But I focus on patients new to taking these drugs, and in some specifications I control for health×molecule fixed effects (so that the information would have to be orthogonal to observable health status). Thus it is unlikely that information drives the relationship between initial prescriptions and plan prices.

Detailing, on the other hand, undoubtedly does contain useful information. Indeed, a great deal of the content of detailing is publicity for positive clinical trials. In Table 1.8, I attempt
to control, in a blunt way, for new information: I control for the number of times each drug is mentioned in publications indexed by PubMed in each year, as well as the lag number of mentions.\footnote{\url{pubmed.org} provides machine-searchable bibliographic information for 22 million citations for biomedical literature. For the 6 statins, I counted all publications mentioning the underlying molecule. For Zetia/Vytorin I searched for mentions of “ezetimibe,” for Niacin-statins, I searched for “Niacin and simvastatin or Niacin and lovastatin.” For Fibrates and BARs, I searched for mention of any of the underlying molecules. For Niacin, I searched for “Niacin and cholesterol” and excluded the combination articles.} Controlling for each drug’s publication count and its lags does not meaningfully affect the coefficient on detailing, suggesting that the response to detailing corresponds to more than just information.

Finally, in a recent paper, Carrera et al. (2013) argue that doctors are imperfectly informed about patient preferences and the prices patients face. Although such information problems are likely present, it is unlikely that they could explain my results, since they would suggest that doctors respond too little to out-of-pocket prices, whereas I find that doctors respond too much to other variables. I conclude that these results provide initial evidence for agency problems in the market for anti-cholesterol drugs.

1.5. A model of doctor-patient interactions

In this section I develop a model of doctor-patient interactions in the market for anti-cholesterol drugs. The model will let me study how compliance and welfare change in counterfactuals as I shut down all or some agency problems, or implement alternative contracts to more closely align doctor and patient preferences.

The model formalizes the notion that compliance decisions reveal patient preferences, as initial prescriptions do for doctors. The model builds on the reduced form regressions in two ways. First, the model allows for doctors matching patients with drugs on the basis of unobserved match quality. This matching can induce a correlation between match quality and prices that might bias the estimated price sensitivity; the model corrects for this bias. Second, the model allows patients to switch drugs as well as comply or not. The patient switches if the doctor prescribes him a drug that he dislikes, when there are better options
available. Switching lets patients veto a very poor initial prescription, and provides some protection against agency. Accounting for switching is therefore important for measuring the welfare consequences of physician agency correct. The model builds on the work of Ellickson et al. (2001), who develop but do not estimate a model of doctor-patients conflicts of interest in which compliance decisions identify patient preferences. I extend their model to allow for patient switching, and take it to the data.

The model is meant to capture physician agency in a simple way that nonetheless lets me study welfare in counterfactuals. It abstracts from several rich features of the market, including uncertainty and learning about match quality, the process of patients matching to doctors, and decision making by pharmaceutical companies, insurers, and pharmacies (and therefore treats all prices and product offerings as endogenous). Exploring these aspects of the market is beyond the scope of this paper, but an important and interesting avenue for future work.

1.5.1. Model details

**Timing** I model doctor patient interactions as a two-period game. In the first period, the patient (he) visits the doctor (she) for the first time, she writes a prescription, and he fills it. In the second period, the patient decides whether to comply, visit the doctor to switch drugs, or quit. If he visits the doctor, she must write a prescription for a new drug, which the patient must fill. The periods are unequal length: the first period action reflects a single month’s purchase, but the second period reflects 11 months of purchase and non-purchase. The second period decision in the model can be viewed as a reduced form for a more complete dynamic model in which the patient decides every month whether to fill his prescription, visit the doctor, or not take the drug.

**Drug preferences** Doctor’s and patient’s action-specific utilities derive from the patient’s utility from drug consumption. Patient $i$ obtains utility from consuming drug $d$ as follows:

$$u_d(\xi_{id}) = -\alpha p_{id} + X_{id}^{P}\beta^P + \mu_d^P + \xi_{id}, \quad (1.3)$$
\( \alpha \) measures price sensitivity, \( p_{id} \) is the price to patient \( i \) of drug \( d \), and \( X_{id}^P \) is a vector of patient or drug characteristics affecting utility from consumption, including for example patient’s health status or direct to consumer advertising. \( \mu_d^P \) is average drug quality as patients perceive it and \( \xi_{id} \) is a patient-drug specific match term. \( \mu_d^P \) is meant to reflect the average benefit of taking an anti-cholesterol drug, so it includes both average side effects/tolerability and effectiveness. People differ, however, in both how well they tolerate drugs and in how they value the health benefits of anti-cholesterol drugs. The idiosyncratic aspect of match quality, \( \xi_{id} \), reflects these patient-drug interactions. I assume that \( \xi_{id} \) is normally distributed with mean zero and variance \( \sigma_d^2 \), independent across drugs. Independence imposes many restrictions on substitution patterns, but these substitution patterns are not the focus of my counterfactuals (as they would be if I were interested in changes in competition or market structure).

**Compliance** The patient’s utility from complying with his prescription consists of the drug utility plus an idiosyncratic shock:

\[
\begin{align*}
  u_c^P(\xi_i, d, \epsilon_{ic}) &= u_d(\xi_{id}) + \epsilon_{ic},
\end{align*}
\]

where \( \epsilon_c \) is a type I extreme value shock. \( \epsilon_c \) reflects the many factors affecting compliance not included in drug utility. For example, it includes the convenience cost of filling a prescription. While I assume that \( \epsilon_{ic} \) is part of people’s utility, an alternative interpretation is that \( \epsilon_{ic} \) reflects whether patients remember to fill their prescription. This interpretation is consistent with the view, articulated forcefully by Baicker et al. (2013), that most non-compliance is a mistake. So long as these mistakes are uncorrelated with drug characteristics, however, they do not affect my empirical analysis. In the welfare calculation, I calculate consumer surplus both treating \( \epsilon_{ic} \) as part of patient utility and not.

**Switching and quitting** If the patient visits the doctor to switch, the doctor writes a new prescription which the patient fills. The patient does not know what drug the doctor will prescribe and only has subjective expectations, \( \hat{P}_r(d' | \xi_i, d) \), that the doctor will prescribe
drug \(d'\) when the initial prescription is \(d\) and match quality \(\xi\). Visiting the doctor to switch yields a net benefit or cost \(\theta_s\), an idiosyncratic type I extreme value shock \(\varepsilon_s\), as well as the expected utility of the new prescription, with the expectation taken over the prescription:

\[
\begin{align*}
  u^p_s(\xi_i, d, \varepsilon_is) &= \theta_s + \left( \sum_{d'} \tilde{P}(d' \mid \xi_i, d) u_{d'}(\xi_id') \right) + \varepsilon_is = \bar{u}_s(\xi_d; d) + \varepsilon_is  \\
  \end{align*}
\]

(1.5)

Mean utility from quitting is normalized to zero, and quit utility consists only of a type I extreme value shock.

\[
\begin{align*}
  u^p_q(\xi_i, d, \varepsilon_{iq}) &= \varepsilon_{iq}.  \\
  \end{align*}
\]

(1.6)

As with the compliance error \(\varepsilon_{ic}, \varepsilon_is\) and \(\varepsilon_{iq}\) can be interpreted either as idiosyncratic convenience shocks or as mistakes, and I provide welfare calculations below for both interpretations.

**Initial prescription and second drug choice** The doctor’s utility from prescribing drug \(d\), as an initial prescription or a second drug choice, is

\[
\begin{align*}
  u^{MD}_d(\xi_d) &= w u_d(\xi_id) + X^{MD}_{id} \beta^{MD} + \alpha^{MD} f_{in_id} + \mu^{MD}_d + \varepsilon^{MD}_{dt}.  \\
  \end{align*}
\]

(1.7)

The doctor places a weight \(w\) on the patient’s utility from consuming the drug, \(u_d\), but she also values other drug characteristics, \(X^{MD}_{id}\) (e.g. detailing) and financial incentives \(f_{in_id}\), which may be explicit or implicit. Finally the doctor may disagree with her patient about average match quality, for example because doctors may disagree with patients about the relative importance of side effects and efficacy. \(\mu^{MD}_d\) is a drug-specific intercept in the doctor’s utility function to reflect such disagreements. \(\varepsilon^{MD}_{dt}\) is a type I extreme value shock, and the \(t\) subscript indicates that the first and second prescription are governed by different draws.

I do not directly observe \(f_{in_id}\). Instead I observe plan prices, the prices to plans of procuring drugs. I assume that plans set incentives to encourage doctors to prescribe cheap drugs, so
that $\text{fin}_{id}$ and $\text{price}_{plan}$ are closely correlated. In particular, I assume that

$$\text{fin}_{id} = (\beta_0 + \beta_1 \text{capitated}) \text{price}_{plan},$$

so that $\text{price}_{plan}$ enters the doctor’s utility function with a coefficient of $\alpha^{MD} \beta_0 + \alpha^{MD} \beta_1 \text{capitated}$. I cannot separately identify $\beta$ and $\alpha^{MD}$, so I normalize $\alpha^{MD} = 1$. This normalization matters in the counterfactuals because I explicitly vary doctor incentives. I return to it below.

**Information** All covariates are common knowledge. The choice-specific errors, the $\varepsilon$s, are unknown by either the doctor or the patient until each decision is made. Match quality $\xi$ is known by the doctor and the patient at the first prescription and is time invariant.

### 1.5.2. Choice probabilities

To solve this game by backwards induction, I first find the doctor’s probability of prescribing each drug if the patient requests a switch with an initial prescription for $d_1$. The logit errors imply that

$$Pr(d_2|d_1; \xi_i) = \frac{\exp \left( w_{ud_2}(\xi_{id_2}) + X_{d_2}^{MD} \beta \right)}{\sum_{d \neq d_1} \exp \left( w_{ud_1}(\xi_{id_1}) + X_{d}^{MD} \beta \right)}. \quad (1.8)$$

In equilibrium, subjective probabilities are correct, so $Pr(d''|\xi_i, d) = Pr(d_2|d_1; \xi_i)$. The patient choice probabilities are therefore

$$Pr(\text{comply}|d_1; \xi_i) = \frac{\exp(u_{d_1}(\xi_{id_1}))}{1 + \exp(u_{d_1}(\xi_{id_1})) + \exp \left( \theta_s + \sum_{d_2} Pr(d_2|\xi_i, d_1) u_{d_2}(\xi_{id_2}) \right)} \quad (1.9)$$

$$Pr(\text{switch}|d_1; \xi_i) = \frac{\exp(\theta_s + \sum_{d_2} Pr(d_2|\xi_i, d_1) u_{d_2}(\xi_{id_2}))}{1 + \exp(u_{d_1}(\xi_{id_1})) + \exp \left( \theta_s + \sum_{d_2} Pr(d_2|\xi_i, d_1) u_{d_2}(\xi_{id_2}) \right)} \quad (1.10)$$

$$Pr(\text{quit}|d_1; \xi_i) = \frac{1}{1 + \exp(u_{d_1}(\xi_{id_1})) + \exp \left( \theta_s + \sum_{d_2} Pr(d_2|\xi_i, d_1) u_{d_2}(\xi_{id_2}) \right)}. \quad (1.11)$$

Note that even in the absence of agency problems, patients may not comply with their prescription, if the best option (i.e. the maximal $u_d(\xi)$) nonetheless provides low utility.
And finally the doctor’s initial prescription choice probabilities are

\[
Pr(d_1|\xi_i) = \frac{\exp(wu_{d_1}(\xi_{id_1}) + X_{d_1}\beta_{MD})}{\sum_d \exp(wu_d(\xi_{id}) + X_d\beta_{MD})}.
\]

These choice probabilities depend on match quality \(\xi\), which the patient and doctor know but the econometrician does not. The likelihood is obtained by integrating over the distribution of \(\xi\).

### 1.5.3. Consumer surplus, social surplus, and health

**Consumer surplus** Expected consumer surplus, before all shocks are known, is

\[
\frac{1}{\alpha} \int \sum_d Pr(d_1 = d|\xi) \times \left( u_{d_1}(\xi) + \delta E_{\xi} \max\{u_{c}^{P}(\xi, d_1, \varepsilon_c), u_{s}^{P}(\xi, d_1, \varepsilon_s), u_{q}^{P}(\xi, d_1, \varepsilon_q)\} \right) dF(\xi).
\]

To understand this expression, work from the inside out. The first term in the large parentheses is the utility from filling the initial prescription. The second term is the option value in the second period associated with having that prescription. The expression in parentheses, therefore, is the patient’s overall utility from having a prescription for \(d_1\). The initial prescription is uncertain from the patient’s perspective, so to find utility we must integrate over the initial prescription choice. Finally, I calculate ex ante expected utility by also integrating over patient heterogeneity, \(\xi\). The \(\frac{1}{\alpha}\) term converts utility into dollars. In practice I set \(\delta = 11\) since the first period corresponds to 1 month but the second period to 11 months.

To operationalize this expression for consumer surplus, I need to calculate the option value of having a prescription for drug \(d_1\) at the beginning of the second period. I consider two approaches. First, I take the traditional view that the utility errors, \(\varepsilon_c, \varepsilon_s, \) and \(\varepsilon_q\), are part
of the patient’s utility, and so the usual log-sum formula applies:

\[
\max \{ u^p_c(\xi, d_1, \varepsilon_c), u^p_s(\xi, d_1, \varepsilon_s), u^p_q(\xi, d_1, \varepsilon_q) \} = \log \left( 1 + \exp(u_{d_1}(\xi)) + \exp(\bar{u}_s(\xi, d_1)) \right).
\]

One interpretation of \( \varepsilon_c, \varepsilon_s \) and \( \varepsilon_q \), however, is that they reflect mistakes in decision making, in which case they do not belong in the patients’ utility function. Under this view, the expected utility at the beginning of period 2 is

\[
\max \{ u^p_c(\xi, d_1, \varepsilon_c), u^p_s(\xi, d_1, \varepsilon_s), u^p_q(\xi, d_1, \varepsilon_q) \} = Pr(c|\xi, d_1)u_{d_1}(\xi) + Pr(s|\xi, d_1)\bar{u}_s(\xi, d_1).
\]

In the counterfactuals below, I report both measures of consumer surplus.

**Net surplus** Consumers are not the only actors in this market. A full accounting of welfare includes spending by insurance plans, insurance premiums, consumer gains from risk protection, and drug production costs (including advertising costs). These responses, though important, are beyond the scope of this paper. Instead I focus on consumer surplus and “net surplus” in the market, defined as consumer surplus less plan drug spending. This measure of surplus is interesting in its own right, because it represents the most money a profit-maximizing insurance plan could extract with a two-part tariff (i.e. an insurance contract with a premium and a copay; see Gaynor et al. (2000)). Thus the loss of surplus due to agency can be interpreted as the most money a profit-maximizing insurer could gain if it could unilaterally eliminate agency, holding fixed all prices.

**Health** These welfare measures do not directly reflect the health gains from taking anti-cholesterol drugs. They indirectly reflect health gains because higher quality drugs, measured by \( \mu_d + \xi_{id} \), are likely to have higher health benefits, and also confer higher utility. Nonetheless, health is an interesting and important outcome in its own right. To include health in the model, suppose that the health benefit of complying with a drug (or switching to it) is given by \( h_d \). This health benefit is constant across people—it does not depend on observed health status or match quality; Appendix A.3 summarizes evidence from clinical
studies roughly consistent with this view, and provides more detail on how I calculate $h_d$ from FDA labels.

A consumer’s expected health benefit is

$$E[H] = \int \sum_d Pr(d_1 = d|\xi) \left( Pr(c|d_1,\xi)h_{d_1} + Pr(s|d_1,\xi) \left( \sum_{d_2 \neq d_1} Pr(d_2|d_1,\xi)h_{d_2} \right) \right) dF(\xi).$$

This expression is the expected value of $h_d$, weighted by the probability that the patient complies with or switches to drug $d$. Implementing this health gain requires knowledge of $h_d$. In practice I use the reduction in LDL cholesterol estimated in clinical trials and measured in mmol/L. Cholesterol reduction is not a direct measure of drug-specific health gains (such as reductions in stroke or heart attack), but it is available separately for each drug in my data. Moreover, Baigent et al. (2005) show that in a meta-analysis of 14 clinical trials of statins, the reduction in stroke and heart attack caused by statin use is linear in the LDL cholesterol reduction of that statin: each mmol/L reduction in LDL cholesterol reduces the risk of heart attack or stroke by 20 percent. Cholesterol reduction is therefore a meaningful health outcome with precise, drug-specific measurements.

One limitation of using reductions in LDL cholesterol, however, is that non-statins primarily target other kinds of cholesterol; Niacin, for example, does very little for LDL cholesterol but increases HDL (good) cholesterol. These measures focus on a narrow aspect of health, and therefore complement the standard welfare metrics, which in principle reflect patient’s overall assessment of drug quality.

**Role of agency** Physician agency in the model arises because doctors may choose drugs to maximize their patients’ utility. The model has four sources of physician agency: Detailing and other drug characteristics, the $X_{d}\beta^{MD}$ terms in the doctor’s utility function; financial incentives as embodied by plan drug prices, $fin_{id}$; disagreement about average drug quality, $\mu_{d}^{MD}$; and idiosyncratic choice errors, $\varepsilon_{d}^{MD}$. These distortions affect patient welfare and health through two channels. First, they make it less likely that patients receive their
preferred drug. This mechanically reduces the patient’s expected utility. Second, by leading to lower drug utility on average, they also reduce compliance, which reduces patients’ health. But if doctors have a relative taste for potent drugs, then eliminating agency causes doctors to substitute towards weaker drugs. The overall effect of agency on health is ambiguous: if the compliance effects dominates, then health rises, but if the substitution effect dominates, then it falls. In the counterfactuals I quantify the sum of these effects but also show how the strength of the initial prescription changes, highlighting the substitution effect.

1.5.4. Discussion of modeling choices

This model contains several nonstandard aspects and simplifications which I now discuss.

**Error structure** The error structure is a logit-normal mixture. I use type I extreme value errors to facilitate computation, as they lead to smooth, closed form choice probabilities conditional on the normally distributed quality. Many other studies use normal mixtures, especially interacted with product characteristics, to obtain realistic substitution patterns (Berry et al., 1995). The normal errors serve a very different role in my context: they provide an explicit link between the initial prescription decision and the comply/switch/quit decision. This link is how I account for the non-random matching of patients to drugs.

**No first period decision for patients** Unfortunately, my data do not contain information on unfilled prescriptions, so I must assume that patients fill the first prescription. In principle, of course, patients could refuse to comply even at this point, and with sufficient data I could extend the model to allow for this margin as well. In that case the doctor must weight her utility by the probability that the patient fills his prescription.

**No outside option** The model has no outside option of “no prescription.” This is, potentially, a further source of agency problems: doctors might not prescribe a drug to patients who in fact want one. Following the literature, I abstract from this element of agency. Including an outside option is particularly difficult because I infer patient preferences from their compliance decision, but patients who are not prescribed have no option to comply,
and are likely to be different from patients with a prescription. In Appendix Table A.5 I show that the initial prescription regression results are robust to including an outside option in the doctor’s initial choice set, suggesting that ignoring the outside option does not overwhelming change the point estimates.

**No dynamics** The model abstracts from dynamics in two ways. Doctors and patients do not learn about the patient’s match quality from his experience with the drugs, and doctors are myopic in deciding which drug to prescribe; they do not consider the patient’s future compliance probability and switch probabilities. A forward looking doctor values compliance because the doctor only realizes $X_d^βMD$ if the patient complies. Since patients are more likely to comply with high-utility drugs, the myopic model conflates doctor altruism (measured by $w$) with forward-looking behavior. Thus $w$ should be viewed as a reduced form object, reflecting both forward-looking behavior and altruism per se. In future versions of this paper, I expect to allow for forward-looking doctors, however.

Crawford and Shum (2005) have shown that learning about match quality is an important part of prescription drug demand. Including these rich dynamics, however, would vastly increase the computational requirements of the model. My interest is in agency problems on the prescription choice margin, and in alternative contract structures that might fix them. For this question, the dynamics do not seem central, so I exclude them. Nonetheless, exploring the trade-off between physician knowledge and physician agency, and how patient learning might ameliorate it, is an interesting topic for future work.
1.5.5. *Maximum Likelihood Estimation via Sparse Grid Integration*

I estimate the model via maximum likelihood. Letting \( \theta \) be the vector of parameters, the contribution of individual \( i \) to the likelihood is

\[
L_i(\theta \mid X) = \int \Pr(d_{1i} \mid \xi_i) \times \\
\Pr(\text{comply} \mid \xi_i, d_{1i})^{\text{comply}} \Pr(\text{switch} \mid \xi_i, d_{1i})^{\text{switch}} \Pr(d_{2i} \mid \xi_i, d_{1i})^{\text{switch}} \times (1.13) \\
\Pr(\text{quit} \mid \xi_i, d_{1i})^{\text{quit}} \Pr(d_{1i} \mid \xi_i, d_{1i})^{\text{switch}} dF(\xi). 
\]

This integral is a multivariate normal integral and does not have a closed form solution. I approximate it using sparse grid integration (SGI; see Heiss and Winschel (2008)), a type of quadrature. Appendix A.4 provides further computational details and an explanation of SGI. Let \( \xi^k \) be a quadrature node and \( \omega_k \) its weight. The approximated likelihood is

\[
L_i(\theta) = \sum_k \omega_k \Pr(d_{1i} \mid \xi^k_i) \Pr(c \mid \xi^k_i, d_{1i})^{\text{comply}} \Pr(v \mid \xi^k_i, d_{1i})^{\text{switch}} \Pr(d_{2i} \mid \xi^k_i, d_{1i})^{\text{switch}} \Pr(\text{quit} \mid \xi^k_i, d_{1i})^{\text{quit}}. 
\]

I compute all standard errors using the heteroskedasticity- and cluster-robust procedure suggested by Wooldridge (2002); the resulting standard errors are robust to arbitrary within-plan autocorrelation.

1.5.6. *Identification*

Although the model is straightforwardly parametrically identified, I discuss informally the features of the data that help identify the key parameters of the model.

**Identification of \( \Sigma \)** The variance of the match quality distribution is the hardest set of parameters to identify. Its identification relies on a kind of exclusion restriction: variables \( Z_d \) that affect the probability that drug \( d \) is chosen but are known not to affect the patient’s utility from consuming \( d \). In this context, the prices and characteristics of the other drugs satisfy this restriction.\(^{13}\)

---

\(^{13}\)This identification argument is similar to the identification of the Roy Model with covariates (Heckman and Honoré, 1990).
To see how these exclusions identify match quality, consider what happens to the compliance rate of drug $d$ when the price of $d'$ increases. Increasing a drug’s own price affects compliance in two ways: first, it has a direct effect, making the drug less appealing. But second, because the drug is now more expensive, doctors will prescribe it only to patients with higher match quality, who are more likely to comply. But when the price of $d'$ changes, only the second effect is present.\footnote{There is an additional effect: patients are differentially likely to switch. But this effect does not influence the quit/comply margin.} The rate at which compliance changes as $Z_d$ changes identifies the density of the match quality distribution. In my application I normalize the variance of Lipitor’s match quality to unity, since I found it difficult to identify the scale of the match quality distribution (in simulations and in the actual data).

**Identification of $\mu^p$, $\theta^v$, and $\mu^{MD}$** Conditional on the match quality variance, patient mean drug utility $\mu^p$ is identified by average compliance rates and average rate of switching away for a given drug. $\theta^v$ is identified by the overall average switch rate. The doctor’s quality, $\mu^{MD}$, is identified by matching the initial prescription and second prescription rate (conditional on switching from the initial prescription). Note that $\mu^{MD}$ and $\mu^p$ are in a sense overidentified, because the same fixed effect is used to match two moments. Because the location of doctor’s utility is not identified, I normalize Lipitor’s mean quality to doctors, $\mu_1^{MD}$, to zero. (On the patient side, Lipitor’s quality is identified by the normalization that mean utility from quitting is equal to zero.)

**Identification of preference parameters and $w$** Patient preferences are identified by the covariance between drug characteristics and compliance probabilities, after correcting for selection, i.e. the fact that among prescribed drugs, drug characteristics are correlated with match quality.\footnote{For a given match quality distribution, however, the correlation takes a known form and so can be corrected. See Ellickson et al. (2001), proposition 2.} Thus identification of price sensitivities and responses to advertising, once the match quality distribution is known, relies on the same variation as the descriptives results. Once patient preferences are known, so is $u_d$, and so the covariance among initial prescription choice probabilities, $u_d$, and $X^{MD}$ identifies $w$ and $\beta^{MD}$. However, $u_d$ is a linear
function of drug characteristics, as is $X^{MD}$. Identification of $w$ therefore requires at least one variable in patient utility that is not in doctor utility. Out-of-pocket price, its interaction with capitation, and direct-to-consumer advertising serves as the primary excluded variables (although I also include dummies for price imputation in the patient’s utility function but not the doctor’s). The coefficients on these excluded variables in patient’s utility are identified by both the compliance and the initial prescription response to them.

1.6. Estimation results

To minimize the computational burden, I estimate the model on a 5% random sample of the data, stratifying on insurance plans (so as not to lose price variation), resulting in a sample of 15,053 people.\footnote{The stratification procedure always selects at least one person from each plan, and hence oversamples people in plans with fewer than 100 people.} I include only a limited number of controls in the model. In the drug utility, I include out of pocket price and plan price, both prices interacted with capitation status, an indicator for imputed price, a main effect for capitation, plan generosity, and the two advertising measures. I let detailing, plan price, and plan price interacted with capitation shift the doctor’s utility differentially. The 45 parameters to estimate consist of 12 coefficients on these variables, the weight the doctor places on patient utility, the mean utility from visiting the doctor, and three parameters for each of the 11 drugs (standard deviation of match quality, and the mean utility to doctors and to patients), less the two normalizations.

1.6.1. Point estimates

Table 1.8 shows the estimated parameters of the utility function, and Table 1.8 shows the parameters of the match quality distribution. The estimated utility functions appear quite reasonable. Patients are price sensitive, and more so in capitated plans. The average elasticity of the initial prescription with respect to plan price is 1.00 (0.863 for uncapitated plans and 1.41 for capitated), and with respect to out-of-pocket price it is 0.79 (0.61 and 1.35). The elasticity of compliance with respect to out-of-pocket price, conditional on the
drug prescribed, is 0.22 (0.16 and 0.38).\textsuperscript{17} This elasticity is about 50% larger than the reduced form estimates from the compliance decision.

To better understand the coefficients, including the weight doctors put on patient utility, consider decreasing the patient’s utility from drug consumption by $1, either by increasing its price or by adjusting its characteristics so that utility falls by the price sensitivity $\alpha$. In order to keep an uncapitated doctor’s utility from prescribing that drug unchanged, its plan price would have to fall by $2.41$ (holding patient utility from plan prices fixed). Alternatively, we could compensate the doctor by holding plan prices fixed but increasing detailing, which would have to rise by 1.02 standard deviations to keep the doctor indifferent.

The second set of estimates in Table 1.8 shows the estimates when I set the standard deviation of match quality to zero. Ignoring match quality biases the point estimates towards zero (on the characteristics that vary across drugs), reducing patient price sensitivity by nearly half, for both capitated and uncapitated patients. Accounting for heterogeneous match quality is clearly important.

Table 1.8 shows the estimated match quality distribution, i.e. mean quality to doctors, mean quality to patients, and the standard deviation of match quality. There are clear differences between doctors and patients in average perceived match quality: $\mu_{d}^{MD}$ is often different from zero, although it is imprecisely estimated in many cases. Doctors appear to over emphasize non-statins and under emphasize combination products. Doctors also place too little weight on Mevacor. The estimates also imply clear heterogeneity in match quality: most of the estimated standard deviations of match quality are significant.

1.6.2. Goodness of fit

Table 1.8 provides some information about goodness of fit. The model fits for the initial prescription and comply decisions are both quite good. The initial prescriptions are all

\textsuperscript{17}The probability of compliance is $\int_{\xi} \sum_{d} Pr(d|\xi)Pr(c|d,\xi)$. To find the compliance elasticity, I differentiated this expression with respect to out-of-pocket price, but holding $Pr(d)$ fixed; this is to keep the elasticity comparable to the reduced form estimate, which holds fixed the initial prescription.
within a percentage point, and the compliance decision is also close for all drugs except BARs (and interestingly, the model estimates here provide better out-of-sample fit than the raw data). The model has some trouble matching the high compliance rate of Mevacor. The model does worse on predicting switching away from drugs, especially for the drugs with low initial prescription rates; since these drugs are rarely prescribed, errors in the rate of switching away from them does not much affect the likelihood.

The final columns, however, show that the model has a fairly hard time matching the rate of switching to drugs, conditional on switching. The model struggles here because the drug fixed effects must match both the switch to rates and the initial prescription rates. The error is much larger for the switch to rates because switching is relatively rare, so errors in the conditional probability do decrease the likelihood enormously. Because the same fixed effects govern the initial prescription and the switch to rates, the model will have a hard time matching differences between these probabilities. To the extent that switching reflects either a need for a stronger drug or a general statin intolerance, the model has trouble picking it up.

To further judge the fit, Figure 1.8 shows the initial prescription, compliance, and switch from rates, by year, for Lipitor, Zocor, and Mevacor, using the holdout sample (which comprises 95% of the data). The solid line shows the actual rates in the holdout data, and the dashed line shows the prediction from the model parameters in the holdout sample data.

The model fits the data reasonably well, although not perfectly. It matches the decline in Lipitor’s popularity, as well as the rise in Zocor. It also roughly tracks the rise in Mevacor’s share from 2001 to 2004, although it does not match the magnitude. The fit for the compliance and switch is also good. Overall, the fit of the model is reasonable, given that I am looking out of sample, and nothing in the model requires that I match the yearly patterns—including the large drop for Lipitor’s initial prescription share and rise for Zocor’s.
1.6.3. Welfare losses from agency

Given the adequate fit of the model, I use it to quantify the importance of physician agency for compliance and welfare. The model allows for four sources of agency problems: financial incentives, detailing, disagreement about mean drug quality (as represented by the doctor fixed effects), and the idiosyncratic logit errors. In Table 1.8, I shut down each of these factors, jointly and individually, to see how they affect compliance rates, patient utility, plan spending, and net surplus, i.e. utility less plan drug spending.\(^{18}\)

The first column of Table 1.8 shows baseline compliance, welfare, and spending. Compliance is 53% and 11% of patients switch. Consumer surplus is just under $1100 per year, while plans spend $563 (over and above out of pocket spending of about $125), resulting in net surplus of $530. Interpreting the choice errors as mistakes, in the row labelled Consumer surplus (mistakes), implies lower consumer surplus, but does not substantially affect the counterfactual welfare changes. The final two rows show the direct health consequences of access to prescription drugs. On average, doctors prescribe drugs which reduce LDL cholesterol by 1.07 mmol/L. Because patients do not always comply with their prescriptions, the realized cholesterol reduction is only 0.71 mmol/L. Baigent et al. suggest that each mmol/L reduction in cholesterol reduces the risk of stroke and heart attack by 20%, so this implies a 14% baseline reduction in risk.

In column (2) I shut down all agency problems by requiring doctors to prescribe the drug that maximizes patient utility. Compliance rises by 6.5 percentage points and, although patients are now better matched to their first choice, the gains from switching are also higher, so switching rises by 1 percentage point. Eliminating agency has two offsetting effects on plan spending: increasing compliance mechanically increases spending, but doctors have a taste for expensive drugs, and so shutting down agency means cheaper drugs are prescribed.

\(^{18}\)Shutting down the logit errors implies that the initial choice probabilities are discontinuous functions of match quality, \(\xi\), and so SGI cannot be used to evaluate the choice probabilities. In all the calculations underlying Table 1.8, therefore, I use crude Monte Carlo integration with 1000 draws. All counterfactual outcomes are linear functions of initial drug choice probabilities, however, so the integration, although noisy, is not biased.
The net effect is a modest increase in plan spending of $13. Eliminating agency results in much higher match quality, improving patient welfare: consumer surplus increases by $148, about 20% of baseline total spending. Because plan spending changes little, net surplus also increases by $136. Because doctors have a slight preference, relative to patients, for stronger drugs, the strength of prescribed drugs falls slightly, but because patients are more likely to stick with their medication, overall cholesterol improvement increases.

Eliminating all agency problems raises compliance, patient welfare, net surplus, and health. But how do the individual aspects of agency contribute to these gains? In columns (2) and (3) I shut down financial incentives and advertising, separately, by setting their coefficient in doctors’ utility to zero. Eliminating these agency problems actually reduces patient utility, although only slightly; compliance also falls slightly. This may seem surprising, but the intuition is straightforward. Although advertising and plan price have only a very slight impact on patient utility, they are correlated with characteristics that patients do value; patients like drugs that happen to be heavily detailed, and they dislike expensive drugs. Eliminating this aspect of agency makes the other aspects—the drug fixed effects and the logit errors—relatively more important in the doctor’s utility function, and so she switches towards drugs that patients like even less.

Although shutting down financial incentives has a small impact on patient utility and on health, it has a big effect on plan spending: eliminating them increases spending by $48. Eliminating detailing has the opposite impact on spending, decreasing it by $42. These results suggest that patients are close to indifferent between drugs that are expensive and cheap for plans to procure, and so the financial incentives are effective in controlling costs without harming patients. The results also suggest that, while doctors are unwilling to harm patients, they are relatively more open to transferring profits between insurers and pharmaceutical companies.

In column (5) I set the molecule fixed effects in the doctor’s utility function to zero. These fixed effects are an important source of disagreement between doctors and patients, so
eliminating them raises patient welfare. And because doctors, on average, have high taste for expensive drugs, shutting down the fixed effects also reduces plan spending considerably, and the net effect is an increase in net surplus of about $80. Finally, column (6) shuts down the logit errors, which effectively puts more weight in the doctor’s decision on patient utility and the observable aspects of drug quality. Compliance and utility increase considerably, but plan spending increases as doctors shift towards higher quality, more heavily advertised drugs. The net effect, again, is a sizeable increase in consumer surplus and net surplus.

1.7. Can better contracts reduce agency problems?

The results in the previous section indicate that eliminating physician agency can increase compliance and patient welfare, reduce prescription drug spending, and raise net surplus between patients and insurers. It is not possible, of course, to simply switch off agency. But agency problems ultimately reflect a contract failure: they arise because the doctor’s incentives are imperfectly aligned with her patient’s interest. A natural solution, therefore, is an alternative contract that better aligns incentives. In this section, I conduct counterfactual simulations to study the impact of alternative contracts on compliance, patient welfare, and net surplus.

In studying the welfare implications of these contracts, I ignore the possibility that additional spending on drugs will be offset by reduced hospital spending in the future. Chandra et al. (2008, 2010) study these offsets in a particularly clean setting, looking at hospital spending among Medicare beneficiaries when drug copays rose. Although they find evidence for offsets, in their heterogeneity analysis they find no evidence for offsets among people with high cholesterol, and they also find much smaller offsets among the relatively young (65-74 year olds). As these groups are likely closest to my sample, I assume that the spending offsets are small enough to ignore in my welfare calculations.

19Making all drugs available over-the-counter might be a solution, but it ignores the possibility that doctors are better informed than patients about the risks and benefits of drugs at the time of the initial prescription. The counterfactuals hold fixed information.
1.7.1. The contracts

One challenge in designing these contracts is that the optimal drug choice depends on the patient’s characteristics (especially his insurance plan), but also on the match quality between patients and drugs. If match quality is observed, it is trivial to write a “forcing contract” that imposes an arbitrarily large penalty on doctors for prescribing all but the best drug. The more interesting, and more realistic, case is when match quality is not observed. I consider three contracts in this context.

**Contract 1: “Quality.”** I first consider a contract that pays the doctor a bonus if she prescribes the drug that is best, based on the average quality drug quality, net of cost. Average quality is defined as the mean utility of drug consumption, gross of match quality, but conditional on patient and drug characteristics. This contract is conceptually straightforward to design and implement. It encourage doctors to prescribe drugs that are cheap for patients and that have high average quality, and can therefore improve utility and compliance. The downside of this contract is that it ignores heterogeneity in match quality and, in an extreme case, can encourage the doctor to prescribe a drug that she knows to be a poor fit, simply because it ranks first on observables.

**Contract 2: “Comply.”** The problem with the “Quality” contract is that it does not take advantage of the physician’s private information about match quality. Contract 2 attempts to avoid this problem by paying doctors a bonus if, ex post, patients end up complying with the prescription. Because compliance decisions reflect match quality, these contracts encourage doctors to prescribe drugs with high match quality.

**Contract 3: “Threshold.”** The “comply” contract leverages the doctor’s private information, but it also rewards inframarginal behavior: more than half of patients comply without any special incentive, so the “comply” contracts could be expensive to implement. Contract 3 instead pays doctors for achieving a threshold compliance rate, in practice 40%.

---

20In actually implementing this contract, I pay the bonus if the expected compliance probability exceeds the threshold. If doctors have many patients and make each prescription decision in isolation (rather than
To implement these contracts, I assume they shift doctor utility in the same way that plan prices do.\textsuperscript{21} The doctor’s utility from initially prescribing drug $d$ is therefore

$$u_{d}^{MD}(\xi_{d}) = \alpha^{MD}(\beta_{0} + \beta_{1} \text{capitated}) \text{price}_{plan}(\text{bonus}_{id} + \text{price}_{plan}^{id})$$

$$+ wu_{d}(\xi_{id}) + X_{id}^{MD} \beta^{MD} + \mu_{d}^{MD} + \varepsilon_{dt}^{MD}.$$ 

The doctor’s utility if the patient switches to drug $d$ is unchanged. The bonus amounts are

$$\text{bonus}_{d}^{best} = b \times \left\{ \bar{u}_{d} = \max_{d'} \bar{u}_{d'} \right\}$$

$$\text{bonus}_{d}^{comply} = b \times \Pr(c|d, \xi_{d})$$

$$\text{bonus}_{d}^{threshold} = b \times 1 \left\{ \Pr(c|d, \xi_{d}) \geq \text{threshold} \right\},$$

where $\bar{u}_{id'}$ is drug utility exclusive of match quality, a function of observable drug and patient characteristics only. The compliance probabilities include patient match quality but they are in principle observable ex post so they can be contracted on.

In the counterfactuals, I vary the size and form of the bonus payment $b$. Note that the impact of bonus payments on doctors’ utility is moderated both by their marginal utility of income, $\alpha^{MD}$, and by $\beta$, the relationship between drug prices and doctor income. These two parameters are not separately identified. In practice I normalize $\beta_{0} = 1$, but the results should be understood as relative to the doctor’s marginal utility of income. Because this is presumably constant across the bonus payments, the counterfactuals show how changes in bonus translate into changes in prescribing patterns, but it cannot speak to how much spending is required to achieve a given level of compliance or patient utility. With this limitation in mind, I do not focus on particular bonus levels; instead I vary bonus amounts over a wide range and examine the performance of each contract at each bonus amount. These results are useful for ranking for the contracts and for seeing what can be achieved at some spending level.

\textsuperscript{21} But with an opposite sign, since plan prices are a loss and the bonuses a gain.
1.7.2. Better contracts increase compliance and consumer surplus

Figure 1.8 shows the results. In Panel A I plot the compliance rate as I increase the yearly bonus from $0 to $600. All three contracts initially increase compliance. The “quality” contract eventually decreases compliance as doctors start to prescribe drugs that they know to be poor matches. The “threshold” contract clearly does best, and with strong enough incentives it achieves the compliance rate that would obtain absent any agency problems. These contracts, unsurprisingly, also markedly increase patient welfare, as panel B shows: consumer surplus rises with all three contracts. With enough incentives, however, the “quality” contract indeed reduces welfare slightly. Under the other contracts, consumer surplus is never decreasing in the bonus amount, but it levels off at about $1150. Although the “threshold” achieves the first-best compliance level, it produces, at best, only about a third of the welfare gain from eliminating agency. This is because these contracts do not improve agency problems for inframarginal patients, patients who would comply with their current prescription but would nonetheless prefer a different one.

The gains in compliance and utility affect plan drug spending in two ways. Increasing compliance for a given drug mechanically increases spending. Offsetting this effect, however, is that patients (relative to doctors) have a preference for cheaper drugs. Mitigating agency encourages doctors to switch patients to cheaper drugs, and this can drive down spending. This effect is especially pronounced for the “quality” contract, since plan price does not enter into patient utility. The “comply” contract also reduces drug spending slightly at all levels, while the “threshold” contract drives it up noticeably (in part because it achieves the highest level of compliance). The cost of the contracts also includes the direct outlays to physicians, however, and these are shown in Panel D. More generous bonuses translate almost linearly into higher compensation.

Whether insurance companies find these new contracts attractive depends on how they

---

22The units here are dollars per patient per month. A $300 bonus is quite large, but recall that the scale of the bonus is not identified and is normalized to 1.
split the surplus with the doctors and patients. Panel E shows consumer surplus less total plan spending (i.e. drug spending plus the bonus payments), and panel F shows total surplus (consumer surplus less drug spending). In all cases, utility less total plan spending (i.e. treating bonuses as a cost) is decreasing with the bonus amount. This result implies that insurance companies will lose money under the new contracts unless they can renegotiate contracts with doctors, exchanging higher bonuses for lower base salaries. Such renegotiation may not be difficult for the “quality” contract as the net losses are quite small for it.

If plans can renegotiate with doctors, then the bonus payments are just a transfer, and total surplus—consumer surplus less drug spending—is the relevant welfare metric. Panel F shows that total surplus is increasing under all contracts, and most strongly for the “quality” contract. “Comply” and “threshold” produce much weaker gains that do not increase rapidly with the bonus amount, despite the increase in patient utility they generate.

The problem is that the “comply” and “threshold” contract, despite exploiting doctor’s private information about match quality, target the wrong patients. The contracts encourage doctors to improve compliance for patients just on the margin of complying and not. But for these patients, improving compliance reduces efficiency. Since their drug purchases are heavily subsidized, their willingness to pay is far below the cost to the plan of procuring the drug. Encouraging compliance among these patients reduces welfare. Alternative contracts can therefore ameliorate but not fully solve the agency problems in the market for anti-cholesterol drugs, but optimal contracts—which likely lie outside of the set of contracts considered here—must encourage compliance only among high-value patients.

1.8. Conclusion and directions for future work

This paper argues that physician agency problems hurt patients, reducing compliance and welfare in the market for anti-cholesterol drugs. My results provide the first quantification

---

23This is because compliance is most sensitive to patient utility for patients who are at the margin of indifference between complying and not, since mean utility from drugs is close to zero.
of the welfare losses from imperfect physician agency. Counterfactual contracts that better align physician and patient incentives can raise compliance, but they cannot achieve the agency-free level of utility. Fundamentally, the agency problem arises because doctors are much better informed about patient health—in this case match quality between patients and drugs—than are insurance companies, and it is difficult to write a contract that induces doctors to reveal this information.

These results relate to the large literature on physician agency, and more specifically conflicts of interest in the prescription drug market. The previous literature has largely demonstrated that doctors respond to factors that do not affect the patient’s utility, including financial incentives. My results build on this literature, establishing physician agency problems by showing that plan prices and advertising also affect the initial prescription decision, but have no direct impact on the utility patients receive from filling their prescription. The model adds to the literature by allowing me to quantify the welfare losses from autonomy and explore the impact of policies to remedy it.

My results suggest several questions for future research. First, one of the key sources of conflicts of interest is that physicians but not patients respond to plan prices. These prices likely reflect a partial solution to a different agency problem: doctors are also imperfect agents of insurance companies, and insurers set incentives to guide doctors towards cost-effective treatments. Like other researchers (Limbrock, 2011; Dickstein, 2012), I observe proxies for these incentives (i.e. plan prices), but not the incentives themselves. Thus, an important question is how these incentives work in practice.

Second, the model estimates indicate that doctors and patients have very different evaluations of drugs, in terms of both willingness to pay and perceived quality. These differences are important for determining the value of new drugs. A large literature studies the consequences of FDA policy and new drug entry for welfare, but this literature has not separated doctor and patient welfare. Thus, an open question is how patients value innovation or increased competition in drug markets.
Third, the model focuses on static efficiency as viewed by insurance companies, i.e. consumer utility less plan spending. A more complete view of efficiency recognizes that drugs are priced well above marginal cost, so that the “moral hazard” I document can actually be efficient. An interesting direction for future work is to account for these mark-ups in getting the welfare question right, and to explore in more detail how insurers set incentives for doctors.

Finally, insurance companies and health care agencies are beginning to encourage prescription drug compliance as a goal for doctors. For example, two of the National Quality Forum endorsed quality measures are compliance with asthma medications and compliance with anti-depressants (National Commitee for Quality Assurance, 2012). My results suggest that these policies will not only encourage doctors to follow up with patients and remind them to take their medication, but may also shift medication towards lower-cost drugs and other drugs that patients prefer.
Table 1: Molecule summary statistics, all years

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Probability of:</th>
<th>Molecule mean:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>Comply if</td>
<td>Switch from if</td>
<td>Switch to if</td>
<td>OOP Price</td>
<td>Plan Price</td>
</tr>
<tr>
<td></td>
<td>Prescription</td>
<td>Prescribed</td>
<td>Prescribed</td>
<td>Switch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipitor</td>
<td>38.3</td>
<td>56.4</td>
<td>9.4</td>
<td>19.5</td>
<td>0.58</td>
<td>2.45</td>
</tr>
<tr>
<td>Zocor</td>
<td>22.2</td>
<td>54.4</td>
<td>10.5</td>
<td>18.1</td>
<td>0.62</td>
<td>3.10</td>
</tr>
<tr>
<td>Mevacor</td>
<td>7.4</td>
<td>60.7</td>
<td>7.6</td>
<td>6.4</td>
<td>0.35</td>
<td>1.52</td>
</tr>
<tr>
<td>Pravachol</td>
<td>6.8</td>
<td>45.2</td>
<td>17.6</td>
<td>7.5</td>
<td>0.58</td>
<td>3.03</td>
</tr>
<tr>
<td>Crestor</td>
<td>6.7</td>
<td>46.3</td>
<td>13.0</td>
<td>9.5</td>
<td>0.78</td>
<td>2.04</td>
</tr>
<tr>
<td>Lescol</td>
<td>2.0</td>
<td>45.3</td>
<td>17.4</td>
<td>2.3</td>
<td>0.72</td>
<td>1.56</td>
</tr>
<tr>
<td>Ezetimibe Combo</td>
<td>5.0</td>
<td>51.0</td>
<td>10.4</td>
<td>18.6</td>
<td>0.69</td>
<td>1.95</td>
</tr>
<tr>
<td>Niacin Combo</td>
<td>0.8</td>
<td>39.2</td>
<td>17.6</td>
<td>1.9</td>
<td>0.67</td>
<td>2.19</td>
</tr>
<tr>
<td>Fibrate</td>
<td>8.3</td>
<td>42.6</td>
<td>17.8</td>
<td>11.4</td>
<td>0.49</td>
<td>1.62</td>
</tr>
<tr>
<td>Niacin</td>
<td>2.4</td>
<td>26.6</td>
<td>23.6</td>
<td>6.6</td>
<td>0.58</td>
<td>1.98</td>
</tr>
<tr>
<td>BAR</td>
<td>2.4</td>
<td>13.7</td>
<td>10.3</td>
<td>2.6</td>
<td>0.59</td>
<td>3.00</td>
</tr>
</tbody>
</table>

The sample consists of 296,760 people with a diagnosis of cholesterol-related disorders, and a prescription for a cholesterol drug, first prescribed in 2001-2005 or 2008-2009 and at least six months after entering the data, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans. Comply is defined as filling the initial prescription at least 6 times during this time period and no other anti-cholesterol drug; switching is defined as filling at least 6 prescriptions, including at least one for another drug than the first. OOP price and plan prices are dollars per day supply; detailing and DTCA are millions of dollars. The percentages do not add up to 100 because all probabilities condition on the drug being available.
Table 2: Multinomial logit estimates of the impact of patient price, plan price, and advertising on initial prescription

<table>
<thead>
<tr>
<th>Specification:</th>
<th>(1) Main</th>
<th>(2) No Molecule Fixed Effects</th>
<th>(3) Mol×Year Fixed Effects</th>
<th>(4) Mol×Health Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out of pocket price</td>
<td>$-1.382$</td>
<td>$-2.184$</td>
<td>$-1.093$</td>
<td>$-1.478$</td>
</tr>
<tr>
<td>Plan price</td>
<td>$-0.590$</td>
<td>$-0.057$</td>
<td>$-0.541$</td>
<td>$-0.580$</td>
</tr>
<tr>
<td>DTCA</td>
<td>$0.359$</td>
<td>$0.700$</td>
<td>$0.360$</td>
<td>$0.360$</td>
</tr>
<tr>
<td>Detailing</td>
<td>$0.761$</td>
<td>$0.695$</td>
<td>$0.737$</td>
<td>$0.737$</td>
</tr>
<tr>
<td>$\varepsilon_{OOP}$</td>
<td>$-0.757$</td>
<td>$-1.197$</td>
<td>$-0.599$</td>
<td>$-0.810$</td>
</tr>
<tr>
<td>$\varepsilon_{plan}$</td>
<td>$-1.189$</td>
<td>$-0.114$</td>
<td>$-1.091$</td>
<td>$-1.168$</td>
</tr>
<tr>
<td># Plans</td>
<td>383</td>
<td>383</td>
<td>383</td>
<td>383</td>
</tr>
<tr>
<td># People</td>
<td>296,760</td>
<td>296,760</td>
<td>296,760</td>
<td>296,760</td>
</tr>
</tbody>
</table>

Table shows the coefficients obtained from a multinomial logit regression of initial prescription choice against the indicated variables. All specifications also include an indicator for imputed price, and all specifications except column (2) and (3) also include molecule fixed effects. Subsequent columns include additional controls as indicated. Molecule×health fixed effects include a full set of interactions between molecule fixed effects and the health variables, measured as a set of dummy variables for female, age risk for heart disease, and cholesterol problems, diabetes, heart disease, and hypertension. Plan generosity is measured as the average price per day supply of drugs in the plan, and molecule×plan generosity includes a full set of interactions between molecule fixed effects and plan generosity. $\varepsilon_{OOP}$ and $\varepsilon_{plan}$ are the average elasticities of initial prescription choice probability with respect to out-of-pocket and plan price. The sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed in 2001-2005 or 2008-2009 and at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans. Robust standard errors, clustered on plan, are in parentheses.
Table 3: Logit estimates of the impact of patient price, plan price, and advertising on compliance

<table>
<thead>
<tr>
<th>Specification</th>
<th>(1) Main</th>
<th>(2) No plan Generosity</th>
<th>(3) No Mol Fixed effects</th>
<th>(4) Plan Fixed effects</th>
<th>(5) Plan FEs Mol×year FEs</th>
<th>(6) Plan FEs Mol×health Mol×generosity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out of pocket price</td>
<td>−0.498</td>
<td>−0.678</td>
<td>−0.888</td>
<td>−0.877</td>
<td>−0.921</td>
<td>−0.896</td>
</tr>
<tr>
<td>Plan price</td>
<td>−0.046</td>
<td>−0.037</td>
<td>−0.113</td>
<td>−0.075</td>
<td>−0.185</td>
<td>−0.072</td>
</tr>
<tr>
<td>DTC</td>
<td>0.047</td>
<td>0.038</td>
<td>0.001</td>
<td>−0.071</td>
<td></td>
<td>−0.067</td>
</tr>
<tr>
<td>Detailing</td>
<td>0.057</td>
<td>0.089</td>
<td>0.518</td>
<td>0.230</td>
<td></td>
<td>0.222</td>
</tr>
<tr>
<td>ε\text{OOP}</td>
<td>−0.141</td>
<td>−0.192</td>
<td>−0.252</td>
<td>−0.249</td>
<td>−0.261</td>
<td>−0.254</td>
</tr>
<tr>
<td>ε\text{Plan}</td>
<td>−0.051</td>
<td>−0.041</td>
<td>−0.126</td>
<td>−0.084</td>
<td>−0.207</td>
<td>−0.080</td>
</tr>
<tr>
<td># Plans</td>
<td>383</td>
<td>383</td>
<td>383</td>
<td>344</td>
<td>344</td>
<td>344</td>
</tr>
<tr>
<td># People</td>
<td>296,760</td>
<td>296,760</td>
<td>296,760</td>
<td>296,700</td>
<td>296,700</td>
<td>296,700</td>
</tr>
</tbody>
</table>

Table shows the coefficients obtained from a logit regression of compliance against the indicated variables. All columns except (2) and (5) also include plan generosity (as measured by the average price per day supply of branded drugs), and all columns except (3) and (4) also include molecule fixed effects. All columns include an indicator for imputed price, and a capitation indicator. The health controls include indicators for female, age-related heart disease risk, and any history of high cholesterol, diabetes, hypertension, or heart disease. ε\text{OOP} and ε\text{Plan} are the average elasticities of compliance probability with respect to out-of-pocket and plan price. The sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed in 2001-2005 or 2008-2009 and at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans. Robust standard errors, clustered on plan, are in parentheses.
Table 4: Differential responses by capitation status

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Initial Prescription</th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Not capitated</td>
<td>(2) Capitated</td>
</tr>
<tr>
<td>Out of pocket price</td>
<td>−1.232 (0.143)</td>
<td>−2.411 (0.745)</td>
</tr>
<tr>
<td>Plan price</td>
<td>−0.453 (0.033)</td>
<td>−0.494 (0.185)</td>
</tr>
<tr>
<td>DTCA</td>
<td>0.319 (0.056)</td>
<td>0.697 (0.195)</td>
</tr>
<tr>
<td>Detailing</td>
<td>0.820 (0.139)</td>
<td>0.544 (0.125)</td>
</tr>
<tr>
<td>ε&lt;sub&gt;OOP&lt;/sub&gt;</td>
<td>−0.669</td>
<td>−1.355</td>
</tr>
<tr>
<td>ε&lt;sub&gt;plan&lt;/sub&gt;</td>
<td>−0.919</td>
<td>−0.974</td>
</tr>
<tr>
<td># Plans</td>
<td>301</td>
<td>82</td>
</tr>
<tr>
<td># People</td>
<td>223,341</td>
<td>73,419</td>
</tr>
</tbody>
</table>

Table shows the coefficients obtained from multinomial logit regressions of the initial prescription choice, and logit regression of compliance, on the indicated variables, estimated separately by capitation status. Additional controls include plan generosity (as measured by the average price per day supply of branded drugs), molecule fixed effects, and an indicator for imputed price. ε<sub>OOP</sub> and ε<sub>plan</sub> are the elasticities of initial choice and compliance probabilities with respect to out-of-pocket and plan price. The sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans. Robust standard errors, clustered on plan, are in parentheses.
Table 5: Robustness to informational content of advertising

<table>
<thead>
<tr>
<th>Specification:</th>
<th>No lags</th>
<th>Three lags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Prescription</td>
<td>Compliance</td>
</tr>
<tr>
<td>Out of pocket price</td>
<td>$-1.370$</td>
<td>$-0.572$</td>
</tr>
<tr>
<td></td>
<td>$(0.250)$</td>
<td>$(0.112)$</td>
</tr>
<tr>
<td>Plan Price</td>
<td>$-0.599$</td>
<td>$-0.095$</td>
</tr>
<tr>
<td></td>
<td>$(0.083)$</td>
<td>$(0.034)$</td>
</tr>
<tr>
<td>DTCA</td>
<td>$0.416$</td>
<td>$0.047$</td>
</tr>
<tr>
<td></td>
<td>$(0.144)$</td>
<td>$(0.096)$</td>
</tr>
<tr>
<td>Detailing</td>
<td>$0.741$</td>
<td>$0.026$</td>
</tr>
<tr>
<td></td>
<td>$(0.231)$</td>
<td>$(0.161)$</td>
</tr>
<tr>
<td>Pubs</td>
<td>$-0.001$</td>
<td>$-0.001$</td>
</tr>
<tr>
<td></td>
<td>$(0.001)$</td>
<td>$(0.000)$</td>
</tr>
<tr>
<td>Pubs_{t-1}</td>
<td>$0.002$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>Pubs_{t-2}</td>
<td>$-0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>Pubs_{t-3}</td>
<td>$-0.002$</td>
<td>$-0.001$</td>
</tr>
<tr>
<td>p-value</td>
<td>$0.202$</td>
<td>$0.000$</td>
</tr>
<tr>
<td># Plans</td>
<td>383</td>
<td>383</td>
</tr>
<tr>
<td># People</td>
<td>296,760</td>
<td>296,760</td>
</tr>
</tbody>
</table>

Table shows the coefficients obtained from multinomial logit regressions of the initial prescription choice, and logit regressions of compliance, on the indicated variables, estimated separately by capitation status. Additional controls include plan generosity (as measured by the average price per day supply of branded drugs), molecule fixed effects, and an indicator for imputed price. The p-value is of the likelihood ratio test that the coefficients on PubMed mentions and its lags are jointly equal to zero. The sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans. Robust standard errors, clustered on plan, are in parentheses.
Table 6: Parameters of doctor’s and patient’s utility function

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Full Model</th>
<th>No Match Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient utility function:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out of pocket price</td>
<td>−0.459 (0.089)</td>
<td>−0.267 (0.116)</td>
</tr>
<tr>
<td>Out of pocket price × capitated</td>
<td>−0.526 (0.146)</td>
<td>−0.334 (0.105)</td>
</tr>
<tr>
<td>Capitated</td>
<td>0.249 (0.169)</td>
<td>0.299 (0.136)</td>
</tr>
<tr>
<td>Plan generosity</td>
<td>−0.403 (0.139)</td>
<td>−0.382 (0.147)</td>
</tr>
<tr>
<td>Imputed price</td>
<td>0.115 (0.237)</td>
<td>0.106 (0.127)</td>
</tr>
<tr>
<td>DTCA</td>
<td>0.144 (0.066)</td>
<td>0.067 (0.049)</td>
</tr>
<tr>
<td>Plan price</td>
<td>−0.032 (0.049)</td>
<td>0.045 (0.048)</td>
</tr>
<tr>
<td>Plan price × capitated</td>
<td>−0.093 (0.072)</td>
<td>−0.075 (0.067)</td>
</tr>
<tr>
<td>Detailing</td>
<td>−0.126 (0.172)</td>
<td>−0.088 (0.141)</td>
</tr>
<tr>
<td>Visit utility</td>
<td>−1.240 (0.062)</td>
<td>−1.315 (0.039)</td>
</tr>
<tr>
<td><strong>Doctor utility function:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight on patient utility</td>
<td>2.431 (0.231)</td>
<td>3.402 (1.276)</td>
</tr>
<tr>
<td>Plan price</td>
<td>−0.459 (0.116)</td>
<td>−0.564 (0.195)</td>
</tr>
<tr>
<td>Plan price × capitated</td>
<td>−0.189 (0.168)</td>
<td>−0.098 (0.285)</td>
</tr>
<tr>
<td>Detailing</td>
<td>1.091 (0.314)</td>
<td>0.947 (0.507)</td>
</tr>
</tbody>
</table>

Table shows the parameters of the doctor’s and patient’s utility function, obtained from simulated maximum likelihood estimation of the model of doctor-patient interactions. The “no match quality” specification imposes that $\sigma_d = 0$ for all drugs. The log likelihood at the solution of the full model is $-43,941$. The sample is a 5% random sample, stratified on plan, drawn from the same sample described in Table 1.8, and consists of 15,053 people in 383 plans. Standard errors, in parentheses, allow for arbitrary heteroskedasticity and autocorrelation within plan.
<table>
<thead>
<tr>
<th>Medicine</th>
<th>$\mu_{MD}$</th>
<th>SE</th>
<th>$\mu_P$</th>
<th>SE</th>
<th>$\sigma$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipitor</td>
<td>0.000</td>
<td></td>
<td>0.652</td>
<td>(0.153)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Zocor</td>
<td>0.156</td>
<td>(0.308)</td>
<td>0.747</td>
<td>(0.222)</td>
<td>0.571</td>
<td>(0.162)</td>
</tr>
<tr>
<td>Mevacor</td>
<td>-1.922</td>
<td>(0.646)</td>
<td>0.871</td>
<td>(0.443)</td>
<td>0.205</td>
<td>(0.476)</td>
</tr>
<tr>
<td>Pravachol</td>
<td>1.185</td>
<td>(0.461)</td>
<td>-0.268</td>
<td>(0.355)</td>
<td>0.810</td>
<td>(0.119)</td>
</tr>
<tr>
<td>Crestor</td>
<td>-0.734</td>
<td>(0.386)</td>
<td>0.275</td>
<td>(0.206)</td>
<td>0.535</td>
<td>(0.100)</td>
</tr>
<tr>
<td>Lescol</td>
<td>-1.365</td>
<td>(0.557)</td>
<td>0.338</td>
<td>(0.337)</td>
<td>0.393</td>
<td>(0.179)</td>
</tr>
<tr>
<td>Zetia/Vytorin</td>
<td>-2.360</td>
<td>(0.424)</td>
<td>0.864</td>
<td>(0.215)</td>
<td>0.357</td>
<td>(0.145)</td>
</tr>
<tr>
<td>Niacin Combo</td>
<td>-2.085</td>
<td>(1.025)</td>
<td>0.311</td>
<td>(0.653)</td>
<td>0.447</td>
<td>(0.280)</td>
</tr>
<tr>
<td>Fibrates</td>
<td>1.090</td>
<td>(0.368)</td>
<td>-0.342</td>
<td>(0.304)</td>
<td>0.762</td>
<td>(0.102)</td>
</tr>
<tr>
<td>Niacin</td>
<td>1.564</td>
<td>(0.504)</td>
<td>-0.837</td>
<td>(0.298)</td>
<td>0.574</td>
<td>(0.112)</td>
</tr>
<tr>
<td>BARs</td>
<td>4.426</td>
<td>(0.899)</td>
<td>-1.664</td>
<td>(0.435)</td>
<td>0.403</td>
<td>(0.164)</td>
</tr>
</tbody>
</table>

The table shows the estimated parameters of the match quality distribution and their standard errors, obtained from simulated maximum likelihood estimation of the model of doctor-patient interactions. $\mu_{MD}$ is the average quality as doctors perceived, over and above the average quality to patients, $\mu_P$. $\sigma$ is the standard deviation of match quality. Standard errors allow for arbitrary error correlation within plan. Note that $\mu_{MD}$ is normalized to zero, and $\sigma_{Lipitor}$ is normalized to 1.
Table 8: Goodness-of-fit

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Pr(Prescribe)</th>
<th>Pr(Comply)</th>
<th>Pr(Switch from)</th>
<th>Pr(Switch to)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data</td>
<td>Model</td>
<td>Data</td>
<td>Model</td>
</tr>
<tr>
<td>Lipitor</td>
<td>0.380</td>
<td>0.372</td>
<td>0.561</td>
<td>0.569</td>
</tr>
<tr>
<td>Zocor</td>
<td>0.223</td>
<td>0.214</td>
<td>0.553</td>
<td>0.537</td>
</tr>
<tr>
<td>Mevacor</td>
<td>0.073</td>
<td>0.070</td>
<td>0.593</td>
<td>0.560</td>
</tr>
<tr>
<td>Pravachol</td>
<td>0.070</td>
<td>0.072</td>
<td>0.471</td>
<td>0.501</td>
</tr>
<tr>
<td>Crestor</td>
<td>0.050</td>
<td>0.049</td>
<td>0.472</td>
<td>0.463</td>
</tr>
<tr>
<td>Lescol</td>
<td>0.020</td>
<td>0.019</td>
<td>0.464</td>
<td>0.442</td>
</tr>
<tr>
<td>Zetia/Vytorin</td>
<td>0.046</td>
<td>0.056</td>
<td>0.512</td>
<td>0.509</td>
</tr>
<tr>
<td>Niacin Combo</td>
<td>0.007</td>
<td>0.008</td>
<td>0.459</td>
<td>0.456</td>
</tr>
<tr>
<td>Fibrates</td>
<td>0.083</td>
<td>0.088</td>
<td>0.428</td>
<td>0.454</td>
</tr>
<tr>
<td>Niacin</td>
<td>0.025</td>
<td>0.030</td>
<td>0.278</td>
<td>0.283</td>
</tr>
<tr>
<td>BARs</td>
<td>0.023</td>
<td>0.023</td>
<td>0.141</td>
<td>0.100</td>
</tr>
</tbody>
</table>

Table shows the probability that each molecule is prescribed and patients comply or switch, conditional on having a prescription for the indicated molecule. The data column gives the raw means from the estimation sample. The model column shows the prediction implied by the model estimates.
Table 9: Impact of agency

<table>
<thead>
<tr>
<th>Scenario:</th>
<th>(1) Baseline</th>
<th>(2) No Agency</th>
<th>(3) No Incentives</th>
<th>(4) No Detailing</th>
<th>(5) No Fixed Effects</th>
<th>(6) No logits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Change relative to baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comply</td>
<td>53.1</td>
<td>6.5</td>
<td>-0.1</td>
<td>-1.8</td>
<td>1.1</td>
<td>3.3</td>
</tr>
<tr>
<td>Switch</td>
<td>11.1</td>
<td>1.1</td>
<td>-0.3</td>
<td>0.1</td>
<td>1.9</td>
<td>1.2</td>
</tr>
<tr>
<td>Quit</td>
<td>35.8</td>
<td>-7.6</td>
<td>0.4</td>
<td>1.7</td>
<td>-2.9</td>
<td>-4.5</td>
</tr>
<tr>
<td>Plan spending</td>
<td>563</td>
<td>12.8</td>
<td>48.4</td>
<td>-42.1</td>
<td>-33.4</td>
<td>45.9</td>
</tr>
<tr>
<td>Out of pocket spending</td>
<td>124.8</td>
<td>0.2</td>
<td>2.4</td>
<td>-7.6</td>
<td>2.7</td>
<td>7.3</td>
</tr>
<tr>
<td>Consumer surplus</td>
<td>1094.3</td>
<td>148.4</td>
<td>-5.2</td>
<td>-34.6</td>
<td>45.8</td>
<td>82.1</td>
</tr>
<tr>
<td>Consumer surplus (mistakes)</td>
<td>278.1</td>
<td>141.8</td>
<td>-3.7</td>
<td>-29.9</td>
<td>35.7</td>
<td>74.3</td>
</tr>
<tr>
<td>Net surplus</td>
<td>531.3</td>
<td>135.5</td>
<td>-53.6</td>
<td>7.5</td>
<td>79.2</td>
<td>36.1</td>
</tr>
<tr>
<td>Δ LDL-C, prescribed</td>
<td>1.07</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.08</td>
<td>-0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Δ LDL-C, realized</td>
<td>0.71</td>
<td>0.05</td>
<td>0.01</td>
<td>-0.05</td>
<td>-0.01</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Column (1) shows the level of the indicated statistic at the baseline parameter estimates, and columns (2)-(6) show the change relative to baseline as I shut down the indicated aspects of physician agency. LDL cholesterol is measured in mmol/L, and higher values indicate a greater decrease. Net surplus is consumer surplus less plan spending.
Figure 1: Out-of-pocket and plan prices, by plan, 2008

Panel A: Lipitor vs. Crestor (OOP)
Panel B: Zocor vs. Mevacor (OOP)
Panel C: Lipitor vs. Crestor (Plan)
Panel D: Zocor vs. Mevacor (Plan)

Figure shows the prices of Lipitor, Crestor, Zocor and Mevacor (Panel B) in 2008.

Figure 2: Out of sample fit for Lipitor, Zocor, and Mevacor

Figure shows the out-of-sample trends for Lipitor, Zocor, and Mevacor. The data are the actual means for the random hold out sample.
Figure 3: Impact of alternative contracts on compliance and patient welfare

Panel A: Compliance

Panel B: Utility

Panel C: Plan drug spending

Panel D: Bonus spending

Panel E: Utility − total plan spending

Panel F: Total surplus

### Bonus

Figure shows several outcomes under counterfactual incentive schemes as I vary the strength of the bonus. The “quality” contract pays doctors a bonus for prescribing the best drug, based on observables. The “comply” contract pays the bonus if the patient complies with the prescription. The “threshold” contract pays the bonus if at least 40% of the doctor’s patients comply. “Total surplus” is equal to patient utility less plan drug spending.
CHAPTER 2: Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test

2.1. Introduction

In a traditional model of workers’ earnings or labor supply choices, individuals optimize their behavior frictionlessly in response to policies that affect their incentives.\(^1\) However, several recent papers have suggested that individuals face frictions in adjusting behavior to policy (Chetty et al., 2009, 2011, 2012, 2013; Chetty, 2012; Kleven and Waseem, 2013). Costs of adjusting behavior help to govern the welfare cost of taxation (Chetty et al., 2009), and they also help to explain heterogeneity across contexts in the observed elasticity of earnings with respect to the net-of-tax rate (Chetty et al., 2011, 2013; Chetty, 2012).\(^2\)

This paper develops evidence on the existence, nature and size of frictions in adjusting earnings in response to policy. The U.S. Social Security Annual Earnings Test (AET) represents a promising environment for studying these questions. This setting provides a useful illustration of many issues—such as the development and application of a methodology for estimating elasticities and adjustment costs simultaneously—that are applicable to studying earnings responses to policy more broadly. The AET reduces Social Security Old Age and Survivors Insurance (OASI) claimants’ current OASI benefits as a proportion of earnings, once an individual earns in excess of an exempt amount. For example, for OASI claimants aged 62-65 in 2013, current OASI benefits are reduced by 50 cents for every extra dollar earned above $15,120. The AET may lead to very large effective benefit reduction rates (BRRs) on earnings above the exempt amount, creating a strong incentive for many individuals to “bunch” at the convex kink in the budget constraint located at the exempt amount (Burtless and Moffitt, 1985; Friedberg, 1998, 2000).\(^3\) Reductions in current benefits due to the AET sometimes lead to increases in later benefits; nonetheless, as we discuss in

---

\(^1\) This chapter is co-authored with Alexander M. Gelber and Damon Jones.  
\(^2\) The net-of-tax rate is defined as one minus the marginal tax rate (MTR). Literature including Altonji and Paxson (1988) examines hours constraints in the context of labor supply.  
\(^3\) Other papers on the AET include Gruber and Orszag (2003) and Song and Manchester (2007).
detail in Section 2.2, several factors may explain why individuals’ earnings still respond to the AET.

The AET is an appealing context for studying earnings adjustment for at least three reasons. First, bunching at the AET kink is easily visible on a graph, allowing credible documentation of behavioral responses.\(^4\) Second, the AET represents one of the few known kinks at which bunching occurs in the U.S.; indeed, our paper represents the first study to find robust evidence of bunching among the non-self-employed at any kink in the U.S.\(^5\) Third, the AET is important to policy-makers in its own right, as it is a significant factor that affects the earnings of the elderly in the U.S.

We make three main contributions to understanding adjustment frictions. First, we document that earnings adjustment frictions exist in the U.S., by showing that in some contexts individuals do not adjust immediately to changes in AET. We focus particularly on cases in which a kink in the effective tax schedule disappears, either because individuals reach an age at which they are no longer subject to the AET, or because legislative changes remove the AET for some groups.\(^6\) We focus on the disappearance of kinks because in the absence of adjustment frictions, removal of a convex kink in the effective tax schedule should immediately lead to a complete lack of bunching at the earnings level associated with the former kink; thus, any observed delay in reaching zero bunching should reflect adjustment frictions. We observe clear evidence of delays in some contexts, consistent with the existence of adjustment frictions. Nonetheless, across several contexts—including both anticipated and unanticipated changes in policy—the vast majority of individuals’ adjustment occurs

---

\(^4\) Other papers have examined bunching in the earnings schedule, including Blundell and Hoynes (2004) and Saez (2010). Saez shows that the amount of bunching can be related to the elasticity of earnings with respect to the net-of-tax rate.

\(^5\) The lack of bunching at other kinks is consistent with the existence of adjustment costs, although this finding could also be explained by other factors such as a low elasticity of earnings with respect to the net-of-tax rate. As we discuss in greater detail in Section 2.6, Chetty et al. (2013) do find evidence of more diffuse earnings responses to the Earned Income Tax Credit among the non-self-employed.

\(^6\) For consistency with the previous literature on kink points that has focused on the effect of taxation, we sometimes use "tax" as shorthand for "tax-and-transfer," while recognizing that the AET reduces Social Security benefits and is not administered through the tax system. The "effective" marginal tax rate is affected by the AET BRR, among other factors.
within at most three years. Adjustment appears even faster in certain contexts.

Second, we assess the mechanisms that underlie the patterns of adjustment we observe, in order to build a model consistent with these descriptive patterns. We assess the extent to which employers play a role in coordinating individual responses to the AET by offering jobs with earnings at the AET exempt amount.\textsuperscript{7} In our main period of study, we find little evidence that those too young to claim benefits (and therefore not subject to the AET) bunch at the kink, suggesting that the primary responses to the AET are mediated by employees’ choices. We also find evidence that the individuals who respond to the removal of the AET are primarily those locating at the kink prior to its removal, suggesting that these individuals are particularly responsive. Others subject to the AET appear to be unresponsive, suggesting heterogeneity in adjustment costs or elasticities in the population.

Third, we specify a model of earnings adjustment consistent with the descriptive evidence that allows us to estimate a fixed adjustment cost and the elasticity of earnings with respect to the effective net-of-tax rate. Recent work demonstrating the importance adjustment costs has raised the question of how to estimate both the elasticity and adjustment cost simultaneously. We develop tractable methods that allow estimation of elasticities and adjustment costs with kinked budget sets. This complements Kleven and Waseem (2013), who develop a method to estimate related parameters in the presence of a notch in the budget set. Our method relies on the fact that the amount of bunching at a kink increases with the elasticity but decreases with the adjustment cost. This prevents estimation of both parameters using a single cross-section—since a small amount of bunching, for example, could be consistent with either a low elasticity or a high adjustment cost—but with with two or more cross-sections of individuals facing different tax rates in the region of the kink, we can specify two or more equations and find the values of two variables (the elasticity and the adjustment cost).\textsuperscript{8} The model Saez (2010) describes how bunching should vary between

\textsuperscript{7}Due to interactions between adjustment costs for workers and hours constraints set by firms, some individuals may bunch at a kink even though they are not directly subject to the policy that creates the kink. Chetty et al. (2011) document that employers play such a role in Denmark.

\textsuperscript{8}Under certain approximations that we later specify, this can yield a system of linear equations that can
two different kinks in a frictionless setting, and the extent to which observed bunching deviates from this pattern is attributed to the adjustment cost. Intuitively, inertia due to an adjustment cost leads to an excess amount of bunching after a kink in the budget set becomes less sharply bent (or disappears altogether). Our primary estimation method uses the degree of such inertia (in combination with the initial amount of bunching at the kink) in estimating the size of the adjustment cost (and elasticity).\(^9\)

We apply our method to data spanning the decrease in the AET benefit reduction rate from 50 percent to 33.33 percent in 1990 for those aged 66 to 69, as well as two settings in which the AET no longer applies for certain groups (at age 70 in the 1990-1999 period, and for ages 66-69 beginning in the year 2000). In a baseline specification examining the 1990 change, we estimate that the fixed adjustment cost is $152.08 (in 2010 dollars)—if the gains exceed this level, then the individual adjusts earnings—and that the earnings elasticity with respect to the net-of-tax share is 0.23. This specification examines data on individuals in 1989 and 1990; thus, our estimated adjustment cost represents the cost of adjusting earnings in the first year after the policy change. Other empirical strategies show results in the same range. By contrast, when we constrain the adjustment cost to be zero in 1990, we estimate a statistically significantly higher earnings elasticity of 0.39 in the baseline specification (69 percent larger than the unconstrained estimate).\(^{10}\) These estimates suggest that while adjustment costs are modest in our setting, they have the potential to change elasticity estimates substantially, thus demonstrating that it can be important to incorporate adjustment costs when estimating elasticities. Nonetheless, our estimates are specific to our setting, and adjustment costs and elasticities may be substantially different (larger or smaller) in other contexts.

---

\(^9\) As we describe in detail later, this intuition applies to our primary empirical approach, the "Sharp Change" approach.

\(^{10}\) In our context, it makes sense that the estimated elasticity is higher when we do not allow for adjustment costs than when we do, as adjustment costs keep individuals bunching at the kink even though tax rates have fallen.
In the course of investigating these issues relating to frictions and earnings adjustment, we build on previous literature on the AET to provide new evidence that enriches our understanding of how the AET affects earnings. First, we systematically investigate each of the major AET policy changes since 1961. Second, we use SSA administrative data with a full sample of 13,612,313 observations on 619,580 individuals, building on certain previous studies that use survey data. Third, our study is the first to estimate bunching in the context of the AET through a method similar to Saez (2010). Fourth, we present evidence on individuals’ earnings reaction to changes in the Delayed Retirement Credit. Fifth, we investigate whether mortality expectations help drive individuals’ earnings responses to the AET by estimating the pattern of life expectancy around the exempt amount. Sixth, we investigate whether individuals change earnings in response to the AET by changing jobs or by changing earnings levels within a job, as well as whether employers coordinate employees on the AET exempt amount. Finally, we show that individuals serially bunch at the exempt amount.

The remainder of the paper is structured as follows. Section 2.2 describes the policies we examine. Section 2.3 presents a framework for analyzing the behavioral response to these policies and describes our empirical strategy for quantifying bunching. Section 2.4 describes our data. Section 2.5 presents empirical evidence on the earnings response to changes in the AET. Section 2.6 explores certain mechanisms underlying the behavioral responses. Section 2.7 specifies a tractable model of earnings adjustment and estimates the fixed adjustment cost and elasticity simultaneously. Section 2.8 concludes with discussion and avenues for future work.

2.2. Policy Environment

Figure 4 shows key features of the AET rules from 1961 to 2009. The AET became less stringent over this period. The dashed line and right vertical axis show the benefit reduction rate. From 1961 to 1989, every dollar of earnings above the exempt amount reduced OASI
benefits by 50 cents (until OASI benefits reached zero).\textsuperscript{11} In 1990 and after, the benefit reduction rate fell to 33.33 percent for beneficiaries above the Normal Retirement Age (NRA).\textsuperscript{12} The figure also shows that the AET applied to a narrower set of ages over time. In 1961, the AET applied to ages 62-71; starting in 1983, the AET was eliminated for 70-71 year-olds; and starting in 2000, the AET was also eliminated for those NRA and above. The solid line and left vertical axis show the real exempt amount. Between 1961 and 1971, the exempt amount rose with price inflation. Beginning in 1972, the exempt amount typically rose faster than inflation. Starting in 1978, the AET had different rules for beneficiaries younger than NRA and those at least NRA but younger than the maximum age subject to the AET. Subsequently, the exempt amount rose much faster on average for beneficiaries NRA and older than for younger beneficiaries.\textsuperscript{13}

We later model the AET as creating a positive implicit marginal tax rate for some individuals, consistent with the empirical finding that some individuals bunch at AET kinks, certain theoretical considerations we describe below, and previous literature. In the empirical section, we explore evidence relating to certain mechanisms that explain this response.\textsuperscript{14}

\textsuperscript{11}In addition to this threshold, until 1972 there was a second, higher earnings threshold over which the benefit reduction rate was 100 percent (Social Security Annual Statistical Supplement 2012). The second threshold is well above the first threshold, ranging from 25 percent to 80 percent higher depending on the year.

\textsuperscript{12}The NRA, the age at which workers can claim their full OASI benefits, is 65 for those born 1937 and before, rises by two months a year for cohorts between 1938 and 1943, is constant at age 66 for cohorts between 1943 and 1954, and rises by two months a year until reaching age 67 for those born in 1960 and later.

\textsuperscript{13}The exempt amount has not been a "focal" earnings level—such as $1,000, $5000, or $10,000—that could lead to bunching at the exempt amount even in the absence of AET. Indeed, in our main period of study we find no evidence of bunching at the exempt amount among those younger than the ages to which the AET applies. In 2000 and subsequently, those in the year of attaining NRA face the AET in the months prior to such attainment, but they are subject to a higher exempt amount and a benefit reduction rate of 33.33 percent.

\textsuperscript{14}In this paper, we focus on the marginal incentives created by the AET and intensive margin responses, following previous literature based on the technique of Saez (2010). Other important decisions could include the choice of whether to earn a positive amount, or the decision to claim OASI. We abstract from the claiming decision by examining a sample of OASI claimants, following previous literature such as Friedberg (1998, 2000); however, it is worth noting that that if the AET affects the claiming decision, there is no a \textit{a priori} reason that this change in claiming should increase or decrease the magnitude of the bunching responses we document among claimants. Moreover, we add to previous literature by showing in Appendix Figure 35 that the hazard of claiming at year $t + 1$ is smooth around the exempt amount at year $t$, indicating no evidence that claimants come disproportionately from close to or far from the kink. We discuss the claiming decision further in the Appendix. We examine the extensive margin response in a companion paper (Gelber et al., 2013). (Cogan (1981) is a classic reference on fixed costs of adjustment in the extensive margin choice.)
When current OASI benefits are lost to the AET, future scheduled benefits are increased in some circumstances. This is sometimes called "benefit enhancement." Benefit enhancement can reduce the effective tax rate associated with the AET, in particular for those individuals considering earning enough to trigger the enhancement in the post-1972 period, as we describe in detail in Appendix A.7 and briefly in this section.

Prior to 1972, the AET caused a pure loss in benefits for those NRA and above, as there was no benefit enhancement for these individuals. For beneficiaries subject to the AET aged NRA and above, a one percent DRC was introduced in 1972, meaning that each year of benefits foregone led to a one percent increase in future yearly benefits. The DRC was raised to three percent in 1982 and gradually rose to eight percent for cohorts reaching NRA from 1990 to 2008 (though the AET was eliminated in 2000 for those above NRA). An increase in future benefits between seven and eight percent is approximately actuarially fair on average, meaning that an individual with no liquidity constraints and average life expectancy should be indifferent between either claiming benefits now or delaying claiming and receiving higher benefits once she begins to collect OASI (as Diamond and Gruber (1999) show with respect to the actuarial adjustment).

As we describe further in Appendix A.7, future benefits are only raised due to the DRC when annual earnings are sufficiently high that the individual loses an entire month’s worth of OASI benefits due to the reductions associated with the AET (Friedberg, 1998; Social Security Administration, 2012). In particular, an entire month’s benefits are lost once the individual earns \( z^* + (MB/\tau) \) or higher, where \( z^* \) is the exempt amount, \( MB \) is the monthly benefit, and \( \tau \) is the AET benefit reduction rate. With a typical monthly benefit of $1,000 and a benefit reduction rate of 33.33 percent, one month’s benefit enhancement occurs when the individual’s annual earnings are $3,000 (= $1000/0.3333) above the exempt amount. For example, if an individual born in 1933-1934 earned at or just above this amount in years when she was subject to the DRC, future benefits were raised by 0.46 percent (but no increase occurs if the individual earns below this amount). As a result, at
or just above the AET threshold, earning an extra dollar does not affect subsequent OASI benefits. Thus, benefit enhancement is only relevant to an individual considering earning substantially in excess of the exempt amount. Indeed, we later describe suggestive evidence of both little systematic bunching reaction to changes in the DRC and little relationship between bunching and life expectancy.\footnote{Later, our empirical specification alternatively assumes that benefit enhancement does not (or does) affect the AET implicit marginal tax rate, and we find similar patterns in both specifications.}

Thus, the AET could affect the earnings decisions of those NRA and above for a number of reasons. As we have discussed, for those to whom benefit enhancement is effectively irrelevant (because they are only considering earning sufficiently near to the AET that they would not receive benefit enhancement through increasing earnings), the marginal incentives they face are not affected by benefit enhancement. For those to whom benefit enhancement is relevant (because they are considering earning in a region well above the AET exempt amount, thus triggering benefit enhancement), the AET could also affect decisions, for several reasons. First, the AET was on average roughly actuarially fair only beginning in the late 1990s. Indeed, prior to 1972, the AET represented a pure loss in benefits for those NRA and above. Furthermore, those whose expected lifespan is shorter than average should expect to collect OASI benefits for less long than average, implying that the AET is more financially punitive (though we ultimately find no evidence consistent with this hypothesis). Liquidity-constrained individuals or those who discount faster than average could also reduce work in response to the AET. Finally, many individuals may also not understand many features of the AET or other aspects of OASI (Liebman and Luttmer, 2011).

For beneficiaries under NRA, the actuarial adjustment raises future benefits whenever an individual earns any amount over the AET exempt amount.\footnote{Social Security Administration (2012), Section 728.2; Gruber and Orszag (2003). Formally, the number of months’ worth of benefit enhancement received by OASI recipients is $\text{floor}(\tau \cdot (z - z^*)/MB)$ for those NRA and above, and $\text{ceiling}(\tau \cdot (z - z^*)/MB)$ for those below NRA. See Appendix A.7 for more details.} Future benefits are raised by 0.55 percent per month of benefits withheld for the first three years of AET assessment.
This creates a notch in the budget set at the AET threshold—as opposed to a simple kink, whose properties we explore in our theory sections. Our discussion of the effects of kinks therefore does not directly apply to pre-NRA ages. Thus, in our estimates of elasticities and adjustment costs, we limit the sample to ages NRA and above, for which the budget set (in the region of the exempt amount) is a kink rather than a notch.

2.3. Initial Bunching Framework

As a preliminary step, we begin with a model with no frictions. This model is well-known and described in detail elsewhere, but we briefly describe it here and in more detail in Appendix A.11. After we have presented our empirical results, we specify a model with frictions that is consistent with the descriptive patterns we document.

Appendix Figure 14 shows the budget constraint and incentives created by the AET for those NRA and above in the frictionless case. Start with a linear tax (Panel A) at a rate of $\tau$. Now, suppose the AET is introduced (on top of pre-existing taxes), so that the marginal net-of-tax rate decreases to $1 - \tau - d\tau$ for earnings above a threshold $z^*$ (Panel B). For small $d\tau$, individuals earning in the neighborhood above $z^*$ reduce their earnings. If ability is smoothly distributed, a range of individuals initially locating between $z^*$ and $z^* + \Delta z$ (as depicted in the density in Panel C) will instead locate exactly at $z^*$, due to the discontinuous jump in the marginal net-of-tax rate at $z^*$. In fact, we find empirically that these individuals locate in the neighborhood of $z^*$, as shown in Panel D.

To measure the amount of bunching, we use a technique similar to Chetty et al. (2011) and Kleven and Waseem (2013), which we illustrate in Appendix Figure 15 and describe further in the Appendix. The x-axis measures before-tax income, $z$, while the y-axis measures the density of earnings. In Panel A, we show that the ex-post density of earnings in the presence of a kink is comprised of a number of groups. Those in the region labeled X in

\[17\] Saez (2010) describes this model in greater detail. This work follows earlier work on estimation of labor supply responses on nonlinear budget sets, including Burtless and Hausman (1978) and Hausman (2004). Moffitt (1990) surveys these methods.
the figure ("bunchers") have optimal earnings above $z^*$ under the lower rate of $\tau$ and at $z^*$ under the higher rate of $\tau + d\tau$. Those in the region labeled Y in the figure consist of individuals whose optimal earnings are below $z^*$ under a lower marginal tax rate of $\tau$, as well as other individuals whose optimal earnings are above $z^*$ under the higher marginal tax rate of $\tau + d\tau$. Panel B shows that to estimate the size of region X, we must estimate the ex post density and subtract the mass associated with Group Y.

As described further in Appendix A.8, we divide the data into $800$ bins and estimate a seven-degree polynomial through the densities associated with the bins. In estimating this polynomial, we control for dummies for being in the seven bins nearest to the kink,\(^{18}\) to capture the bunching near the kink that we wish to ignore when we estimate the counterfactual polynomial density. Our estimate of bunching, $B$, is the difference between the mass in these seven bins and the area under the polynomial in this $5600$-wide region. We estimate confidence intervals through a bootstrap procedure that we describe further in Appendix A.11.8 (and the results are similar under the delta method). We report our bunching amount, $B$, normalized by the share of individuals in the neighborhood $[z^* - \delta, z^* + \delta]$ who belong to Group Y (which we approximate as the area under our polynomial over this range).\(^{19}\)

Some apparent limitations of our approach are worth discussion. First, following previous literature on earnings responses to kinks, we do not take into account other choices that could affect earnings in the long run, such as human capital accumulation. However, human capital accumulation is likely to be less important for the older workers we study than it is for the population as a whole. Second, other programs—such as Medicaid, Supplemental Security Income, Disability Insurance, or taxes such as unemployment insurance payroll taxes—create earnings incentives near the bottom of the earnings distribution. While we

\(^{18}\)This implies that our estimate of excess bunching is driven by individuals locating within $2800$ of the kink (as the central bin runs from $400$ under the kink to $400$ above the kink). We discuss this issue further in the Appendix. As we show in the Appendix, we have also experimented with other bandwidths, which yield similar results.

\(^{19}\)While we show this excess bunching at the kink as arising in a frictionless model here, this technique is also suited to measuring the excess bunching at the kink arising in a model with frictions (as the key in either setting is that there is excess bunching at the kink, which this technique can measure in either case).
acknowledge that other incentives represent a concern in principle—applicable to most of the literature on bunching at kinks—we also note that the kinks created by these programs are typically inapplicable or safely far away from the AET convex kink.\textsuperscript{20} The results show very clear evidence of bunching at the AET kink and no visible, systematic evidence of bunching in other regions close to the AET kink. Third, we follow the previous work and largely do not distinguish among the potential reasons for a response to the AET. Following previous literature, our bunching framework presumes that consistent with the empirical evidence documenting clear responses to the incentives created by the AET, certain individuals treat the AET as creating some effective marginal tax rate above the exempt amount.

Finally, the results are specific to the AET and may not generalize outside of this context. We estimate the speed of adjustment among those initially bunching at a kink, a group that our empirical results suggest is more responsive to the AET than other groups. We therefore believe it is all the more interesting that we still find evidence of modest adjustment frictions among this group whose initial bunching indicates a substantial degree of flexibility (enough to locate at the kink initially). Furthermore, our estimation procedure relies on estimating bunching at more than one kink, and therefore it has the potential to incorporate information on the responses of individuals across a wide range of the income distribution (across multiple kinks).

2.4. Data

We primarily rely on the restricted-access Social Security Administration Master Earnings File (MEF) and Master Beneficiary Record (MBR), described more fully in the Appendix. The data contain a complete longitudinal earnings history with yearly information on earn-

\textsuperscript{20}We have found that many other incentives, including income tax rates, are smooth on average around the AET convex kink. The AET also potentially creates other distortions that we discuss further in Appendix A.7, including: a slight notch for those NRA age and above every time an entire month’s worth of benefits are lost due to the Delayed Retirement Credit; an additional non-convex kink in the budget constraint at the point at which OASI benefits are fully phased out; and a notch for those below the NRA for every month of withheld benefits that triggers the actuarially adjustment described above. However, in the case of those NRA and above that we focus on, these incentives are not likely to be relevant for potential bunchers and do not appear to be empirically relevant (as we discuss elsewhere). For these reasons, we abstract from these additional features but discuss such incentives further when we present our empirical evidence.
ings since 1951; the type and amount of yearly Social Security benefits an individual receives; year of birth; the year (if any) that claiming began; and sex (among other variables). Separate information is available on self-employment earnings and non-self-employment earnings. Prior to 1978, the data measure annual Federal Insurance Contribution Act (FICA) earnings. Starting in 1978, the measure of earnings in the MEF reflects total wage compensation, as reported on Internal Revenue Service (IRS) forms. Our dataset is a one percent random sample of all Social Security numbers in the MEF, keeping all available years of data for each individual sampled.

Several features of the data are worth discussion. First, these administrative data allow large sample sizes and are subject to little measurement error. Second, earnings (as measured in the dataset) are the base for FICA taxes and are not subject to manipulation through tax deductions, credits, or exemptions. Third, because earnings are taken from the W-2 form, they are subject to third-party reporting among the non-self-employed; third-party reporting has been found in the literature to greatly reduce evasion Kleven et al. (2011). This limits the degree to which observed bunching among the non-self-employed—to whom we limit our sample—could reflect reporting issues. Fourth, the data do not contain information on hours worked or amenities at individuals’ jobs.

Table 10 shows summary statistics for the sample of individuals aged 18-75, and for the sample that we typically focus on, all those aged 62-69 who claimed by age 65. In both samples, we exclude those with self-employment income. The larger (smaller) sample has 13,612,313 (1,595,139) observations on 619,580 (545,615) individuals. 56 percent (57 percent) of the sample is male. 50 percent (57 percent) of observations have positive earnings. Mean earnings in the sample (conditional on having positive earnings) is $37,492.28 ($29,485.08). Excluding those with self-employment income reduces the sample size (rela-

\[ \text{Table 10 shows summary statistics for the sample of individuals aged 18-75, and}\]

\[ \text{for the}\]

\[ \text{sample that we typically focus on, all those aged 62-69 who claimed by age 65.}\]

\[ \text{In both}\]

\[ \text{samples, we exclude those with self-employment income. The larger (smaller) sample}\]

\[ \text{has}\]

\[ 13,612,313\ (1,595,139)\ \text{observations on}\]

\[ 619,580\ (545,615)\ \text{individuals. 56 percent (57 percent) of the sample is}\]

\[ \text{male. 50 percent (57 percent) of observations have positive}\]

\[ \text{earnings. Mean earnings in the sample (conditional on having positive earnings) is}\]

\[ $37,492.28\ ($29,485.08).\]
tive to the full population) by 18% among 18-75 year olds and by 12% among 62-69 year olds. Note that median earnings among our main sample of 62-69 year-olds ($17,739.68) is not far from the AET exempt amount; the population our study examines is in a range with a thick density of earnings that is not far from the median.

The second dataset we use is the Longitudinal Employer Household Dynamics (LEHD) dataset of the U.S. Census (McKinney and Villhuber, 2008; Abowd et al., 2009), described further in the Appendix. The data are based on unemployment insurance earnings records and longitudinally follow workers’ earnings over time. The data have information on around nine-tenths of workers in covered states and their employers, though we are only able to use data on a 20 percent random subsample of these individuals. We use these data primarily in order to link employees to employers, as the SSA data that we have access to have no information about individuals' employers. We secondarily use these data because the sample size we are able to obtain in the LEHD is much larger than the (large) sample size we obtain in the SSA data.23

2.5. Earnings Response to Policy

2.5.1. Descriptive Evidence from Policy Variation Across Ages

We first examine the pattern of excess bunching across ages, in order to determine how quickly individuals respond to changes in policy across ages and whether they face delays in responding (consistent with the existence of adjustment frictions). Empirical work often estimates only short-run responses to changes in policy (see Saez et al. (2012), for a review of literature on earnings responses to taxation). If individuals are able to respond more (less) in the long run than in the short run, then this large body of work would under-

---

23 The LEHD lacks information on whether a given individual is claiming OASI. Nonetheless, the ultimate importance of this shortcoming is limited. In our SSA data, 97 percent of people claim by age 69, so it is a safe assumption that the great majority of the individuals observed in the LEHD data of the ages we are interested in (primarily ages 69 and 70) have claimed OASI. The magnitude of the bunching we observe is likely to be slightly understated relative to the magnitude we would measure in the population of OASI claimants, as the results include non-claimants in the sample. However, our primary interest concerns the patterns of responses to the AET across ages and over time, which prove to be visually and statistically clear in the LEHD.
estimate (over-estimate) long-run responses. Moreover, most empirical specifications have related an individual’s tax rate in a given year to the individual’s earnings in that year. In order to choose the appropriate time horizon over which to study behavioral responses to policy, it is necessary to establish how long it takes to respond to policy changes.

Subsequent to 1982, the AET applies to ages 62-69. The policy changes at ages 62 and 70—the imposition and removal of the AET—are "anticipated," by which we refer to changes that would be anticipated by those who have knowledge of the relevant policies. We begin by examining the period 1990-1999. Figure 5 plots earnings histograms for each age from 59 to 73. Earnings are measured along the x-axis, relative to the exempt amount, which is shown by a vertical line.24

Figure 5 shows clear visual evidence of substantial excess bunching from ages 62-71.25 Figure 6 plots the point estimates and 95 percent confidence intervals for bunching at each age. Bunching is statistically significantly different from zero at each age in the 62-71 range (p < 0.01 at all ages). We find no evidence of adjustment in anticipation of future changes in policy, as those younger than 62 do not bunch.26

We do find evidence that unbunching takes more than one year, however, as those ages 70 and 71 show modest bunching. Figure 5 shows that the density near the kink is raised at these ages, and Figure 6 shows that the estimates of bunching are statistically significantly different from zero. Three other considerations also indicate that this reflects excess bunching at these ages. First, we show later that the statistically significant positive estimates are robust to varying the degree of the polynomial, the excluded region, and the bandwidth used in the estimates. Second, the distributions at other ages not affected by the AET

---

24 For ages younger than 62, we define the (placebo) kink in a given year as the kink that applies to pre-NRA individuals in that year. For individuals 70 and above, we define the (placebo) kink in a given year as the kink that applies to post-NRA individuals in that year.

25 As discussed above, in this period individuals aged 62 to 64 faced a notch in the budget set (due to the actuarial adjustment of benefits) at the exempt amount; thus, the incentives they faced were different than those for individuals aged 65 to 69. However, the histograms show no evidence of a spike in earnings just above the kink (as one would predict if they respond to the incentives created by this notch).

26 If the cost of adjustment in each year rose with the size of adjustment and this relationship were convex, we would expect anticipatory adjustment.
that represent reasonable counterfactuals (such as 61 or 73), show nearly perfectly smooth earnings distributions, suggesting that the excess mass near the kink at ages 70 and 71 would not arise in the absence of the AET. Third, Appendix Figure 16 shows that the mean percentage change in earnings from age 70 to 71 shows a modest spike near the exempt amount, consistent with continued earnings adjustment from age 70 to age 71 among those near the kink at age 70. We find it striking that even among the group bunching prior to age 70—that (the data reveal) are able to adjust earnings to the kink—we still find evidence of modest adjustment frictions.

Figure 6 shows that excess bunching is substantially lower at age 65 than surrounding ages. The location of the kink changes substantially from age 64 to age 65; as Figure 4 shows, during this period the exempt amount is much higher for individuals NRA and above than for individuals below NRA. Individuals may have difficulty adjusting their earnings to the new, higher kink within one year. This suggests that individuals also face delays in adjusting in this context.

Similar patterns of adjustment occur when looking at the periods 1972-1982, 1983-1989 and 2000-2006 (Appendix Figures 18 to 20). We find evidence of adjustment delays, as individuals continue to bunch at the kink at ages older than the highest age to which the AET applies. However, in no case does adjustment appear to take more than three years.

Prior to the divergence of the exempt amount for those below and above NRA in 1978, we find no such dip in bunching at age 65. This "placebo" evidence further supports the hypothesis that the dip in bunching at age 65 arises from delayed adjustment to the increase in the exempt amount from ages 64 to 65 that emerges after 1978.

In our context, the only "appearance" of a new kink that we observe is the appearance of a kink at age 62. The amount of time since the appearance of the kink at age 62 is correlated with age, and elasticities and adjustment costs could also be correlated with age—thus confounding analysis of the time necessary to adjust to appearance of a kink. While recognizing these caveats, it is worth noting that the amount of bunching slowly rises from age 62 to 63, which suggests gradual adjustment. In principle, this could also relate to the fact that these graphs show the sample of those who have claimed by age 65, and the probability of claiming at a given age (conditional on claiming by age 65) rises from age 62 to 63. To address this issue, in Appendix Figure 17 we show the results when the sample at a given age consists of those who have claimed by that age, which still shows a substantial increase in excess bunching from age 62 to 63.

It is possible that a small amount of excess bunching occurs at ages 72 or older, but this is statistically insignificant.
2.5.2. Descriptive Evidence from Policy Changes Across Time

We next examine adjustment to a legislated change in AET policy. As shown in Figure 4, the AET was eliminated for those NRA and above in 2000. This policy change was unanticipated prior to the year 2000, as the legislation enacting the policy change was passed in April 2000 and applied to workers’ earnings in October 2000, and discussions prior to 2000 did not widely anticipate these changes.\footnote{The AET was also eliminated in 1983 for individuals aged 70 and 71. However, our results across ages show that individuals bunch at ages 70 and 71 in the 1990-99 period, so that persistent bunching at these ages in the 1983-89 cannot cleanly be interpreted as delayed adjustment to the 1983 change (as opposed to delayed reaction to the disappearance of the kink at age 70).}

Figure 7 shows the results for those aged 66-69. Bunching in the earnings distribution is easily visible in the years prior to 2000. In 2000, however, there is immediately little bunching visible, and this lack of bunching persists after 2000.\footnote{For comparison, Appendix Figure 21 shows that bunching stayed relatively constant for the 62-64 year-old group that experienced no policy change in 2000. While this group faces a notch at the exempt amount rather than a kink (as explained above), the relative comparison is instructive and suggests that the fall in bunching in the 66-69 year-old group in 2000 and subsequent years was due to the removal of the AET for this group in 2000.}

A very small bump in the earnings histogram is visible near the kink, but this proves to be insignificant in these data. Figure 8 also shows the amount of excess bunching estimated by year, along with 95 percent confidence intervals.\footnote{We have estimated this amount of excess bunching using three ways of calculating “placebo” kinks in 2000 and after: 1) by adjusting the exempt amount in 1999 using the CPI-U; 2) by adjusting the exempt amount in 1999 using the Employment Cost Index; 3) by using the exempt amount applicable to individuals in the year of attaining NRA in a given year (which is the same as the exempt amount that had been scheduled prior to the 2000 legislation to apply in each year to those NRA and above). Figures 7 and 8 show the first of these methods, but all of these methods show no significant bunching in these years (which is unsurprising given the lack of bunching visible in the histograms in 2000 and after).}

The amount of bunching is significantly greater than zero in all years prior to 2000, and estimates for 2000 and subsequent years show no significant bunching. Because the change was passed in April 2000 and implemented in October 2000—both after most salaried workers would have learned about their pay that year—the fairly fast reaction suggests that bunching is driven by workers with substantial flexibility in their earnings.\footnote{Due to changes that raised the scheduled exempt amount beginning in 1996, the AET had been scheduled to increase from $15,500 in 1999 to $30,000 by 2002. In principle, this could have affected the amount of bunching in 2000, even absent the elimination of the AET in this year. Nonetheless, bunching is unlikely to have been zero in the absence of the AET elimination, as the quarterly LEHD data discussed above show substantial evidence of bunching in quarterly earnings data prior to the fourth quarter of 2000, when the AET was eliminated.}
Appendix Figure 22 shows bunching in 1999, 2000, and 2001 in the LEHD. A spike at the kink is easily visible in 1999, and a small amount of bunching is visible in 2000 as the two bins on either side of the exempt amount are raised relative to the rest of the density (paralleling the small bump in the earnings histogram in 2000 in the MEF). In fact, excess normalized bunching proves to be significantly different from zero in 2000 in the LEHD ($p < 0.01$). By 2001, there is no clear visual evidence of bunching at the kink, and normalized excess bunching is insignificantly different from zero in the LEHD.\footnote{Since the sample size is much larger in the LEHD than in the MEF, it makes sense that we could estimate a small but statistically significant amount of bunching in the LEHD but not in the MEF in 2000. In principle, residual bunching in 2000 could also reflect individuals who earned money until their earnings reached the exempt amount (in a month prior to October). However, we also investigated the speed of adjustment from quarterly earnings data in the LEHD. (We do not primarily rely on these quarterly data because the AET is assessed yearly, and thus individuals can appear to bunch at the quarterly kink—defined as one-quarter of the earnings level associated with the kink in each year—even though their yearly earnings does not put them at the kink, or vice versa.) These data show a small but significant ($p < 0.01$) amount of bunching in each quarter of 2000 and in the first two quarters of 2001 but no significant bunching in subsequent quarters.}

In 2000, we find weaker evidence of a delay in adjustment—it only appears to occur among a small number of individuals, and it is only statistically significant in the LEHD. Moreover, bunching in the LEHD in 2000 is not necessarily immediately apparent in the earnings density and is therefore substantially less convincing than the residual bunching in the SSA data at ages 70-71. Thus, we do not wish to rely on the finding of residual bunching in the year 2000; instead, we consider this evidence to be merely suggestive of a small amount of residual bunching.

However, a number of facts are clear. First, in at least some contexts—\emph{i.e.} when aging out of the AET at age 70, apparently after the policy change in 1990 that we discuss later, and quite possibly after the policy change in 2000 (though to a smaller extent)—earnings adjustment frictions prevent some individuals from reacting immediately to the removal of a kink. Second, both when changes are anticipated (\emph{i.e.} the changes in policy across age) and unanticipated (\emph{i.e.} the policy change in 2000), adjustment occurs fairly rapidly, with the vast majority occurring within a maximum of three years. It is interesting to note that adjustment appears to be faster in the case in which the change is unanticipated than in the
anticipated case. While this may be surprising, many other differences between the two sets of changes—including differences in the degree to which the changes are publicized, the ages affected, the calendar year, and the distribution of individuals’ earnings—could be responsible for the discrepancy in the speed of adjustment. As we observe only a small number of changes in AET policy and confront several candidate explanations for heterogeneity in the speed of adjustment, we do not explicitly try to distinguish among these explanations.

2.5.3. Other Evidence Relating to Bunching

Figure 8 shows that there is no sharp change in the amount of bunching around the increases in the Delayed Retirement Credit in 1972 or 1982. We consider this suggestive—but not definitive—evidence of little discernable reaction to policy changes in benefit enhancement (particularly in light of our other results suggesting fast adjustment). A general downward trend in the amount of excess bunching is discernable in the 1990s—with the notable exception of a number of years, including 1995—which is coincident in the rise in the DRC through this period. However, we cannot conclusively attribute this potential trend to the influence of the DRC, as it could be due to other factors that changed over this period.\textsuperscript{35} We discuss adjustment to the decrease in the AET marginal tax rate from 50 percent to 33.33 percent in 1990 later.

We also conduct a variety of robustness tests. Appendix Figure 23 uses a bandwidth of $500 instead of $800, which changes our estimates little (as have other bandwidths we have chosen). In Appendix Figure 24, we vary the degree of the polynomial we use between 6 and 8, which shows similar results; other sufficiently rich polynomials we have tried have also shown similar results. In Appendix Figure 25, we vary the region near the kink we exclude when estimating the amount of excess bunching (from $2,000 to $3,000 to $4,000) and again estimate similar results. Limiting the sample to those who have substantial benefits (such

\textsuperscript{35}For example, the AET threshold amount rose much faster in the 1996-1999 period than in the previous period. It is possible that this helps to explain the decrease in the amount of bunching observed in these years, as individuals may find it difficult to adjust earnings to a rapidly-increasing kink. Meanwhile, as we discuss later, the fall in excess bunching after 1990 may relate to adjustment to the reduction in the benefit reduction rate in 1990.
as those with $1,000 or higher in benefits)—so that they are safely far from the concave kink in the budget set created when the AET reduces OASI benefits to zero—also yields very similar results.

Appendix Figure 26 shows that both men and for women bunch at the kink (though interestingly, men show more bunching than women). Previous work has demonstrated very different patterns of bunching among the self-employed and non-self-employed (Chetty et al., 2011) and has shown that bunching at the kink in response to the Earned Income Tax Credit is primarily driven by the self-employed (e.g. Chetty et al. (2013)). Appendix Figure 27 shows histograms for those with self-employment income in 1990-1999—who are excluded from our main sample—who also show an increase in the earnings density near the kink.

2.6. Mechanisms

This section probes the mechanisms that underlie patterns of adjustment to the AET, examining which parts of the earnings distribution adjust to AET changes, whether adjustment relates to age at death, and whether employers or employees drive responses to the AET.

2.6.1. Who Adjusts?

We investigate who adjusts to the AET using the large sample sizes in the LEHD data, which allow us to estimate parameters precisely in relatively small population groups. Specifically, we examine how earnings change as the AET is removed from age 69 to age 70 during 1990-1999, when the AET applied to individuals aged 62-69. As in Appendix Figure 16, Appendix Figure 28 shows the mean percentage change in earnings from age 69 to age 70 (y-axis), against earnings at age 69 (x-axis). The graph shows a large spike at the kink: individuals locating near the kink at age 69 on average increase their earnings substantially from age 69 to age 70.  

Recent literature has documented responses to kinks not captured by bunching,  

The increase near the kink is significantly higher than that in adjacent bins ($p < 0.01$). This spike in earnings growth is interesting in part because it directly documents responses to policy along the intensive margin, which is often found to be very inelastic (e.g. Eissa and Liebman (1996), Meyer and Rosenbaum (2001)). When we examine earnings growth in year $t + 1$ by earnings at year $t$, for ages $t$ younger than 69 we do not observe such a spike at the kink.
including Chetty et al. (2013) in the context of the EITC and Kline and Tartari (2013) in the context of the Connecticut Jobs First program. In the context of the AET, our evidence shows that responses to incentives appear to be concentrated among a group of bunchers at the kink, with little apparent response among others (though we cannot rule out that such changes occur in ways we do not capture, such as responses over a longer time frame).37

This finding suggests that individuals locating near the kink at age 69 are different than other individuals at the same age. Indeed, the AET applies not only to claimants locating at the kink, but also to claimants initially locating above the kink. Thus, if those initially locating at the kink had the same elasticity and adjustment cost as others, we might have expected to see a large increase in earnings in a substantial range of earnings above the kink, as well. The fact that we do not observe this pattern is suggestive of heterogeneity in adjustment costs or elasticities.38 For example, those initially locating at the kink may have low adjustment costs and react to the AET removal quickly, but those who do not bunch at the kink to begin with may have higher adjustment costs.39

In Appendix Figure 29, we show that individuals at the kink tend to follow the kink from year to year. We graph the probability of being at the kink in year $t + 1$, as a function of earnings in year $t$. There are clear spikes at the kink for ages 62-63 and ages 65-68, showing that individuals at the kink in year $t$ are disproportionately likely to be at the kink again the next year.40 We interpret this as further suggestive evidence that certain individuals are

37We further partially addressed the possibility of bunching over a different range by varying the bandwidth that we chose for estimating excess bunching.
38The income effect of the AET also rises with income, which would also lead the mean percentage earnings change to fall as income rises (under the assumption that leisure is a normal good). However, the income effect rises only gradually, whereas the mean percent earnings increase quickly falls just to the right of the exempt amount and remains relatively constant at this lower level as earnings rises—consistent with the hypothesis that those initially locating at the kink are more responsive to the AET. In fact, the data suggest that income effects (if any) are sufficiently small that they do not cause a noticeable systematic decrease in the mean percentage change in earnings as we move increasingly far to the right of the kink.
39The observed pattern is also consistent with such heterogeneity in elasticities or income effects.
40The probability of being near the kink in year $t + 1$ is significantly higher ($p < 0.01$) for those near the kink in year $t$ than in adjacent bins of the year $t$ earnings distribution. For those aged 58-60, who should not be affected by the incentives to bunch at the kink, no such spike occurs—demonstrating that the spike at the kink for ages subject to the AET is not simply an artefact of the natural evolution of the earnings distribution (absent the AET). We define the “placebo” kink for individual aged 58-60 as the kink affecting those aged 62-64.
particularly responsive to the incentives created by the kink, in the sense that they serially bunch at the kink.

To understand which part of the earnings distribution is affected by the AET, we examine more closely how the distribution of earnings differs across adjacent ages that face different AET incentives. Appendix Figure 30 stacks the distributions of earnings at ages 60, 61, and 62, as well as 69, 70, and 71. The earnings distribution changes modestly from year to year due to factors unrelated to the AET, as shown in the Figure from ages 60 to 61. However, the age-62 distribution shows a sharply different pattern than the age-60 or 61 distributions, with a sharp spike at the kink (particularly to the left of the kink), a higher density immediately to the right of the kink, and generally a lower density at earnings levels starting several thousand dollars above the kink.\textsuperscript{41} Similarly, the age-69 earnings distribution shows a sharply higher earnings density than the age 70 distribution in the immediate region of the kink (particularly to its left) but shows a lower density than age 70 at higher earnings levels, eventually reaching a similar earnings density starting around $6,000 above the kink.\textsuperscript{42} We return to this pattern of adjustment when discussing our model of fixed adjustment costs below.

Finally, it is possible that those with short expected lifespan could disproportionately bunch near the kink: the DRC should increase lifetime benefits more for claimants with longer life expectancy, which could lead the AET to be a larger effective tax on those with shorter lifespans (though as we note, the DRC only takes effect at earnings substantially above the exempt amount). In Appendix Figure 31, we show graphs illustrating that life expectancy is smooth near the kink (not significantly different from adjacent bins), suggesting no evidence for such a mechanism.

\textsuperscript{41}The age-61 distribution of earnings conditional on locating in the vicinity of the kink at 62, and the age-70 distribution of earnings conditional on locating in the vicinity of the kink at age 69, show similar patterns.

\textsuperscript{42}Some adjustment to the removal of the AET continues to occur after age 70; the evolution of the income distribution from, for example, age 69 to 72 shows similar patterns.
2.6.2. Employers and the AET

We use the LEHD data to investigate whether employers play a role in mediating responses to the AET. Chetty et al. (2011) argue that employers drive a significant share of the bunching at kink points observed in Denmark. In their context, some individuals bunch at kinks even though they are not directly subject to the policy that creates the kink. Chetty et al. conclude that these individuals bunch at the kink because employers create jobs that have those earnings levels. In other words, some individuals bunch at kinks because their employers present them a limited equilibrium menu of earnings levels (including the kink earnings level), and they would face costs of adjusting earnings to a different level.

We explore this possibility by testing for bunching among workers who are too young to claim OASI benefits and are therefore unaffected by the AET. Above, we have presented evidence indicating that in 1990-1999, individuals at ages earlier than those subject to the AET show little evidence of bunching at the AET kink. Thus, during this period, the evidence is consistent with the hypothesis that responses to the AET are driven by employees’ choices.\(^43\)

We extend this analysis by estimating bunching over the entire age distribution in the pre-1972 period, when the DRC did not exist, as Appendix Figure 32 shows. Appendix Figure 32 shows a small amount of statistically significant excess bunching at some ages younger than those subject to the AET (though not at other ages), suggesting that some employers do coordinate employment responses in this way in the pre-1972 period—though this behavior is small in the aggregate.\(^44\)

In Appendix Figure 33, we graph the probability that individuals change at least one em-

\(^43\)It is possible that employers drive some of the bunching at ages older than those subject to the AET, but we might then expect some degree of employer earnings coordination on the AET exempt amount for ages younger than 62 (and older than 70, including ages 72 and above).

\(^44\)Figure 32 bins three adjacent years of ages (e.g. 18 to 20); doing so in the post-1971 period shows no statistically significant bunching in pre-62 age bins. While individuals could in principle be locating at the kink prior to age 62 in anticipation of facing the AET later, it seems unlikely that they would do so as early as their late 30s, when a small amount of statistically significant bunching first appears—around 25 years before they are first eligible for OASI.
ployer from age $t$ to age $t+1$, against earnings at age $t$ (during 1990-1999 period). At ages when people face the AET, the probability that individuals change jobs across employers is sharply lower for individuals locating near the kink at age $t$ than for individuals with other initial earnings levels.\textsuperscript{45} The probability of changing employers is also sharply lower at the kink when individuals transition from being subject to the AET at age 69 to being no longer subject at age 70, and this is true when we limit the sample to those who increase their earnings from age 69 to age 70. Those initially locating at the kink evidently have sufficiently flexible pay arrangements that they can change their earnings from age 69 to age 70 while typically staying at the same employer.\textsuperscript{46}

2.7. Estimating Elasticities and Adjustment Costs

The results thus far suggest a role for adjustment frictions in individuals’ earnings choices in some contexts. As a first step in incorporating frictions into an estimable model of earnings supply, we build upon the Saez (2010) model (described briefly in the first sections of Appendix A.11), which uses bunching to identify the elasticity of (taxable) earnings with respect to the net-of-tax rate.\textsuperscript{47} We extend this model to allow for a cost of adjusting to tax changes.\textsuperscript{48} We first develop the theory graphically to show how adjustment costs affect bunching. Next, we show that using data on bunching at multiple kinks associated with different jumps in the net-of-tax rate, we can jointly estimate elasticities and adjustment costs. As discussed in Chetty et al. (2009), these parameters are jointly sufficient for welfare calculations in many applications.

\textsuperscript{45}The probability of changing employers near the kink is significantly lower than that in adjacent bins ($p < 0.01$).

\textsuperscript{46}Those locating at the kink might be different from other individuals for reasons—such as different demographics—that lead them to switch across employers less frequently. It is worth emphasizing that we attempt only to document the descriptive pattern that they change employers less frequently. We have also found that the graph of the probability of changing employers at age 59-61 against earnings at age 69 is smooth near the kink, suggesting that absent the AET incentives these individuals do not display noticeably different behavior in this regard.

\textsuperscript{47}Formally, the elasticity of earnings with respect to the net-of-tax rate is defined as $\varepsilon \equiv - (\partial z / z) / (\partial \tau / (1 - \tau))$.

\textsuperscript{48}Following recent literature on bunching— Saez (2010); Chetty et al. (2009, 2011, 2012, 2013); Chetty (2012); Kleven and Waseem (2013)— we specify a static model of earnings choice in each period. As we discuss further in the Conclusion, dynamic considerations represent an important topic for future research.
Our model relies on features of the empirical results that we have documented in the previous two sections. We find evidence of adjustment frictions, which we model through a cost of adjusting earnings. The empirical results also suggest that employees' choices are primarily responsible for patterns of bunching in the main period we study; this motivates a model in which employees choose their earnings rather than a model in which employers coordinate responses.

As described further in Appendix A.11, agents maximize utility $u(c, z; n)$ over consumption and earnings (where greater earnings are associated with greater disutility at the margin), subject to a budget constraint $c = (1 - \tau) z + R$, where $R$ is virtual income and the parameter $n$ reflects the tradeoff between consumption and earnings supply. We assume that in order to change earnings from an initial level, individuals must pay a fixed utility cost of $\phi^*$. This cost could represent the information costs associated with navigating a new tax regime if, for example, individuals only make the effort to understand their earnings incentives when the utility gains from doing so are sufficiently large (e.g. Simon (1955); Chetty et al. (2007); Hoopes et al. (2013)). Alternatively, this cost may represent frictions such as the cost of negotiating a new contract with an employer or the time and financial cost of job search, assuming that these costs do not depend on the size of the desired earnings change.

We model a fixed cost in order to build on recent literature that has focused on fixed costs (e.g. Chetty et al. (2011); Chetty (2012)). The distribution of earnings at ages 62 or 69 is higher in a region surrounding the kink but lower in a region substantially above the kink than at ages 61 or 70, respectively, which is consistent with a simple model with fixed adjustment costs that lead to a region of inaction and a region of adjustment. In Appendix

We describe the model in more detail in Appendix A.11. However, even with a fixed adjustment cost, the AET could in principle cause some individuals to reduce their earnings to levels just above the kink, which in principle could lead to a rise in the density to the right of the kink due to the imposition of the AET. Moreover, the shape of the distribution of earnings at age 70 conditional on locating at the kink at age 69 cannot be predicted a priori, as it should depend among other things on the correlation of the fixed cost of adjustment with the elasticity of earnings with respect to the net-of-tax rate. For example, if individuals with low fixed costs of adjustment tend to have low elasticities, then the conditional earnings distribution at age 70 should be closer to the kink on average than if individuals with low fixed costs of adjustment tend to have high elasticities. As a result of these factors, we cannot use the effect of the AET on such moments of the earnings distribution in estimating elasticities.
A.11.6, we extend our model to a case in which the cost of adjustment is linear in the size of the adjustment.

We develop two different approaches for estimating elasticities and adjustment costs. Our first approach, which we call the "Comparative Static" method, relies on comparing bunching in two separate cross-sections of data. Our second approach, which we call the "Sharp Change" method, additionally relies on attenuation (due to adjustment costs) in the change in bunching among individuals who face a change in the size of the kink over time. As we explain, the Sharp Change method relates more directly to our observation that bunching persists among individuals who formerly faced a kink. We begin by describing the Comparative Static approach because it introduces concepts that the Sharp Change method builds upon.

2.7.1. Estimation: Comparative Static Approach

The Comparative Static approach is best suited to estimating elasticities and adjustment costs from two cross-sections of different individuals who face different policies. Figure 9 Panel A illustrates how a fixed adjustment cost attenuates the level of bunching. Recall that our frictionless model predicts that bunchers have initial earnings (i.e. earnings in the absence of a kink) in the range \([z^*, z^* + \Delta z_1]\). Consider the person with initial earnings \(z_1\) (on the linear budget constraint with tax rate \(\tau_0\)). This individual faces a higher marginal tax rate \(\tau_1\) after the kink is introduced, which increases the marginal tax rate to \(\tau_1\) above earnings level \(z^*\). Because she faces an adjustment cost, she could decide to keep her earnings at \(z_1\) and locate at point 1. Alternatively, with a sufficiently low adjustment cost, she would like to pay the adjustment cost and reduce her earnings to the kink at \(z^*\) marked by point 2. We assume that the benefit of relocating to the kink is increasing in distance from the kink for initial earnings in the range \([z^*, z^* + \Delta z_1]\).\(^{51}\) These assumptions imply

---

\(^{51}\)In general, this requires that the size of the optimal adjustment in earnings increases in \(n\) at a rate faster than the decrease in the marginal utility of consumption. This is true, for example, if utility is quasilinear. We explore the implications of this assumption in Appendix A.11.5.
that above a threshold level of initial earnings, \( z_1 \), individuals adjust their earnings to the kink, and below this threshold individuals remain inert. We have drawn this individual as the marginal buncher who is indifferent between staying at the initial level of earnings \( z_1 \) (at point 1) and moving to the kink earnings level \( z^* \) (point 2) by paying the adjustment cost \( \phi^* \).

In Panel B, we show that the level of bunching is attenuated due to the adjustment cost: only individuals with initial earnings in the range \([z_1, z^* + \Delta z_1]\) bunch at the kink (areas \( ii, iii, iv, \) and \( v \))—whereas in the absence of an adjustment cost, individuals with initial earnings in the range \([z^*, z^* + \Delta z_1]\) bunch (areas \( i, ii, iii, iv, \) and \( v \)). The amount of bunching is equal to the integral of the initial earnings density over the range \([z_1, z^* + \Delta z_1]\):

\[
B_1(\tau_1, z^*; \varepsilon, \phi^*) = \int_{z_1}^{z^* + \Delta z_1} h(\zeta) d\zeta. \tag{2.1}
\]

Bunching therefore depends on the preference parameters \( \varepsilon \) and \( \phi^* \), the tax rates below and above the kink, \( \tau_1 = (\tau_0, \tau_1) \), and the exempt amount \( z^* \). The lower limit of the integral, \( z_1 \), is implicitly defined by the indifference condition drawn in Figure 9, Panel A:

\[
\phi^* = u((1 - \tau_1)z^* + R_1, z^*; \underline{\eta}) - u((1 - \tau_1)z_1 + R_1, z_1; \underline{\eta}) \tag{2.2}
\]

where \( R_1 \) is virtual income, and \( \underline{\eta} \) is the ”ability” level of this marginal buncher. If the marginal tax rate above \( z^* \) were instead \( \tau_2 \), where \( \tau_0 < \tau_2 < \tau_1 \), then bunchers would be comprised only of individuals with initial earnings in the range \([z_2, z^* + \Delta z_2]\) (area \( iii \)), which is again attenuated relative to bunching under a frictionless model (areas \( i, ii, \) and \( iii \)). This generates a second expression for bunching and an indifference condition analogous to 2.1 and 2.2, respectively.

When we later perform our estimates, we make use of a minimum distance estimator described in Appendix A.11.8 to solve this nonlinear system of equations. The key assumption underlying that method is that utility is quasi-linear and isoelastic, which is common in the
bunching literature (see Saez (2010); Chetty et al. (2011); Kleven et al. (2013); Kleven and Waseem (2013), for example). \(^{52}\) If we were to relax the assumption of quasilinearity, we would need to observe wealth, which is not available in the data.

**Intuition and Tractable Approximation**

To build intuition regarding our minimum distance estimation procedure, and to derive an expression relating the elasticity and adjustment cost to the level of bunching that can be easily solved in closed form, we can use a series of approximations to specify a simple system of linear equations. Let \(b \equiv B/h(z^*)\), *i.e.* the amount of bunching scaled by the density of earnings at \(z^*\) when there is no kink. Also assume that \(h(z)\) is uniform and equal to \(h(z^*)\) in the range between \(z\) and \([z^* + \Delta z]\). We show in Appendix A.11.6 that scaled bunching is approximately:

\[
b_1(\tau_1, z^*; \varepsilon, \phi) = \varepsilon \left( z^* \frac{d\tau_1}{1 - \tau_0} \right) - \phi \left( \frac{1}{d\tau_1} \right),
\]

(2.3)

where \(d\tau_1 = \tau_1 - \tau_0\) and \(\phi = \phi^*/u_c\) is the dollar equivalent of the adjustment cost. This equation shows intuitive comparative statics: All else equal, bunching is increasing in the elasticity, decreasing in the adjustment cost, and increasing in the size of the tax change at the kink. This generalizes and nests the formula developed in Saez (2010), which is equivalent in the case in which there is no adjustment cost. Because the amount of bunching is decreasing in the adjustment cost, constraining \(\phi = 0\) and using the Saez (2010) will in general weakly underestimate the elasticity in a single cross-section, since attenuation in bunching is attributed to a small elasticity rather than to the adjustment cost. Note that the expository derivation in (2.3) does not impose quasilinearity but uses the uniform density assumption and a first-order approximation for utility in the neighborhood of the kink.

Equation (2.3) also shows the features of the data that allow us to identify \(\varepsilon\) and \(\phi\). We need to observe bunching at two or more kinks, with variation in the change in tax rate \(d\tau_1\).

---

\(^{52}\) As explained in Appendix A.11.8, in a baseline we also assume that the density of initial earnings \(h(z)\) is uniform over the range \([z^*, z^* + \Delta z^*]\) (as in Chetty et al. (2011) or Kleven and Waseem (2013)), but we alternatively use a lognormal distribution of earnings based on those aged 61 (who are similarly aged but do not face the AET) and find similar results.
If we observe bunching at exactly two kinks of different sizes, then we can solve for $\varepsilon$ and $\phi$ exactly, as we then have a system of two equations and two variables. More generally, we could estimate a regression of $b$ on $z^* (d\tau_1 / (1 - \tau_0))$ and $-1/d\tau_1$, with the constant omitted. The coefficient on the first term is $\varepsilon$, and the coefficient on the second term is $\phi$.

Intuitively, with only a single cross-section of data, the amount of excess bunching increases in the elasticity and decreases in the adjustment cost, and thus it is not possible to identify both. Suppose that instead we have two cross-sections of data featuring different changes in marginal tax rates at the kink. The difference in the amount of bunching from one cross-section to the other will also depend on the elasticity and adjustment cost.\textsuperscript{53} Phrased differently, the Saez (2010) formula describes how bunching should vary between two different kinks in a frictionless model, and the extent to which observed bunching deviates from this pattern is attributed to the adjustment cost. Let $K_1$ and $K_2$ be two kinks that involve jumps at $z^*$ in the marginal tax rate of $d\tau_1 = \tau_1 - \tau_0$ and $d\tau_2 = \tau_2 - \tau_0$, respectively, and assume $d\tau_2 < d\tau_1$. Relative to the frictionless case represented by the Saez model, under the Comparative Static method, the change in bunching from $K_1$ to $K_2$ is larger. In the frictionless model, bunchers comprise areas $i$, $ii$, $iii$, $iv$, and $v$ in Figure 38 under $K_1$ and areas $i$, $ii$, $iii$ under $K_2$ (thus decreasing by areas $iv$ and $v$). Under the Comparative Static method, bunchers comprise areas $ii$, $iii$, $iv$, and $v$ under $K_1$ and area $iii$ under $K_2$ (thus decreasing by areas $ii$, $iv$, and $v$, rather than by only areas $iv$ and $v$ in the frictionless case).

**Heterogeneity in Elasticities and Fixed Costs of Adjustment**

Our empirical results suggest heterogeneity in the elasticity and the fixed cost of adjustment, as some individuals are more responsive to removal of the AET than others. Let $(\varepsilon_i, \phi_i, n_i)$ be jointly distributed according to a smooth CDF $G(\cdot)$, which translates into a smooth, joint distribution of elasticities, fixed costs and earnings $\tilde{H}(z, \varepsilon, \phi)$. As shown in Appendix A.11.6, assuming that the density of earnings, $\tilde{h}(z, \varepsilon, \phi)$, is again constant over the interval\textsuperscript{53}Under the approximations above, (2.3) implies that $\frac{\partial^2 b}{\partial \phi \partial (d\tau)} = \frac{1}{\tau^2} > 0$; as $\phi$ increases, the marginal impact of $d\tau$ on $b$ increases.
\[ [\bar{z}, z^* + \Delta z^*], \] we derive a generalized formula analogous to equation (2.3):

\[
b = \bar{\varepsilon} \left( z^* \frac{d\tau_1}{1 - \tau_0} \right) - \bar{\phi} \left( \frac{1}{d\tau_1} \right),
\]

where \( \bar{\varepsilon} \) and \( \bar{\phi} \) are the average elasticity and adjustment cost for those who bunch at the kink.\(^54\)

2.7.2. Estimation: Sharp Change Approach

The Sharp Change approach is best suited to estimating elasticities and adjustment costs when we are examining a constant population that experiences a change in the marginal tax rate at a kink (which may involve the kink disappearing), as we observe in our empirical applications. Suppose we observe a population that moves from facing a more pronounced \( K_1 \) to facing a less pronounced kink \( K_2 \) (as defined above). Adjustment costs prevent some individuals from "unbunching" from the kink, even though they would prefer to move away from the kink in the absence of an adjustment cost. The fixed adjustment cost therefore attenuates the change in bunching between two cross-sections in response to a reduction in the size of the kink, relative to the Comparative Static approach.\(^55\)

The first source of attenuation in the change is driven by individuals in area \( ii \) of Panel B. They bunch under \( K_1 \) and continue to bunch after transitioning to \( K_2 \). The reason is that the frictionless optimum under \( K_2 \) is \( z^* \) for everyone initially earning in the range \([z^*, z^* + \Delta z_2]\).

The second source of attenuation in the change is driven by individuals in area \( iv \) of Panel B. Panel C of Figure 9 demonstrates this. At point 0, we show an individual's initial earnings \( \bar{z}_0 \in [z^*, z^* + \Delta z_1] \) under a constant marginal tax rate of \( \tau_0 \). We now introduce the first kink, \( K_1 \). The individual responds by bunching at \( z^* \) at point 1. Next, we transition to

\(^{54}\)We are grateful to Henrik Kleven for suggesting the approach that led to this derivation.

\(^{55}\)If \( d\tau_2 > d\tau_1 \) instead – i.e. the kink becomes larger – then additional individuals will be induced to bunch, but the change in bunching will in general be attenuated in the Sharp Change approach relative to the Comparative Static approach (due to the adjustment cost). This is governed by an analogous set of formulas to the case \( d\tau_2 < d\tau_1 \) that we explore.
the muted kink $K_2$. Note that since $z_0 > z^* + \Delta z_2$, this individual would have chosen earnings $z_2 > z^*$ (marked as point 2) under $\tau_2$, if we had gone directly from no kink to $K_2$. However, in order to move to point 2, this individual must pay a fixed cost of $\phi^*$. We have drawn this individual as the marginal buncher who is indifferent between staying at $z^*$ and moving to $z_2$. All individuals with initial earnings in the range $[z^* + \Delta z_2, z_0]$ will remain at the kink.

Thus, bunching under $K_2$ is:

$$B_2(\tau_2, z^*; \varepsilon, \phi^*) = \int_{\tilde{z}_1}^{\tilde{z}_0} h(\zeta) d\zeta. \quad (2.4)$$

It follows that the absolute value of the change in bunching from $K_1$ to $K_2$ under the Sharp Change approach (area $v$ in Panel B) will be smaller than under the Comparative Static approach (areas $ii$, $iv$ and $v$). As discussed in Appendix A.11.7, $\varepsilon$ is still identified by the adjustment of the top-most buncher: $\varepsilon = \frac{z_0 - \tilde{z}_2}{\tilde{z}_2} \frac{(1 - \tau_0)}{d\tau_2}$. The critical earnings level $\tilde{z}_2$ is defined implicitly by the indifference condition in Panel C:

$$\phi^* \equiv u((1 - \tau_2)\tilde{z}_2 + R_2, \tilde{z}_2; \bar{n}) - u((1 - \tau_0)z^* + R_0, z^*; \bar{n}). \quad (2.5)$$

As before, our estimates use the minimum distance estimator described in Appendix A.11.8 to solve the system of nonlinear equations defined by (2.1), (2.4), and (2.5). Intuitively, we rely on a before-and-after comparison of bunching at the same kink, once the jump in marginal tax rates has been reduced. Inertia generates an excess amount of bunching in the period after the policy change. In the extreme case in which a kink has been eliminated, we can attribute any residual bunching to adjustment costs. The amount of residual bunching at the kink, in combination with the amount of bunching prior to the change in the jump in MTRs at the kink, therefore helps to identify both the elasticity and the adjustment cost.$^{57}$

$^{56}$Note that in general $z_2$ may be different from $\tilde{z}_2$.

$^{57}$An approximation explained in the Appendix also helps to build intuition.
Relative to the frictionless case represented by the Saez model, the change in bunching from the larger kink $K_1$ to the smaller kink $K_2$ is attenuated under the Sharp Change method by the adjustment cost (in contrast to the Comparative Static method). As noted above, in the Saez model, bunching decreases by areas $iv$ and $v$ in Figure 38 when moving from $K_1$ to $K_2$. Under the Sharp Change method, areas $ii$, $iii$, $iv$, and $v$ bunch under $K_1$, whereas areas $ii$, $iii$, and $iv$ bunch under $K_2$. Thus, bunching decreases only by area $v$ in the Sharp Change method, rather than both areas $iv$ and $v$ in the frictionless case. The absolute value of the decrease in bunching from $K_1$ to $K_2$ is decreasing in the adjustment cost—$\bar{\epsilon}_0$ is increasing in the adjustment cost, and therefore area $v$ is decreasing in the adjustment cost—helping to provide further intuition for our estimation procedure.

2.7.3. Estimates of Elasticity and Adjustment Cost

To estimate $\epsilon$ and $\phi$, we separately examine several changes in the AET benefit reduction rate, including a reduction in the rate in 1990; the elimination of the AET from ages 69 to 70; and the elimination of the AET in 2000. Our Sharp Change method is applicable in all of these contexts, as we observe a group from before to after these changes in policy. By contrast, the Comparative Static method is more applicable to analyzing changes in bunching when comparing two different groups each with positive (but different) marginal tax rates.

Estimating $\epsilon$ and $\phi$ requires estimates of the implicit marginal tax rate that individuals face. This requires estimates of both the "baseline" marginal tax rate, $\tau$—the rate that individuals near the AET threshold face in the absence of the AET due to federal and state taxes—and estimates of the implicit marginal tax rate associated with the AET. We begin by using a marginal tax rate that incorporates the effects of the AET BRR, as well as the average marginal income and FICA tax rates (including federal and state taxes).

---

58 As in the Comparative Static approach, the amount of bunching at $K_1$ is increasing in the elasticity (ceteris paribus) under the Sharp Change approach.

59 Using TAXSIM and the Statistics of Income individual tax return files, we calculated the mean of the
of benefit enhancement), but recall that Appendix A.7 shows that benefit enhancement is not relevant to an individual’s marginal incentives for earning an extra dollar near the AET exempt amount. This assumption is also consistent with the methodology of Friedberg (1998, 2000), who treats the AET as a pure tax. Moreover, as we have noted, the evidence shows little systematic bunching reaction to changes in the DRC. We vary these assumptions in various dimensions, which show similar results to the baseline: we exclude FICA taxes in the calculation of the baseline tax rate, and we alternatively assume that the benefit enhancement corresponds to a reduction in the effective marginal tax rate.

Before turning to our empirical estimates, we begin with graphical depictions of the patterns driving the estimates. Figure 8 shows excess bunching among 66-69 year-olds, for whom the BRR fell from 50 percent to 33.33 percent in 1990. Excess bunching fell slightly from 1989 to 1990 but fell more subsequent to 1990. For comparison, Appendix Figure 21 shows that bunching stayed relatively constant—both in 1990 and subsequently—for the 62-64 year-old group that experienced no policy change in 1990. While this group faces a notch at the exempt amount rather than a kink (as explained above), the relative comparison is instructive.

Appendix Figure 34 shows that the elasticity we estimate using the Saez (2010) method—constraining the adjustment cost to be zero—rises sharply from 1989 to 1990. This relates directly to our Sharp Change theory, which predicts that following a reduction in the change in the MTR at the kink, there may be excess bunching due to inertia (corresponding to area iv in Figure 9, Panel B). Once we allow for an adjustment cost, this excess bunching
is attributed to optimization frictions. Indeed, in the Appendix we explain that the rise (from just before to just after the policy change) in the elasticity that we estimate using the Saez (2010) method is a telltale sign that we face an adjustment cost as modeled by the Sharp Change method. In a context in which individuals have not yet had a chance to adjust (and the effective marginal tax rate has fallen), frictions may lead to larger elasticity estimates. Interestingly, this is in some sense the opposite of the usual presumption that adjustment frictions should lead to attenuation of elasticity estimates. Our finding also contrasts with the usual presumption that in the presence of adjustment frictions, smaller variation in taxes (i.e. smaller kinks) yield smaller elasticity estimates.

Table 11 presents our results from the Sharp Change method, examining the 1990 change. We estimate an elasticity of 0.23 in Column (1) and a positive adjustment cost of $152.08 in Column (2), both significantly different from zero ($p < 0.01$). This specification examines data in 1989 and 1990; thus, our estimated adjustment cost represents the cost of adjusting earnings in the first year after the policy change. When we constrain the adjustment cost to zero using 1990 data in Column (3), as most previous literature has implicitly done, we estimate a substantially larger elasticity of 0.39. Consistent with our discussion above, it makes sense that the estimated elasticity is higher when we do not allow for adjustment costs than when we do, as adjustment costs keep individuals bunching at the kink even though tax rates have fallen. The difference in the constrained and unconstrained estimates of the elasticity is substantial (69 percent higher in the constrained case) and statistically significant ($p < 0.01$).

We also consider alternative specifications. Using a lognormal earnings density rather than a uniform density (as described in Appendix A.11.8) changes the results little. Adjusting to disappear. Our focus in the Sharp Change estimates is limited to the period just after a policy change, before residual bunching has dissipated.

Specifically, we show that if we actually face the Sharp Change model but mis-specify the model as a frictionless (Saez, 2010) model, and we face a decrease in the jump in the marginal tax rate at the kink, the estimated Saez (2010) elasticity will weakly rise from just before to just after this policy change.

Friedberg (2000) finds uncompensated elasticity estimates of 0.22 and 0.32 in different samples. However, differences in the estimation strategies imply that these results are not directly comparable to ours.
the marginal tax rate to take account of benefit enhancement (applicable to those individuals to whom benefit enhancement is relevant to their earnings choices) raises the estimated elasticity but yields similar qualitative patterns across the constrained and unconstrained estimates. This makes sense: for the same behavioral response, if we assume a less pronounced percentage change in the net-of-tax rate, we infer a larger elasticity. The next rows show other specifications: excluding FICA taxes from the baseline tax rate; other bandwidths; and other years of analysis. Our results are similar under these and other variations.

In Appendix Table 37, we apply the Sharp Change method to the disappearance of the kink at age 70 (in which context we find residual bunching in Figure 5) and find similar (slightly higher) elasticity estimates and somewhat lower adjustment costs (though still in the same range). The constrained estimate of the elasticity is smaller than the unconstrained estimate; this makes sense, because we use data from age 69 to perform the constrained estimate, and adjustment costs attenuate the constrained estimate at those ages because they reduce bunching.

Appendix Tables 38, 39, and 40 present further specifications. In Appendix Table 38, we apply the Sharp Change approach to the 1990 policy change but assume that bunching in 1989 is not attenuated by adjustment frictions (under the rationale that bunching could have reached a ”steady state” in 1989 that is not attenuated by adjustment frictions). We estimate results similar to the baseline. In Appendix Table 39, we use the Sharp Change approach to estimate elasticities and adjustment costs using the disappearance of the kink in the year 2000 for those NRA and above. Given our small point estimate of residual bunching in 2000, it is unsurprising that we find small (though marginally significant)

---

67 The notion that individuals are initially in a steady state in which they have been able to make desired adjustments to the frictionless level of bunching is consistent with the observation that complete adjustment occurs within a few years of policy changes—prior to 1990 the AET policy parameters had last changed in 1983—and with the observation that they face frictions in un-bunching despite initially bunching. Thus, this estimate of the adjustment cost represents the cost of adjustment in the first year after a policy change, under the assumption that individuals are able to make adjustments within a few years of the previous policy change in 1983.
adjustment costs.\textsuperscript{68}

In Appendix Table 40, we apply the Comparative Static to the 1990 policy change. As noted above, the Comparative Static method is most applicable to performing estimates on unrelated cross-sections of individuals—rather than just before and just after a change in policy as in our results on residual bunching.\textsuperscript{69} We use years 1989 and 1993 to make sure that none of the 66-69 year-olds in 1993 were observed in 1989 (to avoid the possibility of residual bunching at the kink among those initially bunching there).\textsuperscript{70} We again find elasticity estimates in the same range, though with lower and typically insignificant adjustment costs. The linear approximation (2.3) shows higher (but still modest) adjustment costs, and higher elasticities than in the baseline specification (which makes sense because higher elasticities are needed to reconcile the higher estimated adjustment cost with the observed change in excess normalized bunching).

2.8. Conclusion

In the context of the Social Security Annual Earnings Test, we investigate the existence, nature and size of earnings adjustment frictions. We develop several related findings. First, we examine the speed of adjustment to the disappearance of convex kinks in the effective tax schedule. We document evidence of delays in adjustment in certain contexts, consistent with the existence of earnings adjustment frictions in the U.S. Nonetheless, we find that adjustment to both anticipated and unanticipated policy changes is quite rapid, as the vast majority of adjustment occurs within at most three years of budget set changes. This suggests that in this context, long-run elasticities are similar to those estimated in a medium-

\textsuperscript{68}We find smaller elasticities than in the baseline specification in part because normalized bunching was relatively small (though still positive and significant) in 1999, as shown in Figure 8.

\textsuperscript{69}The Comparative Static approach is inapplicable to years in which we observe residual bunching at a former kink; in assuming zero marginal tax rate after a kink disappears, the Comparative Static method effectively cannot explain this residual bunching. Thus, the Sharp Change method is more directly motivated by our primary empirical observations.

\textsuperscript{70}Because the Comparative Static approach examines separate cross-sections in 1989 and 1993 in our application, the interpretation of the estimated adjustment cost is different than in the Sharp Change approach (when it represented the cost of adjusting earnings within one year). Thus, there is no reason the estimated adjustment cost should be equal in the Comparative Static and Sharp Change applications we examine.
run time frame of a few years.

Second, we investigate mechanisms that underlie the patterns of adjustment. Adjustment to removal of kinks occurs primarily through substantial earnings growth among those initially locating at the kink, suggesting that they are more responsive than others (due to some combination of different elasticities or adjustment costs). We additionally investigate the extent to which firms may help coordinate bunching. The responses appear to be driven mainly by employees, as those under the minimum age subject to the AET do not bunch at the AET kink during the primary period we examine. Additionally, the bunchers are disproportionately likely to remain with the same employer while under the AET and when responding to the removal of the AET. This combination of evidence suggests that the bunching primarily results from the choices of certain particularly responsive employees who choose themselves to vary their earnings, generally within the same firm.

Third, we specify a model of employees’ earnings adjustment consistent with these findings and use it to estimate the earnings elasticity and the fixed adjustment cost. When we consider the change in bunching associated with the reduction in the AET benefit reduction rate from 50 percent to 33.33 percent for those above NRA from 1989 to 1990, we estimate that the elasticity is 0.23 and the adjustment cost is $152.08. The results are typically similar with other populations, time periods, and methods. When we constrain adjustment costs to zero in the baseline specification, the elasticity we estimate in 1990 (0.39) is substantially (69 percent) larger, demonstrating the potential importance of taking account of adjustment costs. Our estimates demonstrate the applicability of the methodology and the potential importance of allowing for adjustment costs when estimating elasticities. The modest adjustment cost we estimate parallels our empirical observation that bunching responds rapidly to changes in policy.

The analysis leaves open a number of avenues of further inquiry. First, it would further enrich the framework to extend the static analysis here to a dynamic model. We consider our static framework for estimating elasticities and adjustment costs to be a natural first
step in understanding estimation of these parameters (in the spirit of other static papers
such as Saez (2010); Chetty et al. (2009, 2011, 2012); Chetty (2012); Chetty et al. (2013);
Kleven and Waseem (2013)), but incorporating dynamic considerations is an important next
step. The speed at which individuals respond to changes in the AET, the nature of income
effects, and the distinction between anticipated and unanticipated changes—all of which we
have begun exploring in this paper—are three of several possible pieces of evidence that
could help in specifying a model of this sort (perhaps including a stochastic wage arrival
process, and incorporating the benefits over time to adjustment at a given time).

Second, further work distinguishing among the possible reasons for reaction to the AET
(such as misperceptions) remains an important issue. Third, further investigation of ex-
tensive margin and claiming responses to the AET would be valuable. Fourth, following
most previous literature, we have treated the adjustment cost as a ”black box,” without
modeling the process that underlies this cost, such as information acquisition or job search.
Future research could model such processes and distinguish these explanations using data.
Finally, the AET policy environment provides a useful illustration of many issues—such as
a methodology for estimating elasticities and adjustment costs simultaneously—that should
be applicable more broadly to studying adjustment to policy. As elasticities and adjust-
ment costs may be substantially different in other contexts, studying earnings adjustment
to other policies is a high priority.
Note: The right vertical axis measures the benefit reduction rate in OASI payments for every dollar earned beyond the exempt amount. The left vertical axis measures the real value of the exempt amount over time.
Figure 5: Histograms of Earnings, 59-73-year-olds Claiming OASI by Age 65, 1990-1999

Notes: The bin width is $800. The earnings level zero, shown by the vertical lines, denotes the kink. "Claimant" refers to an individual who has claimed by age 65.
Figure 6: Adjustment Across Ages: Normalized Excess Mass, 59-73-year-olds Claiming OASI by Age 65, 1990-1999

Note: The figure shows normalized excess bunching from a one percent random sample of SSA administrative data on Social Security claimants aged 59-73 between 1990 and 1999 (inclusive). Normalized excess bunching is calculated as described in the text. The vertical lines in Panel B show the ages at which the AET first applies (62) and the age at which the AET ceases to apply (70). See other notes to Figure 5.
Figure 7: Adjustment Across Years: Histograms of Earnings and Normalized Excess Mass, 66-69 year olds Claiming OASI by Age 65, 1996-2004

Note: The figure shows histograms of earnings from a one percent random sample of SSA administrative data on Social Security claimants aged 66-69 in each year from 1996 to 2004 (inclusive). In 2000 and after, the (placebo) kink is defined as the kink applying to those in the year of attaining NRA; as we mention in the text, the results are robust to two other ways of defining this placebo kink. See other notes to Figure 5.
Figure 8: Normalized Excess Bunching by Year, 1961-2005

Note: The figure shows normalized excess bunching from a one percent random sample of SSA administrative data on Social Security claimants aged 66-69 in each year between 1961 and 2005. See other notes to Figure 6.
Figure 9: Bunching Responses to a Convex Kink, with Fixed Adjustment Costs

Panel A

After-Tax Income $z - T(z)$

$U(z^*, t_1)$

$U(z_{1}, t_0)$

slope $= 1 - \tau_0$

$U(z_{1}, t_1) = U(z^*, t_1) - \phi^*$

slope $= 1 - \tau_1$

Panel B

Density

$z^*$ $\tilde{z}_1$ $\tilde{z}_2$ $z^* + \Delta z_2$ $\tilde{z}_0$ $z^* + \Delta z_1$

Panel C

After-Tax Income $z - T(z)$

$U(\tilde{z}_0, t_0)$

slope $= 1 - \tau_0$

$U(\tilde{z}_0, t_1)$

slope $= 1 - \tau_2$

$U(z^*, t_1) = U(\tilde{z}_0, t_2) - \phi^*$

slope $= 1 - \tau_1$

Note: See Section 2.7 for an explanation of the figures.
Table 10: Summary Statistics, Social Security Administration Master Earnings File

<table>
<thead>
<tr>
<th></th>
<th>(1) Ages 18-75</th>
<th>(2) Ages 62-69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Earnings</td>
<td>37,492.28</td>
<td>29,485.08</td>
</tr>
<tr>
<td></td>
<td>(282,940.03)</td>
<td>(783,897.87)</td>
</tr>
<tr>
<td>10th Percentile</td>
<td>5,234.74</td>
<td>1,957.78</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>15,291.11</td>
<td>7,291.21</td>
</tr>
<tr>
<td>50th Percentile</td>
<td>32,982.17</td>
<td>17,739.68</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>46,949.86</td>
<td>38,149.90</td>
</tr>
<tr>
<td>90th Percentile</td>
<td>66,370.18</td>
<td>58,417.99</td>
</tr>
<tr>
<td>Fraction with Positive Earnings</td>
<td>0.50</td>
<td>0.57</td>
</tr>
<tr>
<td>Fraction Male</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Observations</td>
<td>13,612,313</td>
<td>1,595,139</td>
</tr>
<tr>
<td>Individuals</td>
<td>619,580</td>
<td>545,615</td>
</tr>
</tbody>
</table>

Note: The data are taken from a one percent random sample of the SSA Master Earnings File and Master Beneficiary Record. The data for ages 18-75 cover those in 1961-2005 who claim by age 65, who do not report self-employment earnings, and who have positive earnings. (However, the fraction with positive earnings is calculated by including those who have zero earnings.) Column 2 covers the same sample but limits the ages to 62-69, the group we examine most often. Earnings are expressed in 2010 dollars. Numbers in parentheses are standard deviations. The standard deviations are large because of very rare, aberrant large values of earnings (as documented in Utendorf (2001/2)); these do not affect our estimates in the figures or tables because they are far above the AET exempt amount. These aberrant values affect mean earnings far less than they affect the standard deviation. The results are robust to winsorizing.
Table 11: Estimates of Elasticity and Adjustment Cost Using Sharp Change Method

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon$</td>
<td>$\phi$</td>
<td>$\varepsilon</td>
<td>\phi = 0$</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.23</td>
<td>$152.08$</td>
<td>0.39</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>[0.20, 0.27]</td>
<td>[49.54, 382.94]</td>
<td>[0.33, 0.48]</td>
<td>[0.18, 0.26]</td>
</tr>
<tr>
<td>Lognormal density</td>
<td>0.26</td>
<td>$165.76$</td>
<td>0.43</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>[0.22, 0.30]</td>
<td>[45.57, 473.52]</td>
<td>[0.35, 0.54]</td>
<td>[0.19, 0.29]</td>
</tr>
<tr>
<td>Benefit Enhancement</td>
<td>0.39</td>
<td>81.52</td>
<td>0.59</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>[0.33, 0.45]</td>
<td>[18.11, 245.88]</td>
<td>[0.49, 0.73]</td>
<td>[0.30, 0.43]</td>
</tr>
<tr>
<td>Excluding FICA</td>
<td>0.32</td>
<td>$129.14$</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>[0.27, 0.36]</td>
<td>[34.18, 360.97]</td>
<td>[0.42, 0.62]</td>
<td>[0.25, 0.35]</td>
</tr>
<tr>
<td>Bandwidth = $500$</td>
<td>0.25</td>
<td>$90.65$</td>
<td>0.38</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>[0.21, 0.30]</td>
<td>[6.04, 319.69]</td>
<td>[0.30, 0.50]</td>
<td>[0.19, 0.30]</td>
</tr>
<tr>
<td>1988 to 1990</td>
<td>0.32</td>
<td>$114.33$</td>
<td>0.50</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>[0.29, 0.35]</td>
<td>[26.49, 307.96]</td>
<td>[0.42, 0.62]</td>
<td>[0.27, 0.34]</td>
</tr>
</tbody>
</table>

Note: The table shows estimates of the elasticity and adjustment cost using the Sharp Change method described in the text. We report bootstrapped confidence intervals shown in parentheses. We investigate the 1990 reduction in the AET BRR from 50 percent to 33.33 percent. The baseline specification assumes a uniform density, calculates the effective MTR by including the effects of the AET BRR and federal and state income and FICA taxes, and uses data from 1989 and 1990. Alternative specifications deviate from the baseline as noted. The estimates that include benefit enhancement use effective marginal tax rates due to the AET based on the authors’ calculations relying on Coile and Gruber (2008) (assuming that individuals are considering earning just enough to trigger benefit enhancement). This translates the BRR before and after the 1990 policy change to 36% and 24%, respectively. Columns (1) and (2) report joint estimates with $\phi \geq 0$ imposed (consistent with theory, as described in the Appendix), while Columns (3) and (4) impose the restriction $\phi = 0$. The constrained estimate in Column (3) only uses data from 1990, whereas that in Column (4) uses only data from 1989 (except in the row where we investigate data from 1988 and 1990, in which case Column 4 uses data from 1988). *** indicates that the left endpoint of the 99 percent confidence interval is greater than zero; ** indicates that this is true for the 95 percent confidence interval; and * for the 90 percent confidence interval.
CHAPTER 3: Health Expenditure Risk and Annuitzation: Evidence from Medigap Coverage

How does health insurance affect annuitization? An emerging theoretical literature explores the impact of medical expenditure risk on annuitization, but there is little consensus on the quantitative importance of this channel. Using an instrumental variables strategy, this paper measures the impact of supplemental Medicare insurance—Medigap—on the annuity demand of the elderly.

An annuity is a financial product that provides a guaranteed stream of income for as long as the purchaser lives; it therefore provides insurance against living too long. Economic theory suggests strongly that elderly individuals should annuitize much of their wealth (Yaari (1965); Davidoff et al. (2005)). At least for individuals without bequest motives, annuities dominate bonds: Annuities provide a high payout while annuitants live, and the forgone payout after death is irrelevant.

Despite their theoretical attractiveness, people in fact annuitize little (e.g. Brown (2007)). A recent literature on health expenditure risk and annuity demand attempts to provide an explanation for this under-annuitization puzzle. Health expenditure risk can help resolve this puzzle, since health risk decreases the demand for annuities by driving up the value of liquidity. Feenberg and Skinner (1994) and French and Jones (2004) show that the elderly face substantial, fat-tailed, and highly autocorrelated medical expenditure risk, and Palumbo (1999) and De Nardi et al. (2010) has a large impact on precautionary savings among the elderly.

If health expenditures require large amounts of cash-on-hand, then annuities, with their slow and steady payouts, would be unattractive relative to more liquid assets. The quantitative importance of the health risk channel, however, is unclear. Sinclair and Smetters (2004), Turra and Mitchell (2008) and Peijnenburg et al. (2011b) all find that medical expenditure risk may decrease the demand for annuities. Pashchenko (2010) and Yogo (2011), however,
suggest that this channel is quantitatively unimportant, and Peijnenburg et al. (2011a) argue that full annuitization remains optimal even in the presence of medical expenditure risk. Pang and Warshawsky (2010) argue that medical expenditure risk actually increases annuity demand. Theory is therefore ambiguous about the importance of health expenditure risk for (under-) annuitization.

Much of what we know about the impact of health risk on annuitization comes from calibrated or estimated structural models. Solving these models requires making untested assumptions about the functional forms for utility and for the distribution of medical expenditure risk, and especially about the institutional structure of retirement financing. The impact of health expenditure risk on annuitization, in particular, appears to depend for example, on the level of the government-provided insurance floor, on the timing of annuity purchases, and on the link between the mortality credit in the annuity market and the equity premium. The theoretical results are sensitive to assumptions which prove difficult to verify, as evidence by the diverse conclusions in this literature.

In this paper, I take an alternative approach to measuring how medical expenditure risk affects annuitization: I directly estimate the impact of health insurance on annuity demand. If indeed medical expenditure risks loom large, then health insurance should complement annuitization. I focus on Medicare Supplemental Insurance, i.e. Medigap insurance, a regulated health insurance product that provides extra coverage beyond Medicare. In principle health insurance coverage can have two distinct effects on annuity demand. First, it may reduce the need for liquid assets, encouraging annuity demand. Second, if health insurance increases access to health care, then it may extend life and therefore also increase the value of an annuity. While Card et al. (2009) show that Medicare reduces mortality for heart attack victims, it is unlikely that the supplemental financial coverage provided by Medigap is life extending. My results are therefore most likely informative about the importance of liquidity risk for annuity demand.

Estimating the impact of Medigap on annuitization is difficult, however, because health
insurance is endogenous. The demand for insurance depends on wealth, risk preferences, and riskiness, and all of these factors also affect savings decisions independently. Indeed, Fang et al. (2008) show that considerable selection characterizes the Medigap market. To address this selection, I use an instrumental variables strategy, using the zip-code level price of Medigap as an instrument for coverage, conditioning on extensive controls for health, wealth, and preferences.

For this empirical strategy to succeed, I require that Medigap prices, conditional on other controls, be uncorrelated with underlying determinants of the demand for annuities. As I argue in Section 3.1, the institutional features of the Medigap market suggest that this condition may be satisfied. By law, Medigap insurance cannot be medically underwritten for individuals purchasing it at age 65. Prices in a given year thus may depend only on the buyer’s location, age, sex, and smoking status. While insurers might charge prices based on average health in an area, it is plausible to assume that, conditional on an individual’s own health and sex, the price of health insurance that she faces is uncorrelated with her underlying annuity demand. The variation in insurance prices underlying my estimates comes from variation in local spending habits that is plausibly unrelated to the individual’s demand for health, wealth, or preferences, and therefore unrelated to her annuity demand.

My empirical strategy requires data on both Medigap prices and annuity data. In Section 3.2, I discuss the data used in this study. I merge zip-code level data on Medigap prices with data from the Health and Retirement Study (HRS) on annuity holdings as well as health, wealth, and preferences. The HRS has excellent, detailed information on annuity income and on health, wealth, and even risk preferences and mortality expectations, making it an excellent setting for exploring annuity demand.

As I show in Section 3.3, there is a very strong and visually clear first stage relationship between Medigap prices and Medigap coverage. Across specifications that include increasingly detailed controls, I estimate a stable price elasticity of demand for Medigap coverage of about -1.4, with large F-statistics that allay weak instrument concerns. The first stage
is not robust, however, to controlling for state fixed effects, because these controls absorb almost all the variation in the instrument. An important weakness of this empirical strategy, therefore, is that I cannot control for unobserved state-level factors affecting Medigap prices and annuity demand. Nonetheless in Section 3.5, I show that the main results are robust to the inclusion of controls for state-level Medicare spending or Medicaid generosity, suggesting that neither overall medical spending patterns nor social insurance generosity drive my results.

As long as the variation in Medigap price is orthogonal to unobserved determinants of annuity demand, I can use Medigap price variation to estimate the impact of Medigap coverage on annuitization. I present two pieces of evidence verifying this identification assumption. First, I show that health insurance prices appear uncorrelated with individual demographics, health, or preferences, conditional on the pricing variables (age, year, sex, and smoking). Second, I implement a placebo test on retirees with employer provided health insurance, and find that their annuity demand is unrelated to the price of Medigap, suggesting that Medigap prices do not reflect unobserved determinants of annuity demand.

In Section 3.4, I examine the impact of Medigap coverage on three measures of annuity demand. First I show that the overall elasticity of annuity income—defined as income from any employer pension or any annuity—with respect to Medigap coverage is about 0.85. Because a large fraction of my sample has zero annuity income, I also examine the extensive margin response of annuity income to Medigap. I find an extensive margin elasticity of about 0.5, suggesting that more than half the response of annuity demand to health insurance occurs through the extensive margin. An alternative interpretation of these numbers is that the cross price elasticity of demand for annuity income with respect to the price of Medigap is about -1.3, with an extensive margin cross price elasticity of about 0.78.

While this measure of annuity income has the virtue of capturing all (non-Social Security) retirement income sources, including defined contribution pension or 401(k) plan balances
taken as annuities, it may be too broad. Some pensions and annuities provide a guaranteed payout for a limited number of years, but do not provide a mortality credit. To see whether “true” annuities respond to Medigap coverage, I look at whether retirees report any income from annuities that last until death and then stop. I find that these true annuities are particularly responsive to Medigap coverage, with an extensive margin elasticity of about 0.9, although this is less precise. These results are all robust to increasingly detailed controls for wealth, health, and preferences, and in Section 3.5 I show that several alternative sets of controls or sample definitions do not change the results.

The main contribution of this paper, therefore, is to provide direct empirical evidence on the relationship between annuity demand and medical expenditure risk. To my knowledge, this is the first paper to present empirical evidence on this question. It also contributes to the large literature on precautionary savings among the elderly. Palumbo (1999) and De Nardi et al. (2010) showed using structural models of retirement savings that medical expenditure risk has a large impact on the savings among older Americans. Goldman and Maestas (2012), using an empirical strategy similar to my own, showed that health expenditure risk increases savings among elderly Americans. Thus this paper provides further evidence on the importance of medical expenditure risk for the overall savings decisions of the elderly.

3.1. Background on Medigap insurance

Elderly Americans in the United States face several sources of medical expenditure risk. Medicare does not cover long term care, and this shortcoming represents a serious financial risk to the elderly (cf. Brown and Finkelstein (2007)). Moreover, the health insurance provided by Medicare is incomplete because of its cost-sharing provisions. Medicare Part A, which covers hospitalizations, requires patients to pay for the first day of hospitalization, and Part B, which covers outpatient services, has a 20 percent coinsurance rate. Medicare did not cover prescription drugs until the introduction of Part D in 2006, and the standard benefit in Part D today includes substantial cost sharing with a deductible and a donut hole with 100 percent coinsurance rate for expenditures above a certain amount.
Seniors may buy supplemental insurance to fill in the holes in Medicare’s coverage. These policies are called Medigap plans and they are tightly regulated by the federal government.¹ Insurers may offer policies from a menu of 10 standardized plans, denoted by letters A through J. CMS (2005) provides a detailed overview of the benefits of each plan letter. ² All plans cover Part A coinsurance and at least part of the Part B coinsurance, but they differ in their coverage otherwise. Because of this standardization, plans of a given letter are not differentiated in their risk protection. They also likely offer uniform service quality. While some health insurance plans might make limit the providers a beneficiary could use, or make it difficult to submit claims, Medigap coverage applies automatically to Medicare charges, suggesting no differentiation on this dimension. In this market, prices therefore likely do not reflect differential quality. ³

A key feature of the Medigap market is that if individuals purchase policies at the time they claim Medicare coverage, then insurers cannot use medical underwriting, nor can enrollment be denied or delayed. That is, prices cannot directly depend on health. Prices, however, can vary with sex, location, and over time.⁴ Many plans are also priced based on the buyer’s age of purchase, so that individuals have a strong incentive to purchase a policy at age 65, and the price they face at age 65 is the price they will face for the rest of their life (if they choose to renew their policies). These plans are called “age-issued.” Plans can also be “attained-age,” in which case prices vary with age, or “community rated,” in which case prices do not vary with age, but vary from year to year according to a local risk rating. I focus on attained-age prices, because these prices at age 65 are likely to have a strong and lasting influence on Medigap coverage throughout retirement.

¹As discussed by Fang et al. (2008), the 1990 Omnibus Budget Reconciliation Act, effective 1992, standardized Medigap plans and established the rules on medical underwriting described here. Individual states began mandating uniform coverage several years earlier, however, and Finkelstein (2004) explores the impact of these mandates.
²After my sample period, the menu of plans was expanded to 12.
³Starc (2012), however, argues that brand preferences and aggressive marketing are important features of this market. If so, then some price differences across areas may reflect differences in brand penetration.
⁴Plans are also allowed to charge different prices to smokers, although in practice this is rare. I do not use this price variation since in many zipcodes there are no plans that charge a different price to smokers. Nonetheless in all results I control for smoking status, so that difference in prices (and hence my results) are not driven by aggregate smoking tendencies.
Insurers setting these prices are prohibited from underwriting based on individual health or mortality expectations. The association between local prices at age 65 and long-term Medigap coverage, along with the fact that prices cannot depend on individual-specific health/wealth/mortality, motivates my instrumental variables identification strategy. Prices in these markets depend instead on the underlying area-level demand for Medical care (which I assume derive from risk preferences, health, and wealth), and on competitive conditions, i.e. differences in administrative costs or the number of active firms. Even conditional on an individual’s demand for medical care, there is substantial variation in prices, since insurance companies cannot perfectly risk adjust (or price discriminate). This variation is plausibly orthogonal to individual annuitization propensities, so it may be useful as an instrument.

3.2. Data: Weiss Ratings and the Health and Retirement Study

The data for this study come from two sources: the Health and Retirement Study conducted by the University of Michigan, and a proprietary insurance pricing database created by Weiss Ratings, Inc. The HRS provides information on demographics, preferences, wealth, annuitization, and insurance coverage. I merge the HRS data at the zip code level to the insurance pricing information.

3.2.1. Weiss Ratings Data

I use data from Weiss Ratings, Inc. on the price of Medigap policies. Weiss Ratings is a market research firm that collects Medigap pricing information from firms in most states. This dataset contains pricing information at the level of year-zip code-firm-plan letter, for 1998-2004. For every zip code it is possible to see every plan offered and the prices charged by each company. There is a fair amount of price dispersion within zip code, as Maestas et al. (2009) show, but very little dispersion in mean price across zip codes, within a state (and given plan letter). Although in principle, prices can differ by sex, in practice mean prices for men and women are extremely highly correlated within zip code ($\rho \approx 0.99$). Prices
are also highly autocorrelated over time. Most of the variation in prices, therefore, is within zipcode or across states, but not across zipcodes within state.

I take the price of Medigap insurance as the mean within-zip code price of an attained-age plan Plan A policy, allowing prices to differ by sex. There is little lost by using mean plan Plan A price rather than a different plan’s price, or some combination of plan prices, since the prices are highly correlated. For example, the correlation between plan Plan A prices and plan Plan F prices is 0.95. Since the age 65 price determines the path of future prices, I use the mean price of plan A plans in an individual’s zip code when she turns 65. In 2001, Plan A was the third most popular plan, with 11% market share, behind Plan F (37%) and Plan C (23%); no other plan had more than 10% market share (Kaiser Family Foundation, 2005). I use Plan A prices, however, because the first stage is slightly stronger for Plan A than for Plan F or Plan C. Using a combination of prices does not strengthen the first stage, because the prices are so highly correlated.

I exclude Massachusetts, Minnesota, and Wisconsin from the analysis because these states have their own set of Medigap plans, so the prices and policies are not comparable to those in other states. I exclude Connecticut, Maine, New Jersey, and New York because they lack age-issued policies in the Weiss data. I also lack data on Arizona and Vermont for 2002-2004, and on Washington for 1999 and 2002-2004, but when I do have data for those states, I include them in the analysis.

3.2.2. Health and Retirement Study

I merge the Weiss pricing data with data from the Health and Retirement Study on annuity holdings, demographics, wealth, health, and preferences. The Health and Retirement Study is a longitudinal survey following older Americans. It began in 1992, collecting detailed information about work, health, and wealth for a representative sample of the cohort born between 1931 and 1941; this is the original “HRS” cohort. In subsequent years, the study added additional cohorts born later and earlier; my analysis here focuses only on heads of
household from the original HRS cohort. The HRS offers a wider range of variables than other longitudinal datasets like the PSID or NLS, and these extra variables are essential to the present study. The HRS also is designed to cover older Americans, the group most likely to have annuity income. The HRS asks detailed questions about pension and annuity income. Pension and annuity income consists of all income from private annuities or employer provided pensions, including both defined-benefit and defined-contribution pensions. These variables are the main outcomes of interest. I also make use of the HRS’s extensive health and preference information as additional controls.

The analysis sample is restricted to heads of household age 65 or older with Medicare coverage and not covered by employer provided health insurance. This sample definition follows Fang et al. (2008), who study Medigap demand extensively; like them, I exclude observations with employer provided health insurance because I expect Medigap coverage to be superfluous for these people. Unlike Fang, Keane and Silverman, I include Medicaid recipients in the sample. While Medicaid coverage is more generous than Medigap, I include these observations because, for my sample, Medicaid coverage may reflect the realization of a severe health expenditure shock that drives assets to very low levels; I do not want to exclude from my sample individuals who realize this risk. Nonetheless in Section 3.5 I show that my results are robust to excluding these dual eligibles. I also exclude observations with imputed values for pension/annuity income, and people with missing values for any of the outcomes or controls. The final sample has 3,162 observations of 1,322 people.

Table 1 shows summary statistics on the key variables for the full sample and for the analysis sample. All dollar amounts are in 2010-constant dollars. I present summary statistics separately by whether the respondent has Medigap coverage, defined as having supplemental insurance and receiving Medicare (again following Fang et al. (2008)). Thirty percent of the sample has purchased Medigap policies. Relative to the uncovered, people with Medigap coverage generally face lower Medigap prices, as expected. Table 1 also shows large differences in annuity income by Medigap status: people with Medigap are more likely to
have annuity income and have more annuity income, on average; they are also more likely to have “true” annuity income, i.e. income from an annuity that lasts exactly until death. These differences do not necessarily reflect the causal impact of Medigap coverage, however. Medigap beneficiaries have fewer children and much higher permanent income than non-beneficiaries. Permanent income is defined as average income prior to age 65. (De Nardi et al. (2010) and Carroll and Samwick (1997) use a similar permanent income concept.)

Medigap beneficiaries are also in better health than non-beneficiaries. They are more likely to report having very good or excellent health, and they are less likely to have had diabetes or cancer. They have lower risk tolerance and expect to live longer. These differences reflect some of the advantageous selection discussed in Fang et al. (2008). Thus determining the impact of Medigap on annuitization requires addressing observed and, likely, unobserved confounding.

3.3. Empirical strategy and first stage results

3.3.1. Empirical strategy

To deal with this endogeneity, I use local price variation in Medigap policies as an instrument for Medigap coverage, conditioning on the individual-level characteristics that may affect savings decisions and correlate with local prices. To see why this strategy is necessary, and why it might be successful, let $y_{it}$ be the outcome of interest for individual $i$ at age $t$ and let $Medigap_{it}$ indicate Medigap coverage at $t$. I assume these variables are related as follows:

$$y_{it} = \alpha Medigap_{it} + X_{it}\beta + \varepsilon_{it}, \quad (3.1)$$

$$Medigap_{it} = \gamma \ln price_i + X_{it}\theta + \omega_{it}. \quad (3.2)$$

\(^{5}\text{Although in fact they report slightly lower probabilities of living to 85, relative to non-beneficiaries.}\)
The object of interest is $\alpha$, the impact of Medigap on annuitization. $X$ is a vector of control affecting both prices and annuity demand, including the pricing variables (age, sex, year, and smoking status) as well as other potential confounders. Annuitzation decisions, however, depend not only on health insurance coverage, however, but also on wealth, preferences (i.e. risk aversion and the intertemporal elasticity of substitution), and health and mortality expectations. But since individuals choose whether to purchase insurance, $Medigap_{it}$ depends on these variables as well, creating an endogeneity problem.

Using local Medigap prices as an instrumental variable potentially overcomes this problem, since identification comes from variation in insurance prices that is plausibly orthogonal to individual wealth, preferences, or health. That is, insurance prices are a function of aggregate (zip-code) level variables only. These variables likely include aggregate (determinants of) demand for medical care (and also medical care costs, administrative costs, and market structure in the Medigap industry). Because similar individuals—i.e. individuals with the same underlying demand for savings products—can live in zip codes with very different demands for medical care, there is a great deal of variation in the price of Medigap coverage, even conditional on annuitization propensities. Local prices can therefore isolate variation in Medigap coverage, independent of health, wealth, or preferences.

One concern in using cross zip-code variation in Medigap prices, however, is that these prices may be systematically correlated with zip-code level characteristics that directly affect annuity demand. For example, it may be that areas in which people have unusually high medical expenses have both high Medigap prices and high savings. In principle, people might decide to live in a given zip code because it has low Medigap prices. Here the control variables are useful: the threat to identification comes from the correlation between prices and individual health, wealth, and preferences. But using the rich HRS data, I can condition on these variables. To the extent that the results are robust to these controls, it is unlikely

---

6Medigap prices can affect annuity demand through income effects alone. However the typical Medigap plan costs about $1000, a small fraction of total income, suggesting that these income effects are unlikely to be important.
that additional unobserved determinants of savings or annuity demand are also correlated with local prices. The instrumental variables estimator therefore relies on variation in prices across people with similar characteristics to identify the impact of insurance on savings or annuitization.

In what follows, I show that, conditional on the variables that prices depend on, Medigap prices are uncorrelated with demographics, permanent health, risk tolerance, or mortality expectations. These results provide some reassurance of the validity of the instrument, since Medigap prices appear uncorrelated with many individual-level determinants of annuity demand. One remaining concern, though, is that high Medigap prices could reflect high administrative costs for insurance companies in that region, and therefore high prices might also signal high annuity prices (i.e. low returns); if so, this could bias estimates towards finding a positive impact of Medigap coverage on annuitization.

Three considerations cast doubt on this possibility. First, in their analysis of annuity premiums from a single large UK insurer in the 1980s and 1990s, Finkelstein and Poterba (2004) report that firms set geographically uniform prices, despite important variation in mortality across regions. This fact suggests that annuity premiums are not likely to be systematically higher in areas with high Medigap prices. It is not clear, however, whether the experience of that firm generalizes to the United States in the 1990s and 2000s. State-level data on annuity prices would let me address this concern directly, but I am unaware of any such data. As a second piece of evidence, I use premium data for several annuity providers collected in 1995 by A.M. Best (examined by Mitchell et al. (1999)). I match this pricing information to the provider's state (as listed by A.M. Best in its ratings) and correlate annuity premiums with state average Medigap premiums (in 1998, the earliest year for which I have data). This exercise reveals that the correlation between the price of a $10/month annuity and state Medigap price is virtually nil (-0.04).8 This procedure is of

7Finkelstein and Poterba (2004) write, “...like other firms in the market, [the firm we study] varies the annuity price on the basis of only age and gender, and not on the basis of the individual’s geographic location,” suggesting that this pricing practice is common at least in the UK.

8A.M. Best reports the monthly payout for a $1000 premium as offered by many providers. I convert this
course far from perfect, but it suggests there is very little geographic correlation between Medigap prices and annuity premiums.

As a final piece of evidence, in Table 18 below, I present evidence from placebo test that examines retirees with employer provided health insurance. Because these people do not buy Medigap, the price of Medigap should not be correlated with their annuity demand, unless it is correlated with some unobservable (such as annuity price) that also affects annuity demand. The placebo results show that people with employer provided health insurance do not respond to Medigap prices, providing further evidence that Medigap prices are not correlated with annuity prices.

3.3.2. First stage results

Table 13 presents the first stage estimates of equation (2). Each column presents the estimate \( \gamma \) obtained from using a different set of controls, as well as the F-statistic on \( \ln\text{price} \), which provides a test of instrument strength. Column (1) reports estimates of \( \gamma \) obtained with a basic set of controls. The basic controls are: year fixed effects, a quartic in age, and indicators for smoking status, sex, race, marital status, and educational attainment (high school, some college, and college or more), and dummy variables for having one, two, or three or more children. These controls serve two functions. First, year, age, smoking status, and sex are directly correlated with Medigap prices, since prices are set yearly and have increased over time. To the extent that there are differences in annuity demand by age, year, or sex, it is therefore important to control for these variables. Second, the demographic variables are closely related to lifetime earnings potential and annuity demand—because they may pick up differences in health, life expectancy, and preferences—and this may have a large impact on both savings and demand for medical care (through income effects). Controlling for these variables helps control for the correlation between local demand for medical care (which influences Medigap prices) and own demand for annuities.

payout into a price (i.e. premium) per $10/month). The correlation between monthly payout and Medigap price is 0.06.
The coefficient on ln price is -0.43 and, with a standard error of 0.07, it is precisely estimated and highly statistically significant. Here and throughout, all standard errors are clustered at the individual level since I pool multiple years of data for each individual. The F-statistic on ln price is 41.4, alleviating any concerns about a possible weak instrument. The sample mean of Medigap is 0.30, implying an average elasticity of about -1.4, somewhat larger than Starc’s (2012) estimate of an own price elasticity of demand for Medigap plans of about -1.1.

Figure 10 provides the graphical analog of the first stage relationship. The figure shows the non-parametric relationship between Medigap coverage and Medigap price, both net of controls. To construct the figure, I regress Medigap coverage and ln price (separately) on the age, year, smoking status, and the demographic controls, form the residuals, and then estimate a local linear regression of residual Medigap coverage on residual ln price; this is exactly the variation underlying the first stage regression. As the figure shows, there is a clear and negative relationship between Medigap coverage and its price. This relationship is strong and roughly linear across the full residual variation in ln price.

The basic controls adjust for standard demographic differences across people, but they may control inadequately for aspects of annuity demand that are also correlated with local prices. In column (2) I add controls for permanent income: a cubic in the log of permanent income, plus log permanent income interacted with age. One concern with the identification strategy is that, because health care and annuities are normal goods, high-income areas will have high annuity demand and high Medigap prices. Controlling for permanent income addresses this concern, and since the estimate of $\gamma$ hardly changes, indeed there is little evidence that own permanent income is closely related to Medigap demand or prices. Permanent income is perhaps an indirect measure of lifetime wealth, but controlling for liquid wealth directly, for example, would be inappropriate since liquid wealth is directly related affected by annuity purchases. Nonetheless, in the robustness section below, I also control for liquid wealth prior to retirement, and the results are unaffected.
The next two columns control for health and then preferences more directly. In column (3), I add controls for health: a set of indicators for health status (excellent, very good, good, fair, or poor), as well as indicators for ever having one of several chronic diseases or medical conditions (high blood pressure, diabetes, cancer, lung disease, heart disease, stroke, or arthritis). In column (4) I control explicitly for the HRS’s measures of risk preferences and mortality expectancy. The risk preference measures are the average value of risk tolerance implied by a person’s answer, over several survey rounds, to questions about large hypothetical gambles. (See Barsky et al. (1997).) Controlling for risk preferences can be important if area-level variation in risk preferences induces low Medigap prices while simultaneously inducing high annuitization. Fang et al. (2008) demonstrate that advantageous selection is an important feature of the Medigap market, hinting at this possibility. To control for mortality expectations, I take advantage of a pair of questions in the HRS that asks respondents the probability that they live until age 75 or until age 85. To correct for measurement and for the mechanical differences in mortality probabilities by age, I regress individuals’ answer to these questions on age and person fixed effects. I take the estimated person fixed effects as the measure of mortality expectations. As columns (3) and (4) show, the first stage results are virtually unchanged by introducing health or preference controls.

The first stage results are therefore highly robust to controlling for individual-level factors that might be correlated with insurance prices while directly affecting savings. The main results are similarly robust, as I show below. However, to the extent that permanent area-level differences in preferences or health may be driving differences in prices across areas, the most direct control is simply to include state (or zip code) fixed effects. As Maestas et al. (2009) show, however, most of the geographic variation in Medigap prices is between (rather than within) states; controlling for state fixed effects eliminates almost all the variation in prices. The results in column (5), which control for state fixed effects, shows this clearly: the coefficient on log price falls and its standard error increases. These state fixed effects eliminate the first stage relationship because there is no variation remaining to (precisely)
identify $\gamma$. Indeed, if I regress the instrument on the other controls, I obtain an $R^2$ of 0.91, but when I include state fixed effects, the $R^2$ rises to 0.97, implying that the state fixed effects absorb about two-thirds the remaining variation in the instrument. Since the first stage F-statistics are so low, state fixed effects cannot be used.

State fixed effects are important to the extent that state-level variables drive both Medigap prices and annuity demand. One such possible variable is state-level Medicaid generosity. In Table 19, I show that my results are robust to the inclusion of these policies. I also show that when I include average state-level Medicare spending, the first stage F-statistic falls considerably, although the main results are largely unaffected. This result confirms that much of the variation in Medigap prices is due to aggregate spending patterns, and that these patterns do not have a direct effect on annuity demand, or at least not a large one.

3.3.3. Assessing instrument validity

The first stage results are essentially unchanged when additional controls are included for permanent income, health status, and preferences. This robustness provides a first hint that the instrument may be valid: since these controls do not affect the first stage estimate of $\gamma$, either they are uncorrelated with Medigap coverage, or they are uncorrelated with $\ln price$. To the extent that observable determinants annuitization are uncorrelated with the instrument, it is possible that the unobservables are also uncorrelated.

Table 14 presents direct evidence that the instrument is uncorrelated with these additional controls. I construct the table by regression $\ln price$ increasingly detailed sets of controls. The table shows F-statistics of the test that the coefficients on the indicated set of controls is jointly equal to zero. Column (1) again has with only the basic controls. Year, age and sex (not shown in the table) are all highly correlated with the instrument. This is expected since prices depend directly on these variables. Smoking status (also not shown), however, is not correlated with $\ln price$ in any specification. The other basic controls which might correlate with insurance demand, health care utilization, or annuity demand are all uncorrelated with
prices, however. They remain uncorrelated as I add successively more controls. Column (2) adds controls for permanent income, a cubic in log permanent income and log permanent income interacted with age. These controls are jointly uncorrelated with the instrument. Column (3) adds controls for self-reported health status and for chronic diseases. Health appears uncorrelated with prices. Finally, column (4) controls for self-assessed mortality and risk tolerance. These preference variables are also uncorrelated with the instrument. The instrument is therefore largely uncorrelated with health, or preferences, except for those variables that directly affect pricing. These results may ease the concern that the instrument is systematically correlated with unobserved determinants of annuitization.

3.4. Results: Medigap coverage increases annuitization

Given the strength of the first stage, and the indication that the instrument may be valid, I now turn to estimating the impact of Medigap coverage on annuitization. I focus on three outcomes. First I study total income from employer pensions and annuities. This covers all permanent retirement income other than Social Security payments and therefore provides a useful summary measure of annuity income. Annuity income is highly skewed and, as Table 12 shows, many people have no annuity income at all. As a second outcome, I therefore focus on the probability of having any annuity or pension income. Finally, while the annuity variable usefully captures the total response of retirement income to Medigap coverage, it may be too broad. It includes pensions and annuities with limited payout horizons, for example, and these are not true life annuities. As a final outcome, I take advantage of the HRS's detailed annuity questions to determine whether an individual reports having any income from a “true” annuity, defined as an annuity that pays out until death and stops then. As Table 1 shows, these annuities are more rare; about three percent of the population has such income.

All these measures exclude income from Social Security. People can effectively buy an annuity by delaying Social Security claiming, which increases their future Social Security payments. In results not reported, I found some evidence that people do indeed earn more
Social Security income and claim later when they face lower Medigap prices, but these results were imprecise and not robust across specifications, so I omit them.

Table 15 shows the impact of Medigap coverage on total income from employer pensions or annuities. Panel A provides the reduced form result, where the coefficient on ln(price) is obtained by regressing annuity income on ln(price) and the indicated controls. The coefficient on ln(price) is a statistically significant -8.46 indicating that increasing the price of Medigap coverage by 10 percent would reduce annuity income by about $850. At the mean level of annuity income, this works out to a cross price elasticity of -1.31. Figure 11 Panel A, constructed analogously to Figure 10, illustrates the reduced form relationship between annuitization coverage and prices: it shows the nonparametric fit between (residual) annuitization probability and (residual) prices (net of controls). There is a clear negative relationship, although it is not as precise as the first stage relationship. This figure shows the exact variation underlying the reduced form results.

Panel B of Table 15 presents the instrumental variables estimate of $\alpha$, the impact of Medigap coverage on annuitization, estimated via two stage least squares with ln(price) as the excluded instrument. The coefficient is 19.6. To interpret this number, it is helpful to calculate the elasticity of (aggregate) total annuity income it implies. With mean Medigap coverage of 0.30 and mean annuity income of $6,510, the elasticity of total annuity income with respect to Medigap coverage is about 0.9.

Columns (2), (3), and (4) include successively more controls. The basic results changes very little with the inclusion of these extra controls. The point estimates fall slightly when controlling for permanent income, and it falls more when controlling for the health variables. Controlling for preferences, however, raises the point estimate almost back to the original. Medigap coverage appears to have a large and robust positive impact on annuitization.

These results reflect the total impact of Medigap on annuity income. Many people have zero annuity income, however, and much of the annuity puzzle is that many households do
not annuitize at all, not that households have relatively low annuity income conditional on having any annuity income. In Table 16, I investigate the extensive margin response, looking at the impact of Medigap on the probability of having any annuity income, estimated using linear probability models.\textsuperscript{9}

The reduced form results in Panel A of Table 16 indicate a strong and negative relationship between Medigap prices and the extensive margin of annuity demand, with a cross price extensive margin elasticity of about 0.78. Figure 11, Panel B shows the visual reduced form. Increasing Medigap prices by 10 percent decreases reduces the probability of having any annuity income by about 2.8 percentage points. The instrumental variables estimate of $\gamma$ in panel B gives the impact of Medigap coverage on the extensive margin of annuity demand. The point estimate of 0.65 implies an aggregate elasticity of about 0.53. Comparing this to the extensive margin elasticity implies that about two-thirds of the impact of Medigap coverage on annuitization comes from the extensive margin, and the remainder from the extensive margin. Adding successive controls in columns (2), (3), and (4) changes the point estimate only very slightly.

These results therefore show an important intensive and extensive margin response of overall annuity income to Medigap coverage. This response reflects the overall change in retirement income to health insurance coverage. To home in on the response from “true”, voluntarily purchased annuities, I examine a final outcome in Table 17: whether respondents report having income from an annuity that lasts until death and stops payment then, which I refer to as true annuity income. Panel A of Table 17 shows the reduced form impact of $\ln\text{price}$ on true annuity income. Figure 11, Panel C shows the visual reduced form. The point estimate of -0.04 is large relative to the main of 0.03, but it is imprecisely estimated. The cost of focussing on such a narrow measure of annuity income is that it likely introduces considerable measurement error, as people may miscategorize their retirement income while nonetheless recording the total amount correctly. While the point estimates remain roughly

\textsuperscript{9}The reduced form estimate of $\partial Pr(1|\text{Has annuity income})/\partial \ln\text{price}$ obtained from the linear probability models is very similar to the marginal effect of $\ln\text{price}$ implied by estimated probit models.
constant once I control for permanent income in column (2), it is no longer statistically significant at traditional levels. Adding additional controls for health in column (3) and preferences in column (4) has little effect on the point estimate, and raises the reduced form coefficient to statistical significance. Across all specifications, the extensive margin elasticity of true annuity income with respect to Medigap coverage is about 0.87.

The results show a clear impact of Medigap coverage on both the extensive and intensive margin of annuity demand. While the instrument is uncorrelated with observed determinants of annuity demand, it might nonetheless be correlated with unobserved factors, including the price of annuities. To test for this possibility, Table 18 presents evidence from a placebo test. This table shows the reduced form estimate of $\ln \text{price}$ on annuity demand, but the sample is modified: it contains only retirees with employer provided health insurance. If the instrument is valid, then for this sample, $\ln \text{price}$ should have no effect on medical expenditure risk, and hence the reduced form coefficient should be zero. If $\ln \text{price}$ is correlated with annuity prices or other unobserved determinants of annuity demand, however, then even retirees with employer provided health insurance will appear to respond to it. As the table shows, the coefficient on $\ln \text{price}$ is statistically insignificant, much smaller in magnitude than in the analysis sample, and, for two outcomes, wrong signed. Unfortunately the results are not precise enough for me to reject the hypothesis that the placebo point estimates are significantly different from the main point estimates. Nonetheless the results suggest that if $\ln \text{price}$ is correlated with unobserved determinants of annuity demand, the correlation is not strong enough to generate the observed relationship between Medigap coverage and annuity demand.

3.5. Robustness

3.5.1. Robustness to State-level confounders

The main threat to identification comes from the correlation between $\ln \text{price}$ and state-level unobservables. In this section, I show that the main results are robust to the inclusion of
two prominent state-level variables: Medicare spending, and the generosity of Medicaid’s implicit long term care insurance.

Table 19 shows the result of controlling for Medicare spending. I measure Medicare spending using 2004 data from the Dartmouth Atlas (Skinner et al. (2011)). In particular, the regressions in Table 19 control for average Medicare payments per enrollee, adjusting for sex and age. This is meant to control for the overall intensity of medical expenditures across states. To the extent that this spending variation is driven by aggregate differences in health across states, then it may be safely excluded from the demand for annuities. There are two reasons, however, that it might be desirable to control for this spending. First, if cross-state spending variation reflects differences in the price of obtaining medical care, then it may actually reflect medical expenditure risk. While it would be interesting to know how annuity demand responds to this risk, using it as an instrument for Medigap coverage would be inappropriate, since it has a direct effect on annuity demand. Second, it may be that this medical spending reflects individual as well as aggregate preferences, over and above the controls I have. In that case controlling for Medicare spending helps control for variation in preferences or health that may also affect annuity demand.

Medicare spending drives much of the variation in ln price, and so when I control for average Medicare spending, the first stage becomes weaker, with the F-statistic falling to 14.5 when I include only the basic controls, and 11.8 when I include the full controls. The results indicate that Medicare spending is positively associated with annuity demand, although the point estimates are generally imprecise, so including Medicare spending generally strengthens the results, although never by a great deal.

I also examine the impact of an alternative program on my estimates, Medicaid. Medicaid is particularly important because Medicaid covers long term care expenses, a major source of risk for the elderly (Brown and Finkelstein (2007)). Medicaid provides imperfect insurance, however. To qualify for Medicaid coverage, individuals must meet both an asset test and an income test. In 1998, the asset test varied between $1000 and $4000, and the income
test ranged between $500 and $1500 per month (Kasner and Shirey (2002)). Some states have “medically needy” rules which allow individuals to spend all income in excess of the allowed amount on their medical expenses in order to qualify.

In Table 20, I examine the robustness of my results to the inclusion of these Medicaid policy variables as they were effect in 1998. The data were collected by AARP (Kasner and Shirey (2002)), who contacted state Medicaid administrators to determine their program eligibility rules. In the table, I control for the log asset test allowed amount, the log income test allowed amount, an indicator for whether the state has “medically needy” rules, and an interaction between this indicator and log income test allowed amount. As the table shows, the main results are robust to the inclusion of the Medicaid policy variables. The table also hints at a surprising impact of Medicaid’s long term care insurance on annuity demand, although the point estimate are generally quite imprecise. In particular the table suggests that increasing the allowed asset amount reduces annuity demand. The table also suggests a very large impact of the income test on annuity demand, and this impact goes to zero for states with “medically needy” rules. These point estimates, however, do not necessarily reflect the causal effect of Medicaid parameters on annuity demand, because they may be correlated with other state policy parameters that also affect annuity demand. The impact of Medicaid rules on annuitization and, more generally, retirement savings is an interesting avenue for future work.

3.5.2. Further robustness

I conducted several additional robustness tests, and I present in the appendix four additional robustness tables. Table 33 shows the results of allowing for a more flexible functional form, a complete polynomial of degree two in risk tolerance, permanent income, and mortality expectations, fully interacted with an indicator for good or excellent health. Table 34 controls for average out-of-pocket medical spending and average liquid assets prior to age 65.

\footnote{Brown and Finkelstein (2008) use data from Kasner and Shirey (2002) to calibrate their model of demand for long term care insurance.}
While these variables may be endogenous—because individuals may anticipate having more
generous health insurance after age 65, or because they may begin to purchase annuities
prior to age 65—they may also soak up some of the remaining heterogeneity in risk or
preferences. Table 35 controls for spousal risk and mortality preference as well as the
respondent’s preferences, as household annuity decisions likely reflect some combination of
these variables. The table controls for spouse preferences linearly and fully interacted with
responded preferences. Finally Table 36 excludes Medicaid recipients from the analysis
sample. Across all robustness tables, the results change little.

3.6. Conclusion

This paper examines the impact of Medigap coverage on annuitization, using local prices
of Medigap coverage as an instrument for coverage. Medigap coverage has a large impact
on annuitization, with an aggregate elasticity of annuity income with respect to Medigap
coverage of about 0.9, and an extensive margin elasticity of about 0.5. The results also
suggest a large response of “true” annuity income, with an elasticity of about 0.87, but this
result is less precise.

These results are robust to extensive individual-level controls for income, health, and pref-
erences, and to controls for aggregate Medicare spending and Medicaid generosity. Overall
the results suggest an important role for health insurance in determining annuity demand.
Part of the underannuitization puzzle, therefore, may be due to uninsured medical expen-
diture risk. This conclusion is important in part because recent research has explored the
importance of behavioral factors such as framing biases in explaining underannuitization
(Brown (2007); Brown et al. (2008), Brown et al. (2011)). That line of research implies
that mandatory annuitization may, by correcting these biases, enhance welfare. To the ex-
tent that underannuitization is a rational response to imperfections in the health insurance
market, however, mandatory annuitization is unlikely to produce welfare gains.
Figure 10: Relationship between Medigap and ln(price)

Figure plots the nonparametric fit between log price of Medigap coverage and an indicator for Medigap coverage, both net of basic controls. Sample is defined in the text. The fit is obtained as a local linear regression of residual Medigap coverage on residual log price. These residuals are obtained from separate regressions of Medigap coverage or log price on year fixed effects, dummies for smoking status, sex, race, and educational attainment, and a quartic in age.
Figure 11: Relationship between endogenous variables and ln price

Notes: Figure plots the nonparametric fit between log price of Medigap coverage and the indicated outcome, both net of of basic controls. See notes to Figure 10.
Table 12: Summary statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1) All</th>
<th>(2) Medigap</th>
<th>(3) No Medigap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has Medigap</td>
<td>0.30</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>Price ($ 1000)</td>
<td>1022</td>
<td>968</td>
<td>1045</td>
</tr>
<tr>
<td></td>
<td>(276)</td>
<td>(261)</td>
<td>(279)</td>
</tr>
<tr>
<td>Has Annuity income</td>
<td>0.36</td>
<td>0.40</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>(0.48)</td>
<td>(0.49)</td>
<td>(0.48)</td>
</tr>
<tr>
<td>Annuity income ($ 1000)</td>
<td>6.48</td>
<td>7.89</td>
<td>5.87</td>
</tr>
<tr>
<td></td>
<td>(45.4)</td>
<td>(76.6)</td>
<td>(21.0)</td>
</tr>
<tr>
<td>Has true annuity</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.20)</td>
<td>(0.17)</td>
</tr>
<tr>
<td># Children</td>
<td>3.20</td>
<td>2.93</td>
<td>3.31</td>
</tr>
<tr>
<td></td>
<td>(2.11)</td>
<td>(1.81)</td>
<td>(2.21)</td>
</tr>
<tr>
<td>Permanent income ($ 1000)</td>
<td>68.1</td>
<td>86.1</td>
<td>60.5</td>
</tr>
<tr>
<td></td>
<td>(87.4)</td>
<td>(99.2)</td>
<td>(80.6)</td>
</tr>
<tr>
<td>High health</td>
<td>0.35</td>
<td>0.40</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>(0.48)</td>
<td>(0.49)</td>
<td>(0.47)</td>
</tr>
<tr>
<td>Ever had diabetes</td>
<td>0.22</td>
<td>0.20</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>(0.41)</td>
<td>(0.40)</td>
<td>(0.42)</td>
</tr>
<tr>
<td>Ever had cancer</td>
<td>0.16</td>
<td>0.14</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>(0.36)</td>
<td>(0.35)</td>
<td>(0.37)</td>
</tr>
<tr>
<td>Risk tolerance</td>
<td>0.39</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>(0.47)</td>
<td>(0.46)</td>
<td>(0.48)</td>
</tr>
<tr>
<td>Pr(live to 75)</td>
<td>63.71</td>
<td>65.14</td>
<td>63.10</td>
</tr>
<tr>
<td></td>
<td>(23.50)</td>
<td>(21.98)</td>
<td>(24.10)</td>
</tr>
<tr>
<td># Observations</td>
<td>3163</td>
<td>868</td>
<td>2295</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Table shows mean and, in parentheses, the standard deviation of select variables, for the analysis sample and by Medigap coverage status.
### Table 13: First stage results

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1) Basic</th>
<th>(2) Permanent inc.</th>
<th>(3) Health</th>
<th>(4) Preferences</th>
<th>(5) State FEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln price</td>
<td>−0.43***</td>
<td>−0.41***</td>
<td>−0.42***</td>
<td>−0.41***</td>
<td>−0.00</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>R²</td>
<td>0.151</td>
<td>0.176</td>
<td>0.187</td>
<td>0.191</td>
<td>0.279</td>
</tr>
<tr>
<td>F-statistic</td>
<td>41.4</td>
<td>38.8</td>
<td>40.3</td>
<td>38.2</td>
<td>0</td>
</tr>
<tr>
<td># Obs</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
<tr>
<td>Controls:</td>
<td>Basic X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Permanent inc. X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Health X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Preferences X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>State FEs X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is an indicator for whether the respondent has Medigap coverage. The sample is described in the main text. Basic controls are a quartic in age, dummy variables for smoking status, sex, race, marital status, and education dummies, and year fixed effects. Permanent income controls for a cubic in log average yearly income, plus log average yearly income interacted with age. Health controls include dummies for chronic conditions and self-reported health status. Preference controls include subjective measures of risk tolerance and mortality. Standard errors clustered on individual in parentheses. ** indicates \( p < 0.01 \).
Table 14: Correlation between instrument and regressors

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic</td>
<td>Permanent inc.</td>
<td>Health</td>
<td>Preferences</td>
</tr>
<tr>
<td>Children</td>
<td>1.25</td>
<td>1.32</td>
<td>1.19</td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.26)</td>
<td>(0.31)</td>
<td>(0.28)</td>
</tr>
<tr>
<td>Black</td>
<td>1.85</td>
<td>1.24</td>
<td>1.45</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.27)</td>
<td>(0.23)</td>
<td>(0.40)</td>
</tr>
<tr>
<td>Education</td>
<td>1.47</td>
<td>1.72</td>
<td>1.77</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.16)</td>
<td>(0.15)</td>
<td>(0.21)</td>
</tr>
<tr>
<td>Permanent income</td>
<td>1.22</td>
<td>1.38</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.30)</td>
<td>(0.24)</td>
<td>(0.21)</td>
<td></td>
</tr>
<tr>
<td>Health status</td>
<td>1.36</td>
<td>1.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic diseases</td>
<td>1.48</td>
<td>1.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality expectations</td>
<td>2.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk Tolerance</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.96)</td>
</tr>
<tr>
<td># Observations</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
</tbody>
</table>

Notes: Table shows the F-statistics on the indicated sets of variables, obtained from a regression of \( \ln price \) on the indicated variables, as well as controls for year, age, smoking, and sex. The sample is described in the main text. P-values are in parentheses and all standard errors are clustered on the individual level.
Table 15: Impact of Medigap coverage on annuity income

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permanent income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panel A: Reduced form</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ln price</strong></td>
<td>−8.46**</td>
<td>−8.04**</td>
<td>−7.35**</td>
<td>−7.93**</td>
</tr>
<tr>
<td></td>
<td>(3.81)</td>
<td>(3.79)</td>
<td>(3.37)</td>
<td>(3.79)</td>
</tr>
<tr>
<td>Panel B: IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has Medigap</td>
<td>19.61**</td>
<td>19.49**</td>
<td>17.54**</td>
<td>19.41**</td>
</tr>
<tr>
<td></td>
<td>(9.44)</td>
<td>(9.87)</td>
<td>(8.65)</td>
<td>(9.90)</td>
</tr>
<tr>
<td>Elasticity</td>
<td>0.90</td>
<td>0.89</td>
<td>0.80</td>
<td>0.89</td>
</tr>
<tr>
<td># Obs</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
<tr>
<td>Basic</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Permanent income</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Health</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Preference</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on total income from employer pensions or annuities (measured in thousands of dollars), and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. See notes to Table 13. Standard errors clustered on individual in parentheses. *, **, and *** indicate $p < 0.1$, $p < 0.05$, and $p < 0.01$. 
Table 16: Impact of Medigap coverage on extensive margin of annuity demand

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>Basic</td>
<td>Permanent inc.</td>
<td>Health</td>
<td>Preferences</td>
</tr>
<tr>
<td>Panel A: Reduced form</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln price</td>
<td>$-0.28^{***}$</td>
<td>$-0.27^{***}$</td>
<td>$-0.27^{***}$</td>
<td>$-0.26^{***}$</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Panel B: IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has Medigap</td>
<td>$0.65^{***}$</td>
<td>$0.66^{***}$</td>
<td>$0.64^{***}$</td>
<td>$0.65^{***}$</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.25)</td>
</tr>
<tr>
<td>Elasticity</td>
<td>0.53</td>
<td>0.55</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td># Obs</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
<tr>
<td>Controls:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Permanent income</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Health</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Preference</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the probability the respondent has any income from employer pensions or annuities, and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. See notes to Table 13. Elasticities are calculated at sample means. Standard errors clustered on individual in parentheses. *, **, and *** indicate $p < 0.1$, $p < 0.05$, and $p < 0.01$. 

131
Table 17: Impact of Medigap coverage on “true” annuity demand, extensive margin

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic</td>
<td>Permanent inc.</td>
<td>Health</td>
<td>Preferences</td>
</tr>
</tbody>
</table>

**Panel A: Reduced form**

<table>
<thead>
<tr>
<th>ln price</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−0.04*</td>
<td>−0.04</td>
<td>−0.04*</td>
<td>−0.04*</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

**Panel B: IV**

<table>
<thead>
<tr>
<th>Has Medigap</th>
<th>0.098*</th>
<th>0.094</th>
<th>0.096</th>
<th>0.101</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.060)</td>
<td>(0.059)</td>
<td>(0.061)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elasticity</th>
<th>0.868</th>
<th>0.830</th>
<th>0.854</th>
<th>0.894</th>
</tr>
</thead>
<tbody>
<tr>
<td># Obs</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
<tr>
<td>Controls:</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permanent income</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Preference</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the probability of having “true” annuity income, and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. The sample is described in the main text. See notes to Table 13. Standard errors clustered on individual in parentheses. *,**, and *** indicate \( p < 0.1, p < 0.05, \) and \( p < 0.01. \)
Table 18: Placebo test for impact on retirees with employer-sponsored health insurance

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1) Basic</th>
<th>(2) Permanent inc.</th>
<th>(3) Health</th>
<th>(4) Preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Outcome is annuity income (Mean: $15.9 thousand)</td>
<td>( \ln price )</td>
<td>1.29</td>
<td>4.02</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( (5.56) )</td>
<td>( (5.32) )</td>
<td>( (5.22) )</td>
</tr>
<tr>
<td>Panel B: Outcome is 1 {Has annuity income} (Mean: 0.64)</td>
<td>( \ln price )</td>
<td>0.02</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( (0.11) )</td>
<td>( (0.11) )</td>
<td>( (0.11) )</td>
</tr>
<tr>
<td>Panel C: Outcome is 1 {Has “true” annuity income} (Mean: 0.037)</td>
<td>( \ln price )</td>
<td>-0.01</td>
<td>-0.00</td>
<td>-0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( (0.05) )</td>
<td>( (0.05) )</td>
<td>( (0.05) )</td>
</tr>
</tbody>
</table>

Notes: The table shows the coefficient on \( \ln price \), obtained by regressing the indicated outcome on \( \ln price \) and the indicated controls. The sample consists of 1769 observations of 750 people with employer provided health insurance who otherwise meet the sample inclusion criteria described in the main text. See notes to Table 13. Standard errors clustered on individual in parentheses. *, **, and *** indicate \( p < 0.1, p < 0.05 \), and \( p < 0.01 \).
Table 19: Include state-level Medicare expenditures

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dep. var.:</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity income</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity income</td>
</tr>
<tr>
<td>ln (\text{price})</td>
<td>(-7.81^{***})</td>
<td>(-0.28^{***})</td>
<td>(-0.07^{***})</td>
<td>(-7.72^{***})</td>
<td>(-0.28^{***})</td>
<td>(-0.07^{***})</td>
</tr>
<tr>
<td></td>
<td>((2.59))</td>
<td>((0.09))</td>
<td>((0.02))</td>
<td>((2.90))</td>
<td>((0.09))</td>
<td>((0.02))</td>
</tr>
</tbody>
</table>

Panel B: IV

<table>
<thead>
<tr>
<th>Has Medigap</th>
<th>25.02^{**}</th>
<th>0.89^{**}</th>
<th>0.21^{**}</th>
<th>28.01^{**}</th>
<th>1.01^{**}</th>
<th>0.24^{**}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>((11.17))</td>
<td>((0.41))</td>
<td>((0.10))</td>
<td>((14.06))</td>
<td>((0.49))</td>
<td>((0.12))</td>
</tr>
<tr>
<td>Medicare spending (log)</td>
<td>6.83</td>
<td>0.30</td>
<td>0.14^{*}</td>
<td>10.23</td>
<td>0.43</td>
<td>0.16^{*}</td>
</tr>
<tr>
<td></td>
<td>((8.88))</td>
<td>((0.30))</td>
<td>((0.08))</td>
<td>((9.85))</td>
<td>((0.35))</td>
<td>((0.10))</td>
</tr>
<tr>
<td>F-stat</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>11.8</td>
<td>11.8</td>
<td>11.8</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the indicated outcomes, and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. ln \(\text{Medicare}\) is log of per beneficiary state Medicare spending in 2004, adjusted for age and sex of beneficiaries. The F-statistic is of the test that the first stage coefficient on ln \(\text{price}\) equals zero. The sample is described in the main text. See notes to Table 13. Standard errors clustered on individual in parentheses. *,**, and *** indicate \(p < 0.1, p < 0.05,\) and \(p < 0.01.\)
Table 20: Include Medicaid policy variables

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls:</td>
<td>Basic</td>
<td>Basic</td>
<td>Basic</td>
<td>Full controls</td>
<td>Full controls</td>
<td>Full controls</td>
</tr>
<tr>
<td>Dep. var.:</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panel A: Reduced form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln price</td>
<td>–7.82**</td>
<td>–0.22**</td>
<td>–0.05*</td>
<td>–7.42*</td>
<td>–0.21**</td>
<td>–0.05*</td>
</tr>
<tr>
<td></td>
<td>(3.61)</td>
<td>(0.09)</td>
<td>(0.03)</td>
<td>(3.84)</td>
<td>(0.09)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Panel B: IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has Medigap</td>
<td>16.19**</td>
<td>0.46**</td>
<td>0.10*</td>
<td>16.50*</td>
<td>0.46**</td>
<td>0.11*</td>
</tr>
<tr>
<td></td>
<td>(8.05)</td>
<td>(0.22)</td>
<td>(0.06)</td>
<td>(9.10)</td>
<td>(0.23)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Asset test (log)</td>
<td>–1.21</td>
<td>–0.12**</td>
<td>0.01</td>
<td>–0.59</td>
<td>–0.11**</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(1.25)</td>
<td>(0.05)</td>
<td>(0.01)</td>
<td>(1.35)</td>
<td>(0.05)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Income test (log)</td>
<td>15.65</td>
<td>0.35</td>
<td>0.08</td>
<td>17.31</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>(28.93)</td>
<td>(1.15)</td>
<td>(0.15)</td>
<td>(24.72)</td>
<td>(1.07)</td>
<td>(0.17)</td>
</tr>
<tr>
<td>Medically needy exemption</td>
<td>14.18</td>
<td>0.40</td>
<td>0.08</td>
<td>16.78</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(31.95)</td>
<td>(1.26)</td>
<td>(0.16)</td>
<td>(26.45)</td>
<td>(1.18)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>Need× income test</td>
<td>–15.41</td>
<td>–0.34</td>
<td>–0.09</td>
<td>–17.85</td>
<td>–0.14</td>
<td>–0.05</td>
</tr>
<tr>
<td></td>
<td>(29.05)</td>
<td>(1.15)</td>
<td>(0.15)</td>
<td>(25.04)</td>
<td>(1.08)</td>
<td>(0.17)</td>
</tr>
<tr>
<td># Obs</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
<tr>
<td>Controls:</td>
<td>Basic X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Permanent inc.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Health</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Preference</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the indicated outcome, and the instrumental variable estimate of Medigap coverage and Medicaid policy variables, with price as the excluded instrument. Standard errors clustered on individual in parentheses. *, **, and *** indicate \( p < 0.1, \) \( p < 0.05, \) and \( p < 0.01. \)
APPENDIX

A.1. Data appendix

A.1.1. Product definition

The claims data contain nearly 2000 unique anti-cholesterol drugs, defined by a National Drug Code (NDC) number. Different NDC numbers correspond to different molecules, different branded status, different generic manufacturers, different strengths, and different formulations (for example, tablet for capsule) as well as other aspects of differentiation (such as standard or extended release). To keep the analysis tractable, I define products at the molecule level, with some further aggregation of uncommon molecules.\(^1\) This procedure results in 11 products. I treat each of the six statins as distinct products. These are Lipitor (atrovastatin), Zocor (simvastatin), Mevacor (lovastatin), Pravachol (pravastatin), Crestor (rosuvastatin), and Lescol (fluvastatin).\(^2\) I aggregate the non-statin molecules by molecular class, so that Fibric acids, Niacin, and Bile Acid Resins (BARs) are distinct products. Last I create product categories for the two combination products available on the market, Advicor (niacin-simvastatin and niacin-lovastatin) and Vytorin (ezetimibe-simvastatin). Because ezetimibe is a unique molecule also available separately, I group it with Vytorin. Not all of these products are available in all years; Zetia/Vytorin and Niacin-Combination became available in 2002, and Crestor in 2003. Throughout the analysis, I limit the choice set to available drugs only.

This aggregation implicitly assumes that branded and non-branded versions of the same molecule are perfect substitutes.\(^3\) This assumption seems reasonable, at least for my sample

\(^1\) Since the product is essentially a molecule, I use the terms “drug” and “molecule” interchangeably throughout the paper.

\(^2\) A seventh statin, Baycol (cerivastatin) was available in the first half of 2001, but was withdrawn from the market after several patients died while taking it. I exclude from my sample any one who ever fills a prescription for Baycol.

\(^3\) In my sample, Fibrates, Niacin, and BARs are always off-patent. Lovastatin’s (Mevacor) patent expired in 2001, and simvastatin’s (Zocor) and pravastatin’s (Pravachol) in 2006, and all three drugs saw considerable generic entry. Atorvastatin (Lipitor), rosuvastatin (Crestor), ezetimibe (Zetia/Vytorin) and simvastatin/niacin (Advicor) were under patent protection throughout the sample. Lescol’s patent expired
period. Consider Zocor’s patent expiration. Zocor was under patent protection until June 2006, so its branded share was 100% prior to 2006. After patent expiration, the FDA grants the first generic entrant a period of six months of exclusivity, so there was limited intramolecular competition in 2006, and in that year Zocor’s branded fell to 64%. In 2007, the first year in which simvastatin was fully off-patent, Zocor’s branded share was 2%. At the same time, simvastatin’s market share, as a fraction of all cholesterol drug sales, rose from 13.7% in 2005 and 2006 to 18% in 2007. The near complete elimination of simvastatin’s branded market share, and the concomitant rise in simvastatin’s overall share, suggests that doctors and patients view branded and generic statins as essentially perfect substitutes. These patterns are also consistent with the overall prevalence of generic drugs: in 2009, 74.5% of all prescriptions dispensed were for generics, and among molecules that faced generic competition, the generic share was 93% (Berndt and Aitken, 2010). In my empirical analysis, I therefore do not differentiate between generic and branded versions of the same molecule. In effect, I assume that when a molecule faces generic competition, doctors’ choice sets include only the generic version, and patients only have a prescription for the generic version (if either).

A.1.2. Additional details on prices

Appendix Table A.5 shows one measure of imputation quality, by year, for formulary and non-formulary plans. Since prices are imputed as the average price per days supplied, they are mechanically correct on average but they understate true price dispersion. To measure the imputation quality, I therefore calculate the $R^2$ obtained from a regression of imputed prices on actual prices; an $R^2$ of one indicates that the imputed prices are perfect, whereas a low $R^2$ indicates that the imputation misses much of the price variation. As the table shows, the imputation for out-of-pocket price in formulary plans is good, with an $R^2$ above 0.75 in every year. In non-formulary plans quality is never this high. The imputation in general is worse for plan prices, because these prices in the raw data are not as tight, even during the sample period but generic entry did not begin until 2012.
for formulary plans.

As an example of the imputation, Appendix Table 21 shows imputed prices and modal prices, for two plans from 2003, and Figure A.5 shows the empirical price distribution for those plans. The Figure shows clear evidence that both plans have a three tier formulary, with a $10/$20/$40 structure for plan 1 and a $10/$20/$35 structure for plan 2. Each row of the table shows the imputed price and mode price for 30 days supplied, for each molecule. The imputed prices track the modal prices, suggesting that the imputation procedure is able to reconstruct formularies reasonably. Imputed prices differ from their modes slightly, primarily reflecting differences in days supplied.

A.1.3. More details on sample creation

Appendix Table A.5 shows the number of unique people in the data as I impose increasingly strong sample selection criteria, and some of their characteristics. The first row begins with all people for whom I ever observe enrolled with health insurance plan identifiers. (Many more people appear in the data without identifiers.) On average people were born in 1972.4, and the sample is 52.3% female. Over my sample period, 2001-2009, people had $15,700 in medical expenditures (the sum of drug spending, outpatient spending, and inpatient spending), of which $2,000 was out of pocket, $3,400 was spend on drugs, $8,500 on outpatient procedures (including office visits), and the remainder on inpatient procedure.

In the second row, I limit the sample to people who are ever at risk for heart disease. These are people who either ever have a diagnosis of a cholesterol-disorder, diabetes, heart disease or hypertension, or who are at risk because of their age (45 and older for men, 55 and older for women, but younger than 65 at some point in my sample). This reduces the sample by about three quarters. This sample, by construction, is sicker than average, and so medical spending is much higher.

---

4These plans were selected because they were moderately large, moderately generous, and from two different companies. Crestor is not in the table because it did not become available until August 2003, and neither plan had any claims for it.

5These prices are nominal prices for 30 days supplied, to keep the numbers round.
In the next several rows I exclude people based on their drug fills. First, I exclude people who ever fill a prescription for Baycol (generic name cerivastatin), a statin which was withdrawn from the market in 2001. Next I exclude people whose first fill occurs less than 6 months after entering the sample, whose first fill occurs before their first heart disease risk factor (including age risk), and whose first fill occurs after turning 64 (so that some fills occur at age 65). These restrictions are intended to select a sample of people who are new to taking cholesterol drugs, so that they do not change their insurance coverage to accommodate their drug needs.

Next, I exclude people whose first fill occurs outside the years 2001-2005 and 2008-2009. I only have advertising data for these years, and I impute advertising at the time of the first fill, so I cannot study people fills outside these ranges. In order to calculate compliance and switching, I require that people be continuously enrolled for 12 months from the first fill date (including that date); this exclusion cuts the sample by almost half.

Finally I exclude people who belong to plans that have no drug spending information, eliminating about 2% of people. These restrictions result in a sample of 945,459 people, including both people not in copay plans and people who never fill a prescription for an anti-cholesterol drugs. In comparison to the original sample, this sample is older and has much higher medical spending, although most of the difference comes from selecting people at risk of heart disease.

Essentially all of the analysis, except for two robustness tests, focuses on people who fill at least one prescription for an anti-cholesterol drug and who belong to copay plans. Just over a third of the sample ever fills a drug, and people who do are older, more likely to be male, and have even higher medical spending (in part by construction, in part because they are sicker). The copay plan sample is also older, sicker, and more likely to be male, although none of these differences is enormous.

---

6For people who never fill a drug, I take the “month of first fill” as the month of the first diagnosis for a heart disease risk factor. The people who never fill a drug are included only in the robustness test that includes an outside option in the initial prescription choice set.
A.1.4. Variable creation

The construction of the non-price variables used in the analysis is described here.

**Initial prescription, quit, comply, switch, and second prescription** I define each patient’s initial prescription as the first drug I see prescribed. I say that a patient complies with his initial prescription if he has at least 180 days supplied for that drug over the next 11 months, and has zero days supplied for other cholesterol drugs. I say that a patient switches his prescription if he fills at least 180 days supplied of any drug, including at least one fill of a drug other than his first. I say that patient quits if he fills fewer than 180 days supplied. I focus on 180 days because this is more than half the possible months filled in the first year (i.e. 6 of 11), but my results are robust to any other choice, as I show in Appendix Table A.5. If the patient switches, his second prescription is the drug he fills most, excluding the first.

**Plan generosity** Because average plan generosity is endogenous, in the sense that patients who are more risk averse or more likely to use health care may select more generous plans, I control for it in my analysis. Prescription drug insurance contracts are very high-dimensional objects so there is no ideal measure of plan generosity. I proxy for plan generosity using the average price paid in that plan and year for branded anti-cholesterol drugs.

**Health status** In some robustness tests, I include controls for health status. I measure health status using chronic heart disease indicators. In particular, from the full set of outpatient and inpatient claims data, I create a set of indicators for whether the patient has ever had a claim with a diagnosis for high cholesterol, diabetes, or heart disease, or hypertension. I also include sex and age among the health status variables, since they predict heart disease as well.

**Physician detailing and direct to consumer advertising** I use data collected by IMS
Health for 2001-2005 and 2008-2009 on total expenditures on physician detailing (excluding retail value of free samples) and total expenditures on direct to consumer advertising (DTCA). These variables are measured at the molecule-year level. Advertising expenditures are highly skewed; for many drug-years there is zero spending, but for the blockbusters, spending is over $100 million per year. In all the regressions, I normalize detailing and DTCA spending by subtracting the mean of of total advertising spending, across all drugs and years, and dividing by the standard deviation of total spending.7

A.2. Robustness appendix

In this Appendix, I present several additional robustness tests of the basic reduced form specifications. In Tables A.5 and A.5 I consider alternative samples. Column (1) includes the baseline sample. In column (2), I consider people in non-formulary plans. The basic pattern of results remains, and most coefficients are similar to the baseline sample. I focus on formulary plans in the main results because I want to ensure that out-of-pocket prices are accurately imputed, but in some plans, not all drugs have many claims. In columns (3)-(5) I apply increasingly stringent sample selection rules, requiring that each available drug have at least 10, 25, or 50 claims. The smaller samples eliminate some price variation and lead to larger standard errors, but otherwise have no important impact on the coefficients.

In Table A.5, I probe the sensitivity of the results to alternative definitions of compliance. The main results all define compliance as filling prescriptions for 180 days supplied (over the 330 days after the first 30). The table varies the definition in 30 day increments, from 90 days supplied to 330. Changing the definition of compliance has no impact on the basic result that compliance responds to out-of-pocket prices but not plan prices or detailing.

In Table A.5, I consider an alternative imputation procedure for out-of-pocket prices. Because I require that patients fill an initial prescription, I observe the actual price paid for the first prescription, so in the compliance regressions I can impute the price paid by the actual

---

7I normalize both variables by the same amount, the standard deviation of total advertising spending, so that the coefficients on them are comparable.
price faced. (A small number of people have missing prices despite filling a first prescription; I drop these people from the regressions using actual prices.) I cannot do the analogous imputation for the initial prescription because patients only receive a single drug, and so the actual price is missing for drugs not chosen. Column (1) shows the baseline results and column (2) shows the point estimates obtained by imputed prices with the actual price paid. The point estimates are extremely similar, suggesting that the imputation procedure does not introduce spurious price variation to drive the results.

In Table A.5, I expand the sample for the initial prescription estimates to include people who never fill a prescription for any anti-cholesterol drug. (I cannot look at compliance for these people, since they have no prescription to comply with.) I otherwise apply the same sample selection criteria for this group. I impute prices and advertising at the time of their first diagnosis with a heart disease risk factor, and require that they be enrolled for at least six months before this diagnosis, and continuously enrolled for at least 12 months after it. As the table shows, including the outside option changes the coefficients (and the coefficient on detailing is now only marginally significant, \( p = 0.06 \)), but detailing and plan price continue to have an important impact on drug choice.

A.3. Clinical evidence on anti-cholesterol drugs

Anti-cholesterol drugs are evaluated in clinical trials, with a focus on two kinds of outcomes: changes in cholesterol, and differences in mortality and major health events such as stroke and heart disease. As part of its approval process, the FDA reviews evidence from clinical trials on these outcomes. Once each drug is approved, the FDA also approves a label for each drug; these labels report the typical health gains from taking the drug. For cholesterol drugs, these labels always report the percent reduction in cholesterol caused by the drug, for each strength (e.g. 10/20/40 mg). However, as I argue below, the clinically relevant measure is the unit reduction in cholesterol. To obtain the unit reduction in cholesterol, I
assume that for each drug the baseline cholesterol level is 1.8.8 Appendix Table A.5 shows the cholesterol reduction caused by the 11 drugs in my sample. Where possible, I match each drug to its modal strength in my data.9

Cholesterol reduction, however, is not necessarily a meaningful health outcome. Several large clinical trials therefore studied the impact of statins on mortality, heart attack, and stroke. These studies find clear evidence that statins improve health. They are underpowered to study specific outcomes (e.g. stroke or heart attack separately) or subgroups (such as people with different conditions or baseline cholesterol levels), however. Baigent et al. (2005) provides a meta-analysis suitable for studying how statins affect specific outcomes and particular subgroups. Pooling data from 14 clinical trials, Baigent et al. find that statins reduce mortality by 12%, heart attack by 23%, and stroke by 17%. More importantly, they find that reductions in heart attack and stroke are roughly linear in the LDL cholesterol reduction induced by the drug: each mmol/L reduction in LDL cholesterol reduces the risk of major coronary events or major vascular events by 20%. Baigent et al. also find, moreover, that the impact of cholesterol reduction on health does not depend on baseline characteristics, including baseline cholesterol.10 Thus I use the absolute gain in cholesterol, across drugs, as a summary measure of the health benefit of taking that medication.

This approach results in a precisely measured, clinically meaningful health outcome that varies across drugs. A downside of this approach is that it is unlikely that Baigent et al.’s results generalize to non-statins. There are two concerns. First, it is unclear whether statins produce all their health benefits through LDL cholesterol reductions, or whether LDL cholesterol reduction is highly correlated with the many mechanisms through which

---

8Cholesterol is measured in molarity, i.e. its concentration in the blood. I report all units as mmol/L. Cholesterol is sometimes also reported as a density, in µg/mL; to convert between them, multiply µg/mL by 0.0245. The baseline cholesterol level is calibrated so that the percent reduction in LDL cholesterol reported on Atorvastatin’s label, 39%, corresponds to a 1.3 mmol/L decrease, as reported in the ASCOT trial (Sever et al., 2003).

9For some drugs and strengths, the labels do not report cholesterol reductions.

10But note that the outcome is percent reduction in heart attack or stroke risk; people with worse health presumably also have worse baseline heart attack or stroke risk.
statins affect health. Generalizing to a different drug class could overstate the gains from LDL reduction if part of statins’ impact occurs through a different channel. For example, the ENHANCE trial finds that ezetimibe reduces LDL cholesterol but fails to reduce thickness of arterial walls, making it unlikely that it would improve health (Kastelein et al., 2008). A second issue is that other non-statins are intended to target other forms of cholesterol. Niacin, for example, increases HDL cholesterol. Focusing on LDL cholesterol reduction can understate the gains from these drugs.

A natural alternative, therefore, would be to look at the clinical evidence for each drug, and focus simply on meaningful outcomes such as mortality. Unfortunately for many drugs, especially non-statins, this evidence is scant: it is much less expensive to study cholesterol than mortality, since cholesterol improvement can be detected in a few hundred people in as little as six weeks. Studying mortality requires thousands of people over several years. Second, the extant studies of clinically meaningful outcomes focus on diverse groups, making it hard to know if differences in treatment effects represent drug differences or population differences. Thus I focus on cholesterol changes because it is available for each drug, although given these concerns the results should be interpreted carefully.

A.4. Computational Notes

This section provides more detail about the numerical integration and the solver for the Maximum Likelihood Estimator.

A.4.1. Numerical integration

I use sparse grid integration with Gauss-Hermite nodes (Heiss and Winschel, 2008), using code provided by Heiss and Winschel. Sparse grid integration extends single dimensional quadrature based on polynomial rules to multidimensional integration problems. A polynomial rule of degree $d$ is a set of $N$ quadrature nodes $q_n$ and weights $\omega_n$ such that if $f$ is a polynomial of degree $d$ or less, then $\int f(x)dx = \sum_n \omega_n f(q_n)$; i.e. these rules are exact for degree $d$ and less. They are therefore highly accurate. One way to extend the polyno-
mial rules to higher dimensions is to use the tensor product, but this suffers from a curse of dimensionality, since the number of nodes required is $N^d$. Approximately, the tensor product extension of a 3-degree rule to two dimensions is inefficient, however, because it includes terms for $x_1^3x_2^3$ as well as $x_1^3$ and $x_2^3$, but the higher order interactions drop out of a Taylor series approximation and so are not needed. Sparse grid integration avoids the curse of dimensionality by extending polynomial rules to higher dimensions without using all the interaction terms.

Heiss and Winschel (2008), Skrainka and Judd (2011), and Skrainka (2012) present Monte Carlo evidence suggesting that sparse grid integration outperforms Monte Carlo or quasi-Monte Carlo (e.g. Halton draws) integration by orders of magnitude. That is, obtaining a given level of accuracy requires 10-100 times fewer function evaluations using SGI than using crude Monte Carlo integration.

Although sparse grid integration is very helpful in the numerical integration, it leads to two new problems. First, as Heiss and Winschel note, some of the weights for the sparse grid are negative, and so it is possible to calculate negative probabilities for some observations; as they argue, this reflects a poorly approximated integral, and in fact it also indicates a very low probability. If a parameter vector generates very low probabilities, it is unlikely to be the correct choice, so this is only a problem as the solver searches for an optimum. To deal with this problem, I set the probability for any observation with a negative value at 1/10 the lowest non-negative probability in the dataset, and I verify that at my solution, no observation has a negative probability. (This approach makes analytic derivatives inaccurate, so in practice I use numerical derivatives until obtaining convergence with a loose tolerance, and then use a tighter tolerance and analytic derivatives.)

A second problem arises because sparse grid integration samples accurately from the tails of distributions. This guarantees that some observations will have very high and very low match quality draws, and (especially at trial parameter values with high standard deviations), this leads to overflow in the evaluation of the logit choice probabilities. That
is, at some parameter guess and nodes, the patient’s utility from complying can be 1000 or more, and
\[ \frac{\exp(1000)}{X + \exp(1000)} \]
evaluates to not-a-number in double precision arithmetic, since \( \exp(1000) \) evaluates to infinity from overflow. I addressed this problem by converting the overflow to underflow (by normalizing by the utility of the highest-utility option). The cost of preventing overflow, however, is underflow; shifting utility this way leads some choices to have very negative utility, and \( \exp(-1000) \) evaluates to 0. Zero probability, however, is not a problem here. Quadrature nodes with extreme values have very small weight, so the zeros contribute very little to the average choice probability (and since the true probability is on the order of \( 10^{-20} \), the error is very small).

A.4.2. Maximization and numerical difficulties

I use the open source program Octave for all computational work. Octave is an open-source version of MATLAB with nearly identical syntax, but lacks some of MATLAB’s more recent features, and does not have any parallelization implemented. I use Octave’s \( \text{fminunc} \) command to maximize the likelihood \( L(\theta) \). The solver uses Newton’s method with line search, which means in each iteration \( t \) it uses Newton’s method to find a step direction, \( \vec{\theta}_t \), constructs a low-order approximation to \( L(\theta) \), and chooses a step size \( h \) to maximize \( L(\theta + h \vec{\theta}_t) \).

I provide the solver with an analytic gradient, which is very expensive to compute because it must be computed for each person and each SGI node. Using the gradient slows down each iteration considerably but appears to reduce the total number of iterations needed. The solver takes several days to converge to a solution; the exact time depends on the starting guess. Octave and its implementation of \( \text{fminunc} \) are not the gold standard of computational software. I used them because of legal limitations that I keep the data on the NBER’s servers. Student licenses of state-of-the-art software such as KNITRO (or even
MATLAB) are not available for use on a server.

The solver had difficulty finding an exact maximum of the likelihood. In all runs, it exited with a flag that indicated that, although the gradient of the objective function was positive, the function could not be decreased in the search direction, i.e. the step size $h$ is close to zero. This can occur if the low-order approximation of $L$ is not very high quality.

These exit flags can mean that the solver has failed even to come close to a maximum. To investigate this possibility further, I conducted several Monte Carlo studies. These studies were intended to explore the computational properties of my estimator when the data are generated according to the model, to see whether the computational problems are properties of the model or the result of misspecification. When I simulated data according to the model, and then estimated the model using MATLAB’s `fminunc` algorithm, I obtained point estimates close to the truth, but with a positive gradient and an exit flag indicating failure. When I estimated the model on the same data using KNITRO, I obtained nearly identical point estimates (off by about $10^{-5}$), but a gradient of zero.

The Monte Carlo performance depends on the distribution of covariates used. For example, more variation in the instruments (plan price and advertising) makes it easier to find a solution, since they help identification and lead to a sharper optimum. To get the distribution of covariates correct, I conducted a large scale Monte Carlo in which the distribution of all independent variables exactly matches their distribution in the estimation data. I performed 50 replications. In each replication, I simulated data using the point estimates reported in the text as the truth and as starting values. In each estimation I used Matlab’s `fminunc` with line search, and each took about four hours on my desktop (a Dell Optiplex 9010 with 16 GB of RAM).

The results from the Monte Carlo, shown in Appendix Tables A.5 and A.5, were quite encouraging. First, there is a small amount of bias, especially evident in the patient utility functions and the match quality terms for Zocor, Lipitor, and Pravachol. Simulated Max-
Table 21: Imputed and mode prices for two plans, 2003

<table>
<thead>
<tr>
<th>Price:</th>
<th>Plan 1</th>
<th>Plan 2</th>
<th>Plan 1</th>
<th>Plan 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OOP,</td>
<td>OOP,</td>
<td>Plan,</td>
<td>OOP,</td>
</tr>
<tr>
<td>imputed</td>
<td>mode</td>
<td>imputed</td>
<td>imputed</td>
<td>imputed</td>
</tr>
<tr>
<td>Lipitor</td>
<td>19.6</td>
<td>20.0</td>
<td>75.9</td>
<td>19.7</td>
</tr>
<tr>
<td>Zocor</td>
<td>19.7</td>
<td>20.0</td>
<td>92.0</td>
<td>19.5</td>
</tr>
<tr>
<td>Mevacor</td>
<td>9.8</td>
<td>10.0</td>
<td>78.8</td>
<td>23.0</td>
</tr>
<tr>
<td>Pravachol</td>
<td>19.5</td>
<td>20.0</td>
<td>102.8</td>
<td>19.7</td>
</tr>
<tr>
<td>Lescol</td>
<td>40.2</td>
<td>40.0</td>
<td>57.4</td>
<td>34.9</td>
</tr>
<tr>
<td>Ezetimibe Combo</td>
<td>40.0</td>
<td>40.0</td>
<td>68.1</td>
<td>34.9</td>
</tr>
<tr>
<td>Niacin Combo</td>
<td>40.0</td>
<td>40.0</td>
<td>44.9</td>
<td>19.8</td>
</tr>
<tr>
<td>Fibrates</td>
<td>13.7</td>
<td>10.0</td>
<td>37.6</td>
<td>16.5</td>
</tr>
<tr>
<td>Niacin</td>
<td>19.8</td>
<td>20.0</td>
<td>52.6</td>
<td>19.4</td>
</tr>
<tr>
<td>BARs</td>
<td>22.2</td>
<td>10.0</td>
<td>73.1</td>
<td>26.2</td>
</tr>
</tbody>
</table>

Formulary: 10/20/40 10/20/35

Table shows, for two plans, the (nominal) imputed out-of-pocket price and plan price for each molecule (defined as the average price per day supply, excluding claims with zero prices), and the (nominal) mode out of pocket price for 30 days supplied.

Histological Likelihood is known to be biased, so it is reassuring that all of the point estimates except the match quality terms for Zocor are close to the truth. The 90% confidence intervals (again clustered on plan) also have about the correct coverage rate. This matters for two reasons. First, it suggests that my inference is correct. Second, it rules out an alternative interpretation of the unbiased point estimates. It could be that the point estimates are unbiased because I use the truth at the starting value, and the solver just doesn’t move very much. But if this were true, the coverage rate would be too high. Since I cover only 90% of the time, I am more confident that starting values have not falsely reduced the bias. Finally, I note that in every replication the solver exited with a flag indicating that the step size was too small. Taken together, these Monte Carlo results suggest that, despite the worrisome exit flags, my estimates are close to the likelihood maximizing parameters.

A.5. Appendix tables and figures
Table 22: Predictive power of imputed prices

<table>
<thead>
<tr>
<th>Year:</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panel A: Formulary Plans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OOP price</td>
<td>0.775</td>
<td>0.837</td>
<td>0.857</td>
<td>0.823</td>
<td>0.759</td>
<td>0.788</td>
<td>0.819</td>
</tr>
<tr>
<td>Plan price</td>
<td>0.582</td>
<td>0.626</td>
<td>0.638</td>
<td>0.661</td>
<td>0.688</td>
<td>0.853</td>
<td>0.881</td>
</tr>
<tr>
<td># Claims</td>
<td>597,146</td>
<td>1,405,862</td>
<td>1,530,874</td>
<td>1,962,384</td>
<td>1,891,554</td>
<td>1,181,447</td>
<td>1,277,102</td>
</tr>
<tr>
<td>Panel B: Non-Formulary Plans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OOP price</td>
<td>0.317</td>
<td>0.577</td>
<td>0.728</td>
<td>0.511</td>
<td>0.545</td>
<td>0.737</td>
<td>0.74</td>
</tr>
<tr>
<td>Plan price</td>
<td>0.527</td>
<td>0.628</td>
<td>0.621</td>
<td>0.620</td>
<td>0.617</td>
<td>0.804</td>
<td>0.869</td>
</tr>
<tr>
<td># Claims</td>
<td>15,9576</td>
<td>98,807</td>
<td>116,847</td>
<td>193,551</td>
<td>328,233</td>
<td>768,561</td>
<td>845,182</td>
</tr>
</tbody>
</table>

Table shows the $R^2$ obtained from a regression of actual prices, as recorded on each claim, on the imputed prices, as defined in the text. The correlation is calculated for the indicated years, using the full set of claims, separately for formulary and non-formulary plans.
<table>
<thead>
<tr>
<th># People</th>
<th>Birth Year</th>
<th>% Female</th>
<th>All</th>
<th>Drug</th>
<th>Outpatient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>OOP</td>
<td>Total</td>
<td>OOP</td>
<td>Total</td>
</tr>
<tr>
<td>Full sample</td>
<td>15,765,277</td>
<td>1972</td>
<td>52</td>
<td>15.7</td>
<td>2.0</td>
</tr>
<tr>
<td>At risk for heart disease</td>
<td>4,846,140</td>
<td>1956</td>
<td>52</td>
<td>34.7</td>
<td>4.0</td>
</tr>
<tr>
<td>Drop Baycol</td>
<td>4,829,739</td>
<td>1956</td>
<td>52</td>
<td>34.6</td>
<td>8.3</td>
</tr>
<tr>
<td>First fill 6+ months after entering sample</td>
<td>4,077,343</td>
<td>1958</td>
<td>54</td>
<td>32.6</td>
<td>3.8</td>
</tr>
<tr>
<td>First fill occurs while at risk</td>
<td>3,995,770</td>
<td>1957</td>
<td>53</td>
<td>32.5</td>
<td>3.8</td>
</tr>
<tr>
<td>First fill occurs before age 65</td>
<td>3,968,298</td>
<td>1958</td>
<td>53</td>
<td>32.6</td>
<td>3.8</td>
</tr>
<tr>
<td>Continuously enrolled for 12 months</td>
<td>2,004,991</td>
<td>1957</td>
<td>53</td>
<td>38.1</td>
<td>4.4</td>
</tr>
<tr>
<td>Not missing plan information</td>
<td>965,265</td>
<td>1957</td>
<td>53</td>
<td>38.4</td>
<td>4.4</td>
</tr>
<tr>
<td>Plan is in drug data</td>
<td>945,549</td>
<td>1957</td>
<td>53</td>
<td>38.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Has drug</td>
<td>348,160</td>
<td>1951</td>
<td>47</td>
<td>55.1</td>
<td>6.1</td>
</tr>
<tr>
<td>Has formulary plan</td>
<td>296,760</td>
<td>1951</td>
<td>47</td>
<td>56.2</td>
<td>6.0</td>
</tr>
</tbody>
</table>

The table shows the sample size and average values of indicated variables as I impose increasingly stringent selection criteria. The continuous enrollment and non-missing plan information criteria must be met at the time of the initial fill (or initial risk factor, for people who never fill a prescription). The initial sample is the full Marketscan sample of people for whom I ever observe a non-missing value for insurance plan identifier.
Table 24: Patient summary statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics and health:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>51.9</td>
<td>7.9</td>
</tr>
<tr>
<td>Female (%)</td>
<td>47.4</td>
<td>49.9</td>
</tr>
<tr>
<td>Cholesterol (%)</td>
<td>68.0</td>
<td>46.7</td>
</tr>
<tr>
<td>Diabetes (%)</td>
<td>24.2</td>
<td>42.8</td>
</tr>
<tr>
<td>Heart disease (%)</td>
<td>22.2</td>
<td>41.6</td>
</tr>
<tr>
<td>Hypertension (%)</td>
<td>51.7</td>
<td>50.0</td>
</tr>
<tr>
<td><strong>Health spending and drug utilization:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol drug spending ($)</td>
<td>139.8</td>
<td>133.7</td>
</tr>
<tr>
<td>Other drug spending ($)</td>
<td>424.6</td>
<td>477.1</td>
</tr>
<tr>
<td>Health spending ($)</td>
<td>601.6</td>
<td>1318.3</td>
</tr>
<tr>
<td><strong>Insurance:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total claims (1000s)</td>
<td>92.8</td>
<td>78.3</td>
</tr>
<tr>
<td>% of claims among 4 modes</td>
<td>97.8</td>
<td>2.9</td>
</tr>
<tr>
<td>Average branded price ($) /30 DS</td>
<td>17.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Capitated (%)</td>
<td>24.7</td>
<td>43.2</td>
</tr>
<tr>
<td>Uncapitated, restricts provider choice (%)</td>
<td>61.8</td>
<td>48.6</td>
</tr>
<tr>
<td>Other plan (%)</td>
<td>13.4</td>
<td>34.1</td>
</tr>
<tr>
<td><strong>Prescribed drug:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out of pocket price ($) /DS</td>
<td>16.1</td>
<td>11.3</td>
</tr>
<tr>
<td>Statin</td>
<td>81.4</td>
<td>38.9</td>
</tr>
<tr>
<td>Comply (%)</td>
<td>51.5</td>
<td>50.0</td>
</tr>
<tr>
<td>Switch (%)</td>
<td>11.6</td>
<td>32.0</td>
</tr>
<tr>
<td># People</td>
<td>296,760</td>
<td></td>
</tr>
<tr>
<td># Plans</td>
<td>383</td>
<td></td>
</tr>
</tbody>
</table>

The sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans.
Table 25: Robustness of initial prescription estimates to alternative samples

<table>
<thead>
<tr>
<th>Sample:</th>
<th>Baseline</th>
<th>Non-formulary plans</th>
<th>10+ claims</th>
<th>25+ claims</th>
<th>50+ claims</th>
<th>100+ claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOP Price</td>
<td>$-1.382$</td>
<td>$-0.816$</td>
<td>$-1.568$</td>
<td>$-1.622$</td>
<td>$-1.600$</td>
<td>$-1.690$</td>
</tr>
<tr>
<td></td>
<td>(0.245)</td>
<td>(0.114)</td>
<td>(0.274)</td>
<td>(0.297)</td>
<td>(0.321)</td>
<td>(0.373)</td>
</tr>
<tr>
<td>Plan Price</td>
<td>$-0.590$</td>
<td>$-0.302$</td>
<td>$-0.516$</td>
<td>$-0.518$</td>
<td>$-0.518$</td>
<td>$-0.515$</td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td>(0.055)</td>
<td>(0.089)</td>
<td>(0.094)</td>
<td>(0.108)</td>
<td>(0.121)</td>
</tr>
<tr>
<td>DTCA</td>
<td>$0.359$</td>
<td>$0.089$</td>
<td>$0.290$</td>
<td>$0.292$</td>
<td>$0.268$</td>
<td>$0.173$</td>
</tr>
<tr>
<td></td>
<td>(0.128)</td>
<td>(0.075)</td>
<td>(0.164)</td>
<td>(0.175)</td>
<td>(0.201)</td>
<td>(0.243)</td>
</tr>
<tr>
<td>Detailing</td>
<td>$0.761$</td>
<td>$0.455$</td>
<td>$0.983$</td>
<td>$1.004$</td>
<td>$0.942$</td>
<td>$0.961$</td>
</tr>
<tr>
<td></td>
<td>(0.226)</td>
<td>(0.097)</td>
<td>(0.350)</td>
<td>(0.379)</td>
<td>(0.457)</td>
<td>(0.556)</td>
</tr>
</tbody>
</table>

N Plans | 383 | 273 | 177 | 128 | 87 | 69 |
N People | 296,760 | 51,400 | 202,721 | 195,705 | 171,050 | 144,187 |

Table shows the coefficients obtained from a multinomial logit regression of initial prescription against the indicated variables, as well as molecule fixed effects, and an indicator for imputed price. The sample always consists of people with a diagnosis of a cholesterol-related disorder and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription. Column (1) includes only people enrolled in formulary-based drug insurance plans, and column (2) instead limits the sample to people not in such plans. Columns (3)-(5) include only formulary plans in which each available drug has at least the indicated number of claims. Robust standard errors, clustered on plan, are in parentheses.
Table 26: Robustness of compliance estimates to alternative samples

<table>
<thead>
<tr>
<th>Sample:</th>
<th>Baseline</th>
<th>Non-formulary plans</th>
<th>10+ claims</th>
<th>25+ claims</th>
<th>50+ claims</th>
<th>100+ claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOP Price</td>
<td>−0.498</td>
<td>−0.581</td>
<td>−0.485</td>
<td>−0.470</td>
<td>−0.490</td>
<td>−0.764</td>
</tr>
<tr>
<td></td>
<td>(0.125)</td>
<td>(0.095)</td>
<td>(0.154)</td>
<td>(0.163)</td>
<td>(0.181)</td>
<td>(0.095)</td>
</tr>
<tr>
<td>Plan Price</td>
<td>−0.046</td>
<td>0.045</td>
<td>−0.005</td>
<td>−0.006</td>
<td>−0.058</td>
<td>−0.119</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.038)</td>
<td>(0.052)</td>
<td>(0.055)</td>
<td>(0.065)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>DTCA</td>
<td>0.047</td>
<td>0.253</td>
<td>0.082</td>
<td>0.094</td>
<td>0.011</td>
<td>−0.148</td>
</tr>
<tr>
<td></td>
<td>(0.099)</td>
<td>(0.104)</td>
<td>(0.127)</td>
<td>(0.135)</td>
<td>(0.156)</td>
<td>(0.102)</td>
</tr>
<tr>
<td>Detailing</td>
<td>0.057</td>
<td>−0.210</td>
<td>−0.053</td>
<td>−0.064</td>
<td>0.092</td>
<td>0.460</td>
</tr>
<tr>
<td></td>
<td>(0.171)</td>
<td>(0.153)</td>
<td>(0.260)</td>
<td>(0.280)</td>
<td>(0.351)</td>
<td>(0.208)</td>
</tr>
<tr>
<td># Plans</td>
<td>383</td>
<td>273</td>
<td>177</td>
<td>128</td>
<td>87</td>
<td>69</td>
</tr>
<tr>
<td># People</td>
<td>296,760</td>
<td>51,400</td>
<td>202,721</td>
<td>195,705</td>
<td>171,050</td>
<td>144,187</td>
</tr>
</tbody>
</table>

Table shows the coefficients obtained from a logit regression of compliance against the indicated variables, as well as controls for plan generosity (as measured by the average price per day supply of branded drugs), molecule fixed effects, an indicator for imputed price, and a capitation indicator. The sample always consists of people with a diagnosis of a cholesterol-related disorder and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription. Column (1) includes only people enrolled in formulary-based drug insurance plans, and column (2) instead limits the sample to people not in such plans. Columns (3)-(5) include only formulary plans in which each available drug has at least the indicated number of claims. Robust standard errors, clustered on plan, are in parentheses.
Table 27: Compliance regression: robustness to alternative definitions of compliance

<table>
<thead>
<tr>
<th># Of days supplied for compliance:</th>
<th>90</th>
<th>120</th>
<th>150</th>
<th>180</th>
<th>210</th>
<th>240</th>
<th>270</th>
<th>300</th>
<th>330</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOP Price</td>
<td>-0.56</td>
<td>-0.55</td>
<td>-0.53</td>
<td>-0.50</td>
<td>-0.49</td>
<td>-0.46</td>
<td>-0.39</td>
<td>-0.34</td>
<td>-0.27</td>
</tr>
<tr>
<td></td>
<td>(0.13</td>
<td>(0.13</td>
<td>(0.12</td>
<td>(0.13</td>
<td>(0.12</td>
<td>(0.12</td>
<td>(0.13</td>
<td>(0.13</td>
<td>(0.15</td>
</tr>
<tr>
<td>Plan Price</td>
<td>-0.08</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.04</td>
<td>(0.05</td>
</tr>
<tr>
<td>DTCA</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.00</td>
<td>0.01</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.11</td>
<td>(0.10</td>
<td>(0.10</td>
<td>(0.10</td>
<td>(0.10</td>
<td>(0.10</td>
<td>(0.11</td>
<td>(0.11</td>
<td></td>
</tr>
<tr>
<td>Detailing</td>
<td>0.12</td>
<td>0.02</td>
<td>0.06</td>
<td>0.09</td>
<td>0.07</td>
<td>0.10</td>
<td>0.09</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.18</td>
<td>(0.17</td>
<td>(0.17</td>
<td>(0.18</td>
<td>(0.18</td>
<td>(0.18</td>
<td>(0.21</td>
<td>(0.27</td>
<td></td>
</tr>
<tr>
<td>Mean DV</td>
<td>63.6</td>
<td>59.2</td>
<td>56.0</td>
<td>51.5</td>
<td>47.7</td>
<td>43.1</td>
<td>37.0</td>
<td>28.7</td>
<td>15.3</td>
</tr>
</tbody>
</table>

Table shows the coefficients obtained from a logit regression of compliance against the indicated variables, where compliance is defined as never switching and filling prescriptions for at least the indicated number of days supplied. All columns include the same set of additional controls: plan generosity (as measured by the average price per day supply of branded drugs), molecule fixed effects, an indicator for imputed price, and a capitation indicator. $\varepsilon_{\text{OOP}}$ and $\varepsilon_{\text{plan}}$ are the average elasticities of compliance probability with respect to out-of-pocket and plan price. The sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription, and enrolled in formulary-based drug insurance plans. Robust standard errors, clustered on plan, are in parentheses.
Table 28: Relationship between compliance and actual prices

<table>
<thead>
<tr>
<th>Price variable</th>
<th>(1) Imputed</th>
<th>(2) Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOP Price</td>
<td>−0.498</td>
<td>−0.517</td>
</tr>
<tr>
<td></td>
<td>(0.125)</td>
<td>(0.052)</td>
</tr>
<tr>
<td>Plan Price</td>
<td>−0.046</td>
<td>−0.008</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>DTCA</td>
<td>0.047</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>(0.099)</td>
<td>(0.085)</td>
</tr>
<tr>
<td>Detailing</td>
<td>−0.057</td>
<td>−0.011</td>
</tr>
<tr>
<td></td>
<td>(0.171)</td>
<td>(0.078)</td>
</tr>
</tbody>
</table>

# Plans          383  383
# People 296,760  294,123

Table shows the coefficients obtained from a logit regression of compliance against the indicated variables, as well as controls for plan generosity (as measured by the average price per day supply of branded drugs), molecule fixed effects, an indicator for imputed price, and a capitation indicator. In column (1), out-of-pocket price is imputed as described in the text, while in column (2) price is imputed as the actual price paid for the initial prescription.

Table 29: Robustness to including outside option

<table>
<thead>
<tr>
<th>Sample:</th>
<th>(1) Main</th>
<th>(2) Include Outside option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out of pocket price</td>
<td>−1.382</td>
<td>−1.025</td>
</tr>
<tr>
<td></td>
<td>(0.245)</td>
<td>(0.221)</td>
</tr>
<tr>
<td>Plan price</td>
<td>−0.590</td>
<td>−0.368</td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td>(0.073)</td>
</tr>
<tr>
<td>DTCA</td>
<td>0.359</td>
<td>0.538</td>
</tr>
<tr>
<td></td>
<td>(0.128)</td>
<td>(0.105)</td>
</tr>
<tr>
<td>Detailing</td>
<td>0.761</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td>(0.226)</td>
<td>(0.190)</td>
</tr>
</tbody>
</table>

\( \varepsilon_{OOP} \) −0.757    −0.572
\( \varepsilon_{plan} \) −1.189    −0.718
# Plans          383  403
# People 296,760  787,798

Table shows the coefficients obtained from a multinomial logit regression of initial prescription against the indicated variables, as well as molecule fixed effects, and an indicator for imputed price. The main sample consists of people with a diagnosis of cholesterol-related disorders and a prescription for a cholesterol drug, first prescribed at least six months after entering the sample, continuously enrolled for at least 12 months after their first prescription. The “include outside option” sample further includes people who never received a prescription for an anti-cholesterol drug but otherwise meet the inclusion criteria. Robust standard errors, clustered on plan, are in parentheses.
Table 30: LDL cholesterol reduction, by drug

<table>
<thead>
<tr>
<th>Generic name</th>
<th>Brand name</th>
<th>Modal strength</th>
<th>LDL cholesterol reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>atorvastatin</td>
<td>Lipitor</td>
<td>10 mg</td>
<td>1.30</td>
</tr>
<tr>
<td>simvastatin</td>
<td>Zocor</td>
<td>20 mg</td>
<td>1.23</td>
</tr>
<tr>
<td>lovastatin</td>
<td>Mevacor</td>
<td>40 mg</td>
<td>0.87</td>
</tr>
<tr>
<td>pravastatin</td>
<td>Pravachol</td>
<td>40 mg</td>
<td>0.92</td>
</tr>
<tr>
<td>rosuvastatin</td>
<td>Crestor</td>
<td>10 mg</td>
<td>1.53</td>
</tr>
<tr>
<td>fluvastatin</td>
<td>Lescol</td>
<td>80 mg</td>
<td>1.17</td>
</tr>
<tr>
<td>ezetimibe/combo</td>
<td>Zetia/Vytorin</td>
<td>10 mg</td>
<td>0.60</td>
</tr>
<tr>
<td>niacin and lovastatin</td>
<td>Advicor, others</td>
<td>500 mg - 20 mg</td>
<td>1.00</td>
</tr>
<tr>
<td>fibrates</td>
<td>Many</td>
<td>145 mg</td>
<td>0.61</td>
</tr>
<tr>
<td>niacin</td>
<td>Many</td>
<td>500 mg</td>
<td>0.07</td>
</tr>
<tr>
<td>BARs</td>
<td>Many</td>
<td>625 mg</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Modal strength is the modal strength, by days supplied, in my data. LDL cholesterol reduction is reported on FDA labels; see drugs FDA. See Appendix A.3 for more detail.

Table 31: Results from the Monte Carlo Experiment: Utility Function Estimates

<table>
<thead>
<tr>
<th>Parameter:</th>
<th>Truth</th>
<th>$E[\beta]$</th>
<th>$SD[\beta]$</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient utility function:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OOP Price</td>
<td>1.130</td>
<td>0.919</td>
<td>0.088</td>
<td>0.30</td>
</tr>
<tr>
<td>OOP Price $\times$ capitated</td>
<td>-1.258</td>
<td>-0.911</td>
<td>0.100</td>
<td>0.14</td>
</tr>
<tr>
<td>Capitated</td>
<td>0.552</td>
<td>0.364</td>
<td>0.070</td>
<td>0.18</td>
</tr>
<tr>
<td>Average price of branded drugs</td>
<td>0.597</td>
<td>0.502</td>
<td>0.114</td>
<td>0.74</td>
</tr>
<tr>
<td>Imputed price</td>
<td>0.244</td>
<td>0.260</td>
<td>0.074</td>
<td>1.00</td>
</tr>
<tr>
<td>Visit utility</td>
<td>-1.075</td>
<td>-1.138</td>
<td>0.039</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Doctor utility function:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight on patient utility</td>
<td>0.568</td>
<td>0.580</td>
<td>0.032</td>
<td>0.78</td>
</tr>
<tr>
<td>Plan price</td>
<td>-0.399</td>
<td>-0.372</td>
<td>0.022</td>
<td>0.66</td>
</tr>
<tr>
<td>Plan price $\times$ capitated</td>
<td>-0.595</td>
<td>-0.599</td>
<td>0.021</td>
<td>0.92</td>
</tr>
<tr>
<td>Detailing</td>
<td>0.903</td>
<td>0.861</td>
<td>0.051</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table shows the results from a Monte Carlo experiment in which the data were simulated 50 times and the model estimated in each simulated data set. The first column gives the parameter value in the simulation. The column labelled $E[\beta]$ gives the average point estimate and $SD[\beta]$ gives their standard deviation across data sets. The final column gives the coverage rate for 90% confidence intervals. These confidence intervals are constructed from robust standard errors that allow for arbitrary within-plan correlation.
Table 32: Results from the Monte Carlo Experiment: Match Quality Distribution

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Doctor mean</th>
<th>Patient mean</th>
<th>Patient SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipitor</td>
<td>0</td>
<td>0.016</td>
<td>0.044</td>
</tr>
<tr>
<td>Zocor</td>
<td>0.31</td>
<td>-1.811</td>
<td>0.077</td>
</tr>
<tr>
<td>Mevacor</td>
<td>-2.067</td>
<td>1.738</td>
<td>0.085</td>
</tr>
<tr>
<td>Pravachol</td>
<td>-1.268</td>
<td>-1.314</td>
<td>0.059</td>
</tr>
<tr>
<td>Crestor</td>
<td>-2.305</td>
<td>-2.338</td>
<td>0.077</td>
</tr>
<tr>
<td>Lescol</td>
<td>-1.158</td>
<td>-1.089</td>
<td>0.076</td>
</tr>
<tr>
<td>Zetia/Vytorin</td>
<td>-2.518</td>
<td>-2.587</td>
<td>0.095</td>
</tr>
<tr>
<td>Niacin Combo</td>
<td>0.849</td>
<td>0.901</td>
<td>0.074</td>
</tr>
<tr>
<td>Fibrates</td>
<td>-0.226</td>
<td>-0.234</td>
<td>0.094</td>
</tr>
<tr>
<td>Niacin</td>
<td>-0.767</td>
<td>-0.842</td>
<td>0.087</td>
</tr>
<tr>
<td>BARs</td>
<td>-0.266</td>
<td>-0.234</td>
<td>0.094</td>
</tr>
</tbody>
</table>

Table shows the results from a Monte Carlo experiment in which the data were simulated 50 times and the model estimated in each simulated data set. The first column gives the parameter value in the simulation. The column labelled $E[\beta]$ gives the average point estimate and $SD[\beta]$ gives their standard deviation across data sets. The final column gives the coverage rate for 90% confidence intervals. These confidence intervals are constructed from robust standard errors that allow for arbitrary within-plan correlation.
Figure 12: Distribution of prices in two plans, 2003

The figure shows the empirical distribution of prices paid for 30 days supplied of anti-cholesterol drugs, in two plans from 2003.

A.6. Appendix tables for Chapter 3
Figure 13: Empirical distribution of prices in 2001

Figure shows the empirical distribution of out-of-pocket prices for 30 days supplied of anti-cholesterol drugs, for the 24 largest plans in 2001. These plans contained 90.6% of all claims for 30 days supplied in 2001. The plotted distributions are left censored at zero and right censored at the 99.5th percentile of prices paid. Each panel shows the distribution for a separate plan, and reports the number of claims for that plan, the percent of claims at one of the top four modes, and the percent of claims that are censored. The dashed and solid lines show the generic and branded price reported by the plan documentation.
Table 33: Robustness to CP(2) in permanent income and preferences

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Annuity income</th>
<th>Has annuity income</th>
<th>Has “true” annuity income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Reduced form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln price</td>
<td>(-8.55^{**})</td>
<td>(-0.28^{***})</td>
<td>(-0.04^{*})</td>
</tr>
<tr>
<td></td>
<td>((3.81))</td>
<td>((0.08))</td>
<td>((0.02))</td>
</tr>
<tr>
<td>Panel B: IV results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has Medigap</td>
<td>(20.09^{**})</td>
<td>(0.66^{***})</td>
<td>(0.10^{*})</td>
</tr>
<tr>
<td></td>
<td>((9.57))</td>
<td>((0.24))</td>
<td>((0.06))</td>
</tr>
<tr>
<td>Elasticity</td>
<td>(.92)</td>
<td>(.54)</td>
<td>(.86)</td>
</tr>
<tr>
<td># Obs</td>
<td>3163</td>
<td>3163</td>
<td>3163</td>
</tr>
<tr>
<td># People</td>
<td>1322</td>
<td>1322</td>
<td>1322</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the indicated outcome, and the instrumental variable estimate of Medigap coverage and Medicaid policy variables, with price as the excluded instrument. The sample is described in the main text. Additional controls are a quartic in age; sex, race, marital status, and education dummies; year fixed effects; a cubic in log average yearly income, plus log average yearly income interacted with age; dummies for chronic conditions and self-reported health status, plus a complete polynomial of degree two in log permanent income, subjective risk tolerance, subjective mortality, interacted with an indicator for good or excellent health. Elasticities are calculated at sample means. Standard errors clustered on individual in parentheses. *, **, and *** indicate \(p < 0.1, p < 0.05\), and \(p < 0.01\).
Table 34: Control for pre-age 65 endogenous variables

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1) Basic</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6) Full controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable:</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity income</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity income</td>
</tr>
<tr>
<td>ln price</td>
<td>(-9.14^{**})</td>
<td>(-0.29^{***})</td>
<td>(-0.04^{*})</td>
<td>(-8.32^{**})</td>
<td>(-0.26^{***})</td>
<td>(-0.04^{*})</td>
</tr>
<tr>
<td></td>
<td>(3.91)</td>
<td>(0.09)</td>
<td>(0.02)</td>
<td>(4.00)</td>
<td>(0.08)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Panel A: Reduced form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has Medigap</td>
<td>21.52^{**}</td>
<td>0.67^{***}</td>
<td>0.10^{*}</td>
<td>20.89^{**}</td>
<td>0.66^{***}</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>(9.96)</td>
<td>(0.25)</td>
<td>(0.06)</td>
<td>(10.58)</td>
<td>(0.25)</td>
<td>(0.06)</td>
</tr>
<tr>
<td># Obs</td>
<td>3138</td>
<td>3138</td>
<td>3138</td>
<td>3138</td>
<td>3138</td>
<td>3138</td>
</tr>
<tr>
<td># People</td>
<td>1312</td>
<td>1312</td>
<td>1312</td>
<td>1312</td>
<td>1312</td>
<td>1312</td>
</tr>
<tr>
<td>Controls:</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permanent inc.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Health</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Preference</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the indicated outcome, and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. The sample is described in the main text. Basic controls are a quartic in age, sex, race, marital status, and education dummies, and year fixed effects. The sample is described in the main text. Basic controls are a quartic in age, sex, race, marital status, and education dummies, and year fixed effects. Full controls add controls for a cubic in log average yearly income, plus log average yearly income interacted with age; dummies for chronic conditions and self-reported health status; and subjective measures of risk tolerance and mortality. Standard errors clustered on individual in parentheses. *, **, and *** indicate $p < 0.1$, $p < 0.05$, and $p < 0.01$. 
## Table 35: Control for spouse’s preferences

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls:</td>
<td>Spouse preferences</td>
<td>Spouse preferences, interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent variable:</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity</td>
<td>Annuity income</td>
<td>Has annuity income</td>
<td>Has “true” annuity</td>
</tr>
<tr>
<td><code>ln price</code></td>
<td>$-10.07^*$</td>
<td>$-0.29^{***}$</td>
<td>$-0.05^*$</td>
<td>$-9.89^*$</td>
<td>$-0.28^{***}$</td>
<td>$-0.05^*$</td>
</tr>
<tr>
<td></td>
<td>(5.21)</td>
<td>(0.09)</td>
<td>(0.03)</td>
<td>(5.35)</td>
<td>(0.09)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Panel A: Reduced form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>Has Medigap</code></td>
<td>$27.02^*$</td>
<td>$0.77^{***}$</td>
<td>$0.13^*$</td>
<td>$26.60^*$</td>
<td>$0.75^{***}$</td>
<td>$0.12^*$</td>
</tr>
<tr>
<td></td>
<td>(14.85)</td>
<td>(0.29)</td>
<td>(0.07)</td>
<td>(15.18)</td>
<td>(0.29)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Elasticity</td>
<td>1.21</td>
<td>.62</td>
<td>1.15</td>
<td>1.2</td>
<td>.61</td>
<td>1.09</td>
</tr>
<tr>
<td># Obs</td>
<td>2949</td>
<td>2949</td>
<td>2949</td>
<td>2949</td>
<td>2949</td>
<td>2949</td>
</tr>
<tr>
<td># People</td>
<td>1231</td>
<td>1231</td>
<td>1231</td>
<td>1231</td>
<td>1231</td>
<td>1231</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the indicated outcome, and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. Spouse preferences are spouse’s subjective risk tolerance and mortality. The interactions are a complete polynomial of degree 2 in spouse risk tolerance, spouse mortality, own risk tolerance, and own mortality. The sample is described in the main text. Basic controls are a quartic in age, sex, race, marital status, and education dummies, and year fixed effects. The sample is described in the main text. Basic controls are a quartic in age, sex, race, marital status, and education dummies, and year fixed effects. Full controls add controls for a cubic in log average yearly income, plus log average yearly income interacted with age; dummies for chronic conditions and self-reported health status; and subjective measures of risk tolerance and mortality. Standard errors clustered on individual in parentheses. *, **, and *** indicate $p < 0.1$, $p < 0.05$, and $p < 0.01$. 
Table 36: Exclude Medicaid recipients

<table>
<thead>
<tr>
<th>Controls:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent variable:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annuity income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has annuity income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has “true” annuity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel A: Reduced form

<table>
<thead>
<tr>
<th>ln price</th>
<th>−9.36**</th>
<th>−0.28***</th>
<th>−0.05</th>
<th>−8.61**</th>
<th>−0.28***</th>
<th>−0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(4.62)</td>
<td>(0.10)</td>
<td>(0.03)</td>
<td>(4.37)</td>
<td>(0.09)</td>
<td>(0.03)</td>
</tr>
</tbody>
</table>

Panel B: IV

<table>
<thead>
<tr>
<th>Has Medigap</th>
<th>21.50*</th>
<th>0.65**</th>
<th>0.10</th>
<th>20.10*</th>
<th>0.64**</th>
<th>0.11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(11.44)</td>
<td>(0.27)</td>
<td>(0.07)</td>
<td>(10.91)</td>
<td>(0.27)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elasticity</th>
<th>0.97</th>
<th>0.54</th>
<th>0.90</th>
<th>0.90</th>
<th>0.53</th>
<th>0.91</th>
</tr>
</thead>
<tbody>
<tr>
<td># Obs</td>
<td>2705</td>
<td>2705</td>
<td>2705</td>
<td>2705</td>
<td>2705</td>
<td>2705</td>
</tr>
<tr>
<td># People</td>
<td>1179</td>
<td>1179</td>
<td>1179</td>
<td>1179</td>
<td>1179</td>
<td>1179</td>
</tr>
</tbody>
</table>

Notes: The table shows the reduced form impact of Medigap price on the indicated outcome, and the instrumental variable estimate of Medigap coverage, with price as the excluded instrument. The sample is described in the main text, except dropping people with Medicaid coverage. The sample is described in the main text. Basic controls are a quartic in age, sex, race, marital status, and education dummies, and year fixed effects. Full controls add controls for a cubic in log average yearly income, plus log average yearly income interacted with age; dummies for chronic conditions and self-reported health status; and subjective measures of risk tolerance and mortality. Standard errors clustered on individual in parentheses. *, **, and *** indicate p < 0.1, p < 0.05, and p < 0.01.
A.7. Appendix: Additional Features of Annual Earnings Test

When current benefits are lost to the AET, future scheduled benefits are increased in some circumstances. This is sometimes called ”benefit enhancement.” As we describe, for workers NRA or older in the pre-2000 period (when they faced the AET), benefit enhancement attenuates the effective AET BRR for individuals considering earning enough to trigger the benefit enhancement, but it does not attenuate the effective AET BRR for those considering earning less than this amount.

The benefit enhancement rules have varied over time, and they depend on whether the beneficiary is above or below NRA. Prior to 1972, there was no benefit enhancement for people aged NRA and older. In these years, the AET represented a pure loss in benefits for those NRA and above (equivalent to a pure tax). For beneficiaries NRA and older, a one percent Delayed Retirement Credit (DRC) was introduced in 1972. The DRC was intended to compensate beneficiaries who delayed claiming beyond age 65, but they also apply to earnings lost to the AET. For individuals above NRA, benefits are increased 1/12 of 1 percent for each month between ages 65 and 72 for which no benefits received after 1972 (Social Security, 2012, Table 2.A.20).

This language indicates that each month’s worth of foregone benefits—either because of delayed claiming or because of the AET—results in increased future benefits. A beneficiary has to forego an entire month of benefits in order to receive the DRC; if, for example, she earns slightly over the exempt amount and loses only a small amount of benefits to the AET, then her future benefits are not adjusted. Thus, the DRC provides no marginal relief from the AET for a claimant who is considering earning near the exempt amount: no benefit enhancement occurs when she earns a marginal dollar at or near the AET earnings threshold. Meanwhile, if she earns enough to forego an entire month’s worth of benefits

---

11 This section is based on table 2.A.20 of the Annual Statistical Supplement of the Social Security bulletin, as well as extensive email correspondence with numerous officials at the Social Security Administration.
12 The size of the DRC was increased to three percent per year in 1982, and then increased steadily throughout the 1990s, reaching eight percent for each year of foregone benefits in 2008. Starting in 1983, benefit enhancement only applied through age 69.
(but not when a smaller amount of benefits is lost due to the AET), future benefits are increased by 1/12 of 1 percent.

As a result of these rules, future benefits are enhanced when the individual’s yearly earnings are over \( z^* + (MB/\tau) \), where \( z^* \) is the exempt amount, \( MB \) is the monthly benefit, and \( \tau \) is the AET benefit reduction rate.\(^{13}\) For example, with a typical monthly benefit of $1,000 and a benefit reduction rate of 33.33 percent, benefit enhancement occurs when the individual’s yearly earnings are $3,000 (=$1000/0.3333) above the exempt amount. Benefit enhancement corresponding to one more month of reduced earnings occurs once annual earnings reaches $6,000 above the exempt amount, and so forth. Thus, benefit enhancement is only relevant to an individual considering earning substantially in excess of the exempt amount and is therefore not relevant to marginal earnings decisions at the exempt amount. Indeed, this theoretical presumption is consistent with suggestive evidence we describe that indicates little systematic bunching reaction to changes in the DRC and that mean age at death is smooth near the exempt amount.

The AET is implemented in a number of stages. First, SSA must determine that a claimant is expected to exceed the exempt amount, or that she has already done so. Claimants can notify SSA in advance if they expect to exceed the exempt amount, or they can report their earnings \textit{ex post facto} at any point in the year. In addition, SSA uses W-2 records at the end of the year to determine if the AET threshold has been crossed (for those who have a W-2). Second, SSA withholds OASI benefits in monthly increments, until enough benefits have been withheld to cover the AET penalty amount. For example, assume an individual aged 66 with a monthly benefit of $1,200 earns $1,800 dollars beyond the AET exempt amount in 1992, when the benefit reduction rate was 0.3333. This individual should receive a yearly benefit reduction of $1,800 \times 0.3333 = $600. SSA withholds an entire month’s check, $1,200, in order to collect the $600. Finally, at the end of the year, SSA refunds any

\(^{13}\)Another month’s benefit enhancement would occur if the individual earns more than \( z^* +(2MB/\tau) \); a third month’s benefit enhancement would occur if she earns more than \( z^* +(3MB/\tau) \); and so on. Note that this creates 12 notches in the budget set, the final one at \( z^* +(12MB/\tau) \).
overwithheld benefits. In the same example, at the end of the year SSA would return $600 in overwithheld benefits. Importantly, the DRC is not applied to future benefits in this case—less than a month’s worth of benefits, $600, was ultimately collected by SSA, after factoring overwithholding and refunds. After considering both withholding and refunds, the AET is ultimately applied at a yearly level—much in the same way that the Earned Income Tax Credit (EITC) is applied at the annual level but the receipt of the credit depends on one’s withholding patterns and when the income tax return is filed.

In sum, for people NRA and older, the AET effectively acts as a kink for those earning close enough to the exempt amount, and benefit enhancement does not attenuate the marginal work disincentives associated with the AET in this range of earnings. However, the DRC is relevant to an individual considering earning enough to reduce her OASI benefit by at least a month’s worth (i.e. at least $z^* + MB/\tau$). Empirically, we find that limiting the sample to those with substantial OASI benefits—for whom this earnings level is several thousand dollars above the AET exempt amount, and for whom the notch created by the DRC is therefore less relevant—yields very similar results to those we have shown. Our empirical specification alternatively assumes that benefit enhancement does not affect the AET implicit marginal tax rate (or does), and we find similar patterns in both specifications.

Note that in our empirical estimation, the region we ”dummy out” near the kink $z^*$ is $[z^* - 2800, z^* + 2800]$. Thus, if all of the ”bunchers” arrived at the kink from initial earnings levels between $z^*$ and $z^* + 2800$, we could find zero bunching despite substantial actual bunching. However, the earnings densities clearly show that the polynomial does not substantially overpredict bunching in the region above the exempt amount. Moreover, the densities also show no evidence of bunching near notches in the budget set created by the DRC. We alternatively use a bandwidth of $500 and find similar results.

Individuals younger than NRA are subject to different rules for benefit adjustment, called ”actuarial adjustment.” The rule for this younger group was introduced in the legislation allowing people younger than 65 to claim early benefits (in 1956 for women and in 1961 for
men). For those younger than NRA, future benefits are reduced 5/9 of 1 percent for each month under age 65 in which an individual claims benefits (Social Security, 2012, Table 2.A.20). This implies that if a beneficiary has any income withheld under the AET in a given month, then she receives a full benefit enhancement for that month. On the other hand, if a beneficiary does not have any income withheld under the AET in a given month, then she receives no benefit enhancement for that month. This creates a notch at the exempt amount—a discontinuous increase in future benefits when moving from just under the exempt amount to just over it—creating incentives to bunch just above the exempt amount in order to receive the monthly (or yearly) benefit adjustment. We find no evidence for this kind of behavior; in fact, as we document in Gelber, Jones, and Sacks (2013), people tend to have earnings just below the exempt amount, exactly opposite the behavior we would expect if people were responding to the notch just described. Because benefit enhancement occurs at exactly the same earnings level that the AET begins to apply for those under NRA, we focus on the group NRA and above.

Formally, the number of months’ worth of benefit enhancement enjoyed by OASI recipients is therefore

\[ \text{floor}(\tau \cdot (z - z^*)/MB) \]

for those NRA and above, and

\[ \text{ceiling}(\tau \cdot (z - z^*)/MB) \]

for those below NRA.

Benefit enhancement is actuarially fair if the net present value of the benefit enhancement equals the benefits lost due to the AET. The actuarial adjustment is approximately actuarially fair in the sense that delaying OASI claiming an extra year is approximately actuarially fair; however, this does not imply that benefit enhancement is actuarially fair when an additional dollar of benefits is withheld due to the AET. For example, actuarial adjustment is not actuarially fair for (among others) those with positive OASI benefits considering earning an additional amount above the AET exempt amount, because this does not result in additional benefit enhancement. Similar considerations apply to the DRC: additional marginal increments of earnings are not compensated through benefit enhancement (except in the case when an individual goes from earning just under to just over \( z^* + (MB/\tau) \) (or
one of the other 12 thresholds).

The AET applies to an individual’s earnings; spouses’ earnings do not count in the earnings total to which the AET is applied. For a retired worker (i.e. primary) beneficiary whose spouse collects spousal benefits, the AET reduces the family’s OASI benefit by the amounts we have described. The family benefit is also reduced when the spouse (separately) earns more than the AET threshold. For a retired worker beneficiary whose spouse is collecting benefits on his or her own earnings record, the AET reduces the retired worker beneficiary’s benefits by the amounts described while not affecting the spouse’s benefits. Thus, following previous literature (e.g. Friedberg 1998, 2000), we model the AET as creating the MTRs associated with the BRRs described, because the AET reduces family benefits by these amounts (all else equal). Our data do not contain the information necessary to link spouses (except when one spouse is claiming OASI benefits on the other spouse’s record).

It is also worth noting how the actuarial adjustment and DRC interact with incentives for claiming OASI. Under the actuarial adjustment, the full benefit enhancement occurs when the individual earnings over the threshold level. Thus, the individual could in principle claim OASI; earn just over this threshold level; collect nearly her entire OASI benefit in this year (since the AET only reduces current OASI benefits at the margin); and later enjoy full benefit enhancement. This illustrates the more general point that it can be in an individual’s interest to claim OASI even if the individual faces the AET. More generally, for individuals for whom the AET reduces OASI benefits sufficiently little, and for whom current OASI benefits are sufficiently important, it can be in their interest to claim OASI even if they face the AET. Appendix Figure 35 shows that among the sample of individuals who have not claimed by year $t$, the hazard of claiming at year $t + 1$ is smooth near the kink, indicating no evidence that claimants come disproportionately from close to or far from the kink.
A.8. Appendix: Procedure for Estimating Normalized Excess Bunching (for online publication)

In order to estimate excess normalized bunching, we use the following procedure. For each centered-earnings bin \( z_i \), we calculate \( p_i \), the proportion of all people with earnings in the range \( [z_i - k/2, z_i + k/2] \) (in a given time period and for a given age group). For example, underlying the first panel in Figure 5 is the probability \( p_i \) of earning in various bins \( z_i \) for 62 year-olds in the 1990-1999 period. The earnings bins are normalized by distance-to-kink, so that for \( z_i = 0 \), \( p_i \) is the fraction of people with earnings in the range \( [-k/2, k/2] \). To estimate bunching, we assume that \( p_i \) can be written as

\[
p_i = \sum_{d=0}^{D} \beta_d(z_i)^d + \sum_{n=-k}^{k} \gamma_k 1\{z_i = k\delta\} + \varepsilon_i
\]

(A.1)

and run this regression. This equation expresses the earnings distribution as a degree \( D \) polynomial, plus a set of indicators for each bin within \( k\delta \) of the kink, where \( \delta \) is the binwidth. In our empirical application, we choose \( D = 7, \delta = 800 \) and \( k = 3 \) (so that seven bins are excluded from the polynomial estimation, including the bin centered at the kink).

We show that our results to alternative choices of \( D, \delta, \) and \( k \).

Our measure of excess mass is \( \hat{EM} = \sum_{n=-k}^{k} \hat{\gamma}_k \), the estimated excess probability of locating at the kink (relative to the polynomial term). This measure depends on the counterfactual density near the kink, so to obtain a measure of excess mass that is comparable at the kink, we scale by the predicted density that would obtain if there were no bunching. This is just the constant term in the polynomial, since the \( z_i \) is distance to zero. So our estimate of normalized excess mass is

\[
\hat{B} = \frac{\hat{EM}}{\beta_0}.
\]

(A.2)

We consider two approaches for constructing standard errors. First, from Equation (A.2), it is straightforward to apply the Delta method. Second, we employ the parametric bootstrap procedure of Chetty, Friedman, Olsen and Pistaferri (2011). This bootstrap draws with
replacement from the estimated distribution of errors $\varepsilon_i$ from Equation (A.1). For each set of draws, we get a new value of $p_i$ and use these new values to re-estimate $B$. The standard deviation across draws of $B$ is our measure of the standard error $\hat{B}$. In practice these two procedures produced extremely similar results, so we only report standard errors from the bootstrap.

A.9. Appendix: Social Security Data

Our data come from the Social Security Master Earnings File (MEF), which is described more extensively in Song and Manchester (2007). The MEF is a longitudinal history of Social Security taxable earnings for all Social Security Numbers (SSNs) in the U.S. Our data are a one percent random sample of SSNs; we randomly extract SSNs from the database and follow each of these individuals over the full time period. The AET is based on earnings as measured in this dataset. Prior to 1978, the data have information on annual FICA earnings; since 1978, the data have information on uncapped wage compensation. Before 1978, the data do not clearly distinguish between earnings from self-employment and non-self-employment earnings, but we are able to distinguish them in the data starting in 1978. The data also contain information on date of birth, date of death, and sex.

We supplement the MEF with information from the Master Beneficiary Record (MBR) file, which contains data on the day, month, and year that people began to claim Social Security (and other variables). The majority of workers excluded from OASDI coverage are in four main categories: (1) federal civilian employees hired before January 1, 1984; (2) agricultural workers and domestic workers whose earnings do not meet certain minimum requirements; (3) individuals with very low net earnings from self-employment (generally less than $400 per year); and (4) employees of several state and local governments. However, civil service and other government workers are covered by Medicare and are therefore present in the MBR.

In choosing our main sample, we take into account a number of considerations. It is desirable
to show a constant sample in making comparisons of earnings densities. Meanwhile, the AET only affects people who claim OASI, and thus we wish to focus on claimants. However, many individuals claim OASI at ages over the Early Entitlement Age (62), implying that they have not claimed at younger ages but have claimed by older ages. This implies that to investigate a constant sample, we cannot simply limit the sample to claimants at each age (because many people move from not claiming to claiming). To balance these considerations, our main sample at each age and year consists of individuals who ultimately claim in the year they turn 65 or earlier. We show that the results are robust to other sample definitions. Because we focus on the intensive margin response (consistent with Saez (2010) and subsequent papers on bunching), we further limit the sample to observations with positive earnings in our main analysis.

Information on AET parameters is from table 2.A.20 and 2.A.29 of the Annual Statistical Supplement to the Social Security Bulletin. Friedberg (1998, 2000) provides a thorough description of these rules. All dollar amounts are deflated to 2010 dollars using the CPI-U.

The standard deviation is large because of very rare aberrant large values of earnings (as documented in Utendorf 2001/2); these do not affect our estimates in the figures or tables because they are far above the AET exempt amount, and they affect mean earnings far less than they affect the standard deviation. The results are robust to winsorizing.

In 1983-1999, the AET is assessed on earnings until the month in which the individual turns age 70. For simplicity, in our baseline sample we measure age as calendar year minus year of birth. Thus, if an individual turns age 70 later in the year—in the extreme case, on December 31—she will have had an incentive to bunch at the kink during nearly the entire year when she is classified as age 70 in our data. As a result, her yearly earnings may appear to be located at or near the kink even though she is bunching at the kink applicable to 69-year-olds through almost all of the calendar year over which her earnings are observed. However, the figure shows that significant bunching occurs at age 71, which cannot be due to this coarse measure of birth dates. Thus, the results do show a delay in
complete adjustment. We have also found substantial and significant ($p < 0.01$) bunching at age 70 among those born in January, who no longer face the AET immediately in January of the year they turn 70 and therefore should not show excess bunching at this age in the absence of adjustment frictions. Likewise, we find a spike in mean earnings growth from age 70 to age 71 among those born in January. In our sample period, the AET applied to ages 62-71 before 1983, and it applied to ages under NRA in 2000 and after. In these time periods, examining only those born in January also shows a delay in responding to the removal of the AET.

Since 1978, the earnings test has been assessed on yearly earnings, implying that we analyze the appropriate time period, *i.e.* earnings in a calendar year. Prior to 1978, the earnings test was assessed on quarterly earnings. While there is likely some error in measuring the amount of bunching pre-1978, we believe that this is not a major issue: the patterns of bunching in the pre-1978 period are visually clear and appear unlikely to be changed in a qualitative sense by an examination of quarterly data. Moreover, Figure 8 shows that the amount of excess bunching falls from 1977 to 1978 and subsequent years, rather than rising as we might expect if we hypothetically measured bunching more accurately starting in 1978.

A.10. Appendix: Longitudinal Employer Household Dynamics

We use the Longitudinal Employer Household Dynamics (LEHD) dataset, which contains wage data available from state-level unemployment insurance (UI) programs. These data measure uncapped quarterly earnings for employees covered by state unemployment insurance systems, estimated to cover over 95 percent of private sector employment. Although coverage laws vary slightly from state to state, UI programs do not cover federal employees, the self-employed, and many agricultural workers, domestic workers, churches, nonprofits, and state and local government employees. We examine a 20 percent random sample of the original LEHD file, as this was the largest amount of data that our available server space could handle.
These administrative earnings records are linked across quarters to create individual work histories. In addition to earnings, information on gender and date of birth are available. The data on employees are linked to data on firms. Each firm at which an individual works in a given quarter is identified through a firm identifier. We consider an employee to have changed employers from year $t$ to year $t + 1$ if at least one of the federal employer IDs at which the employee works in year $t$ is different in year $t + 1$. However, when the individual works at one or more employer in year $t$ and does not work at any employer in year $t + 1$, we drop this individual from the sample. The results are similar when we treat these individuals as if they changed employers.

We select data from 1990-1999. During this period, the AET explicit benefit reduction rate was constant. 1990-1999 also represent natural years to investigate because large sample sizes are not available in the LEHD prior to 1990. When we include other years and age groups in the LEHD sample, we find similar results to those reported here. Note that the population we investigate is not constant over this period, because (among other reasons) an increasingly broad set of states is included in the LEHD over time. Data are available on 13 states in 1990, climbing to 28 states by 1999. In a given quarter, we include in our sample all states whose data are available. Holding the sample constant yields very similar results.

A.11. Appendix: Model of Earnings Response

A.11.1. Baseline Model

We start with a baseline, frictionless model of earnings, following Saez (2010). We briefly sketch the key features of this model for comparison to our model with a fixed cost of adjustment. In Saez (2010), individuals maximize utility over consumption, $c$, and costly earnings, $z$:

$$ u(c, z; n) $$
Heterogeneity is parameterized by an "ability" parameter \( n \), which is distributed according to the smooth cdf \( F(\cdot) \). Individuals maximize utility subject to the following budget constraint:

\[
c = (1 - \tau) z + R
\]

where \( R \) is virtual income. This leads to the first order condition:

\[
-\frac{u_z(c, z; n)}{u_c(c, z; n)} = (1 - \tau),
\]

which implicitly defines an earnings supply function: \( z(1 - \tau, R, n) \).

When necessary, we will use a quasi-linear and iso-elastic utility function:

\[
u(c, z; n) = c - \frac{n}{1 + 1/\varepsilon} \left( \frac{z}{n} \right)^{1+1/\varepsilon}\]

Under this assumption, the first order condition simplifies to:

\[
(1 - \tau) - \left( \frac{z}{n} \right)^{\frac{1}{\varepsilon}} = 0,
\]

which implies this earnings supply function:

\[
z = n (1 - \tau)^\varepsilon.
\]

**A.11.2. Linear Tax Schedule**

Consider first a linear tax schedule with a constant marginal tax rate \( \tau_0 \). Observe that with a smooth distribution of skills \( n \), we have a smooth distribution of earnings that is monotonic in skill, provided we make the typical Spence-Mirlees assumption. Let \( H_0(\cdot) \) denote the cumulative distribution function (CDF) of earnings under the constant marginal tax rate, and let \( h_0(\cdot) = H'_0(\cdot) \) denote the density of this distribution. Under quasilinear
utility, we have:

\[ H_0(z) = F \left( \frac{z}{(1 - \tau_0)} \right). \]

Define \( H_1(\cdot) \) and \( h_1(\cdot) \) as the smooth CDF and density of earnings under a higher, constant marginal tax rate \( \tau_1 \); \( H_1 \) is defined similarly as a function of \( \tau_1 \).

### A.11.3. Kinked Tax Schedule

Now consider a piecewise linear tax schedule with a convex kink: the marginal tax rate below earnings level \( z^* \) is \( \tau_0 \), and the marginal tax rate above \( z^* \) is \( \tau_1 > \tau_0 \). Given the tax schedule, individuals bunch at the kink point \( z^* \); as explained in Saez (2010), the realized density in earnings has an excess mass at \( z^* \). Denote the realized distribution of earnings once the kink has been introduced at \( z^* \) as \( H(\cdot) \):

\[
H(z) = \begin{cases} 
H_0(z) & \text{if } z < z^* \\
H_1(z) & \text{if } z \geq z^*
\end{cases}
\]

Denote the density of this realized distribution as \( h(\cdot) = H'(\cdot) \). In general there is now a discrete jump in the earnings density at \( z^* \):

\[
h(z) = \begin{cases} 
h_0(z) & \text{if } z < z^* \\
h_1(z) & \text{if } z > z^*
\end{cases}
\]

The share of people who relocate to the kink is:

\[
B = \int_{z^*}^{z^* + \Delta z^*} h_0(\zeta) \, d\zeta
\]

These "bunchers" are those whose \textit{ex ante} earnings lie in the range \([z^*, z^* + \Delta z^*] \), who are induced to locate at the kink by the rise in the MTR above the kink point. For relatively small changes in the tax rate, we can relate the elasticity of earnings with respect to the net-of-tax rate to the earnings change \( \triangle z^* \) for the individual with the highest \textit{ex ante} earnings.
who bunches \textit{ex post}:

\[ \varepsilon = \frac{\Delta z^*/z^*}{d\tau_1/(1-\tau_0)} \]

where \( d\tau_1 = \tau_1 - \tau_0 \).\(^{14}\)

\textbf{A.11.4. Graphical Exposition}

Appendix Figure 14 depicts a setting in which a kinked budget set is introduced. In Panels A and B, the x-axis shows before-tax-and-transfer income, \( z \), and the y-axis shows after-tax-and-transfer consumption, \( z - T(z) \). Consider first a linear tax (Panel A) at a rate of \( \tau \). An individual optimally locates at a point of tangency, where the marginal rate of substitution (MRS) between earnings and consumption equals the net-of-tax rate, \( 1 - \tau \). The figure shows indifference curves and earnings levels for low- and high-earning agents (labeled \( L \) and \( H \), respectively). The low earner has an earnings level of \( z^* \), while the high earner receives \( z^* + \Delta z \).

Suppose the AET is introduced (on top of pre-existing taxes), so that the marginal net-of-tax rate decreases to \( 1 - \tau - d\tau \) for earnings above a threshold \( z^* \). For small \( d\tau \), individuals earning in the neighborhood above \( z^* \) will reduce their earnings. If ability is smoothly distributed, a range of individuals will locate exactly at \( z^* \), due to the discontinuous jump in the marginal net-of-tax rate at \( z^* \). In Panels A and B, individual \( L \) has the lowest \textit{ex ante} earnings among those who bunch at \( z^* \), individual \( H \) has the highest \textit{ex ante} earnings in this group, and all others previously earning between \( z^* \) and \( z^* + \Delta z \) also bunch at \( z^* \). Those with \textit{ex ante} earnings higher than \( z^* + \Delta z \) reduce their earnings to a level greater than \( z^* \).

Panels C and D of Figure 14 depict densities of earnings we would expect to observe in the absence and presence of the AET, respectively. The x-axis shows before-tax earnings, \( z \), and the y-axis measures the density of earnings. In Panel C, the density is continuous at \( z^* \),

\( ^{14}\)This formula holds if there is a single elasticity \( \varepsilon \) in the population. Under heterogeneity, the method returns \( \bar{\varepsilon} \), the average elasticity among bunchers. We investigate cases with heterogeneity below.
reflecting a smooth distribution of ability. The blue region represents the set of individuals who bunch at \( z^* \) in the presence of the AET, \textit{i.e.} those earning in \([z^*, z^* + \Delta z]\) in the absence of the AET. Panel D shows that once the AET is introduced, these individuals locate in the neighborhood of \( z^* \). However, rather than depicting a mass point exactly at \( z^* \), we have shown bunching in the region at and surrounding \( z^* \), reflecting the fact that individuals often cannot bunch exactly at the kink point (as discussed, for example, in Saez, 2010).

\section*{A.11.5. Fixed Adjustment Costs}

We now extend the model to include a fixed cost of adjusting earnings. We assume that the adjustment cost reflects a disutility of \( \phi^* \) of increasing or decreasing earnings from some initial earnings level. We begin by analyzing the response to a change in the marginal tax rate from \( \tau_0 \) to \( \tau_1 \), where the tax schedule is linear in both cases, in order to build intuition for the case with a kinked budget set. We assume that following a change in tax rates from \( \tau_0 \) to \( \tau_1 \), the gain (absent adjustment costs) to reoptimizing is increasing in \( n \). In general, this requires that the size of the optimal earnings adjustment increases in \( n \) at a rate faster than the decrease in the marginal utility of consumption.\footnote{To see this, note that the utility gain from reoptimizing is \( u((1 - \tau_1)z_1 + R_1, z_1; n) - u((1 - \tau_1)z_0 + R_0, z_0; n) \approx u_c \cdot (1 - \tau_1)(z_1 - z_0) + u_z \cdot \tau_1 \cdot (z_1 - z_0) = u_c \cdot (1 - \tau_0)(z_0 - z_1), \) where in the first expression, we have used a first-order approximation for utility at \((1 - \tau_0)z_0 + R_0, z_0)\) and in the second expression we have used the first order condition \( u_z = -u_c (1 - \tau_0) \). The gain in utility is approximately equal to an expression that depends on the marginal utility of consumption, the change in tax rates, and the size of the earnings adjustment. The first term, \( u_c \), is decreasing as \( n \) (and therefore initial earnings \( z_0 \)) increases. Thus, in order for the gain in utility to be increasing in \( n \), we need the size of earnings adjustment \([z_0 - z_1]\) to increase at a rate that dominates.} This is true, for example, if utility is quasilinear.

If the gain in utility is monotonically increasing in initial earnings, and the cost of adjustment is fixed, there exists a unique level of initial earnings at which the agent is indifferent between adjusting and staying at the initial earnings level. We formally state the implications in the following result:

\textbf{Remark 1} (Linear Tax Change and Adjustment Costs).
After a change in linear tax rates from $\tau_0$ to $\tau_1$, if there is a constant adjustment cost of $\phi^*$ and the size of the optimal earnings adjustment increases in $n$ at a rate faster than the decrease in the marginal utility of consumption, then:

1. There is a unique threshold of initial earnings, $z_{0,\phi}$, above which all individuals will adjust their earnings in response to the tax change. Those initially locating below the threshold will not adjust.

2. The threshold level of earnings satisfies the following identity:

$$u((1 - \tau_1)z_{1,\phi} + R_1, z_{1,\phi}) - u((1 - \tau_1)z_{0,\phi} + R_1, z_{0,\phi}) \equiv \phi^*$$

where $z_{1,\phi}$ is the ex post earnings level of the individual who initially locates at $z_{0,\phi}$.

In other words, at the threshold level, the gain in $u$ from adjusting earnings is exactly equal to the adjustment cost $\phi^*$.

3. In the case of quasilinear utility, the threshold level of earnings is:

$$z_{0,\phi} = \frac{\phi^*}{\alpha(\varepsilon, \tau_0, \tau_1)}$$

where

$$\alpha(\varepsilon, \tau_0, \tau_1) \equiv \frac{1 - \tau_1}{1 + \varepsilon} \left[ \left( \frac{1 - \tau_1}{1 - \tau_0} \right)\varepsilon - 1 + \varepsilon \left( \frac{\tau_1 - \tau_0}{1 - \tau_1} \right) \right].$$

178
4. The ex post distribution of earnings is:

\[ H(z) = \begin{cases} 
H_0(z) & \text{if } z < z_{1,\phi} \\
H_0(z) + H_1(z) - H_0(z_{0,\phi}) & \text{if } z \in [z_{1,\phi}, z_{0,\phi}] \\
H_1(z) & \text{if } z > z_{0,\phi} 
\end{cases} \]

\[ h(z) = \begin{cases} 
h_0(z) & \text{if } z < z_{1,\phi} \\
h_0(z) + h_1(z) & \text{if } z \in [z_{1,\phi}, z_{0,\phi}] \\
h_1(z) & \text{if } z > z_{0,\phi} 
\end{cases} \]

where \( H_0(\cdot) \) and \( H_1(\cdot) \) are the CDFs of earnings in the presence of linear tax rates \( \tau_0 \) and \( \tau_1 \), respectively.

Next, consider choices in the presence of adjustment costs on a budget set with a convex kink. Consider again an initial linear tax schedule with marginal tax rate \( \tau_0 \). Now, introduce a higher MTR \( \tau_1 > \tau_0 \) for earnings above \( z^* \). We again assume that the gain to reoptimizing is increasing in initial earnings over the range \([z^*, z^* + \Delta z^*]\). Using the same logic as above—the gain in utility is monotonically increasing in initial earnings, and the cost of adjustment is fixed—there exists a unique level of initial earnings at which the agent is indifferent between adjusting and staying at the initial earnings level. Thus, we have the following result:

**Remark 2 (Non-Linear Tax and Adjustment Costs).**

When a kink is introduced in the budget set (i.e. a jump in marginal tax rates from \( \tau_0 \) below \( z^* \) to \( \tau_1 \) above \( z^* \)), there is a fixed adjustment cost of \( \phi^* \), and \( z^* \geq z_{1,\phi} \), then:

1. Individuals with initial earnings below a unique threshold \( \bar{z} \) do not adjust their earnings.

\[^{16}z_{1,\phi} \text{ is again the ex post level of earnings for the individual who initially locates at } z_{0,\phi} \text{— where } z_{0,\phi} \text{ is the initial earnings level over which individuals adjust their earnings defined above in Remark (1). Note that } \bar{z} \text{ denotes this threshold in the non-linear budget set case, whereas } z_{0,\phi} \text{ denotes this threshold in the linear budget set case.}\]
2. The threshold level of earnings is implicitly defined by the following:

\[
\begin{align*}
    u((1 - \tau_1)z^* + R_1, z^*) - u((1 - \tau_1)\bar{z} + R_1, \bar{z}) & \equiv \phi^* \\
    z^* & \leq \bar{z} \leq z^* + \Delta z^*.
\end{align*}
\]

3. Individuals with initial earnings in \([\bar{z}, z^* + \Delta z^*]\) bunch at the kink point \(z^*\).

4. Individuals with initial earnings above \(z^* + \Delta z^*\) reduce their earnings to a new level of earnings higher than \(z^*\).

5. The ex post distribution of earnings is:

\[
H(z) = \begin{cases} 
    H_0(z) & \text{if } z < z^* \\
    H_0(z) + H_1(z) - H_0(\bar{z}) & \text{if } z \in [z^*, \bar{z}] \\
    H_1(z) & \text{if } z > \bar{z}
\end{cases}
\]

\[
h(z) = \begin{cases} 
    h_0(z) & \text{if } z < z^* \\
    h_0(z) + h_1(z) & \text{if } z \in [z^*, \bar{z}] \\
    h_1(z) & \text{if } z > \bar{z}
\end{cases}
\]

6. Excess bunching at \(z^*\) is given by:

\[
B = \int_{\bar{z}}^{z^* + \Delta z^*} h_0(\zeta) \, d\zeta
\]

If the kink point \(z^*\) is lower than \(z_{1,\phi}\), then:

1. Individuals only adjust their earnings if their initial earnings level is above the threshold \(z_{0,\phi}\).

2. There is no bunching at \(z^*\).
3. The ex post distribution of earnings is the same as in the case of a change in a linear tax rate from $\tau_0$ to $\tau_1$ described in Remark (1).

A.11.6. Derivation of Closed-Form Solution for Elasticity and Adjustment Cost (Comparative Static Approach)

As we discuss in Remark (2) above and in Section 2.7 in the text, the amount of bunching in the presence of a fixed adjustment cost is equal to the integral of the initial earnings density over the range $[\bar{z}, z^* + \Delta z^*]$: 

$$B(\tau, z^*; \varepsilon, \phi^*) = \int_{\bar{z}}^{z^*+\Delta z^*} h(\zeta) \, d\zeta,$$  

where $\tau \equiv (\tau_0, \tau_1)$. If the density is locally uniform, the integral in (A.4) is:

$$B(\tau, z^*; \varepsilon, \phi^*) \approx h(\bar{z})(z^* + \Delta z^* - \bar{z}),$$  

(A.5)

Taking a first-order Taylor approximation of $u((1 - \tau_1)\bar{z} + R_1, \bar{z}, \bar{n})$ and $u((1 - \tau_1)z^* + R_1, z^*, \bar{n})$ at $((1 - \tau_0)\bar{z} + R_0, \bar{z}, \bar{n})$, and using the first order condition for initial earnings, $(1 - \tau_0)u_c = -u_z$, we have from (A.3):

$$\phi^* \approx u_c \cdot (1 - \tau_1) |z^* - \bar{z}| + u_z \cdot (z^* - \bar{z})$$

$$\Rightarrow \bar{z} \approx z^* + \frac{\phi^*/u_c}{(\tau_1 - \tau_0)}$$

$$= z^* + \frac{\phi}{d\tau_1},$$

where $d\tau_1 = \tau_1 - \tau_0$ and $\phi = \phi^*/u_c$ is the dollar equivalent of the disutility associated with adjusting earnings. Substituting this expression for $\bar{z}$ into (A.5), we have

$$B(\tau, z^*; \varepsilon, \phi) = h(\bar{z})(\Delta z - \phi/d\tau_1),$$
where bunching now depends on the dollar-denominated cost of adjusting, rather than
the utility cost. Finally, for small \(d\tau\), \(\Delta z\) is small and
\( h(\tilde{z}) \approx h(z^* + \Delta z) \approx h(z^*) \). Let \( b \equiv B/h(z^*) \), and note that \( \Delta z^* = z^*(d\tau_1/(1-\tau_0))\varepsilon \). The excess mass at the kink can now be expressed as a linear function of the parameters:

\[
b(\tau, z^*; \varepsilon, \phi) = \varepsilon \left( z^* \frac{d\tau_1}{1-\tau_0} \right) - \phi \left( \frac{1}{d\tau_1} \right).
\]

(A.6)

**Derivation of Formula for Bunching with Heterogeneity**

We derive the formula for bunching \( B \) in the presence of heterogeneity under the Comparative Static approach as follows:

\[
B = \int \int \int z^* + \Delta z^* \tilde{h}(\zeta, \epsilon, \phi) d\zeta d\epsilon d\phi
\]

\[
\approx \int \int [z^* + \Delta z^* - \tilde{z}] \tilde{h}(z^*, \epsilon, \phi) d\epsilon d\phi
\]

\[
\approx \int \int \left[ \varepsilon \left( z^* \frac{d\tau_1}{1-\tau_0} \right) - \varphi \left( \frac{1}{d\tau_1} \right) \right] \tilde{h}(z^*, \epsilon, \phi) d\epsilon d\phi
\]

\[
= h(z^*) \cdot \left[ \left( \int \int \frac{\tilde{h}(z^*, \epsilon, \phi)}{h(z^*)} d\epsilon d\phi \right) \left( z^* \frac{d\tau_1}{1-\tau_0} \right) - \left( \int \int \frac{\varphi(z^*, \epsilon, \phi)}{h(z^*)} d\epsilon d\phi \right) \left( \frac{1}{d\tau_1} \right) \right]
\]

\[
= h(z^*) \cdot \left[ \tilde{\epsilon} \left( z^* \frac{d\tau_1}{1-\tau_0} \right) - \tilde{\phi} \left( \frac{1}{d\tau_1} \right) \right],
\]

where we have used the assumption of constant \( \tilde{h}(\cdot) \) and the approximations for \( \Delta z^* \) and \( \tilde{z} \) in Section 2.7. Here \( h(z^*) = \int \int \tilde{h}(z^*, \epsilon, \phi) d\epsilon d\phi \), and \( \tilde{\epsilon} \) and \( \tilde{\phi} \) are the average elasticity and adjustment cost, respectively.

**Linear Adjustment Costs**

We now introduce an adjustment cost that increases linearly in the size of the adjustment. Assume that given an initial level of earnings \( z_0 \), agents must pay a cost of \( \phi^* \cdot |z - z_0| \)
when they change their earnings to a new level \( z \). Utility \( \tilde{u} \) at the new earnings level can be represented as:

\[
\tilde{u}(c, z; n, z_0) = u(c, z; n) - \phi^* \cdot |z - z_0| .
\]

The first order condition for earnings can be characterized as:

\[
- \frac{u_z(c, z; n)}{u_c(c, z; n)} = (1 - \tau - \phi^*/\lambda^* \cdot \text{sgn}(z - z_0))
\]

\[
= \begin{cases} 
  (1 - \tau - \phi) & \text{if } z > z_0 \\
  (1 - \tau + \phi) & \text{if } z < z_0 
\end{cases},
\]

where \( \lambda^* = u_c(c^*, z^*; n) \) is the Lagrange multiplier and \( \phi = \phi^*/\lambda^* \) is the dollar equivalent of the linear adjustment cost \( \phi^* \).

The individual chooses earnings as if he faces an effective marginal tax rate of \( \tilde{\tau} = \tau + \phi \cdot \text{sgn}(z - z_0) \). It follows that our predictions about earnings adjustment are similar to our previous predictions, except that the effective marginal tax rate \( \tilde{\tau} \) appears, rather than \( \tau \).

Thus, we can solve for the elasticity of earnings as a function of the change in earnings \( \Delta z^* \) due to introduction of a kink in the tax schedule and the jump in marginal tax rate \( d\tau_1 \):

\[
\varepsilon = \frac{\Delta z^*/z^*}{d\tilde{\tau}_1/(1 - \tilde{\tau}_0)}
\]

\[
= \frac{\Delta z^*/z^*}{(d\tau_1 - 2\phi)/(1 - \tau_0 - \phi)}.
\]

Since the right-hand side is increasing in \( \phi \), the estimate of the elasticity increases as the linear adjustment cost increases. This makes intuitive sense: the adjustment cost attenuates bunching, so holding constant the level of bunching, the elasticity must be higher as the adjustment cost increases.

Now assume that when an individual adjusts his earnings, he incurs a linear adjustment cost \( \phi^{*L} \) for every unit of change in earnings, as well as a fixed cost \( \phi^{*F} \) associated with any change in earnings. Consider again bunching at \( z^* \), with a tax rate jump of \( d\tau_1 = \tau_1 - \tau_0 \).
at earnings level $z^*$. We have the following set of expressions for excess mass:

$$B = \int_{z}^{z^*+\Delta z^*} h(\zeta) \, d\zeta$$

$$\varepsilon = \frac{\Delta z^*/z^*}{(d\tau_1 - 2\phi^L)/(1 - \tau_0 - \phi^L)}$$

$$\phi^F + \phi^L \cdot (\bar{z} - z^*) = u\left((1 - \tau_1)z^* + R', z^*; n\right) - u\left((1 - \tau_1)\bar{z} + R'; \bar{z}; n\right).$$

Using a left rectangle approximation for the integral, we have:

$$b \equiv \frac{B}{h(z^*)}$$

$$= z^* + \Delta z^* - \bar{z}$$

$$= z^* \left(\frac{d\tau_1 - 2\phi^L}{1 - \tau_0 - \phi^L\varepsilon} + 1\right) - \bar{z}.$$  

We can further apply an approximation for $\bar{z}$ similar to the approximation we used in Section 2.7, i.e. $\bar{z} = z^* + \phi^F/\left(d\tau_1 - 2\phi^L\right)$. Thus, the expression for bunching can be simplified to:

$$b = \varepsilon \left(\frac{z^*}{d\tau_1 - 2\phi^L} - \frac{\phi^F}{(d\tau_1 - \phi^L)}\right),$$

where $(\phi^F, \phi^L) = (\phi^*F/\lambda^*, \phi^*L/\lambda^*)$. In this case, we need at least three kinks to separately identify $(\varepsilon, \phi^F, \phi^L)$. Because we do not examine a setting in which one can compare bunching under three different positive tax rates, we are not able to estimate these parameters using data (or to fruitfully estimate the parameters in the non-linear case since we do not have a credible source of variation to identify them).

### A.11.7. Derivation of Formula for Bunching with a Pre-Existing Kink (Sharp Change Approach)

The Comparative Static approach abstracts from a key feature of our empirical setting. In particular, the Comparative Static approach models the transition from a budget set with no kink to one with a kink. This approach facilitates our basic intuition and provides a
transparent bridge between our approach and existing bunching methods in the presence of a kink. However, in our context, we conduct analysis using data just before and just after the benefit reduction rate was decreased (in 1990 from 50 percent to 33.33 percent for 66-69 year olds; or from 33.33 percent to zero for this group in 2000; or from 33.33 percent for 69-year-olds to zero for 70-year-olds in 1990-1999). These changes involve moving from an initial state with a kink to a new state with a smaller kink. In a frictionless model, the distinction is immaterial. However, as we show, this matters in the presence of a fixed adjustment cost. In particular, when the kink becomes more muted, the change in bunching will be attenuated due to the fixed adjustment cost.

We will assume that in the initial state, bunching is characterized as in Remark (2). Let the initial kink, $K_1$, be characterized by a lower marginal tax rate, $\tau_0$, to the left of $z^*$, and a higher marginal tax rate, $\tau_1$, to the right of $z^*$. The initial level of bunching is:

$$B_1 = \int_{z_1}^{z^* + \Delta z_1^*} h(\zeta) \, d\zeta$$

Now, consider a change in the kink to $K_2$, which retains the lower marginal tax rate $\tau_0$ to the left of $z^*$ but reduces the marginal tax rate to the right of $z^*$ to $\tau_2 < \tau_1$. Had we begun with no kink and introduced $K_2$, bunching would have been:

$$B_2 = \int_{z_2}^{z^* + \Delta z_2^*} h(\zeta) \, d\zeta$$

Note that relative to $K_1$, $K_2$ provides a weaker incentive to bunch, when starting from a baseline tax schedule with no kink. Formally, we have $z_2 \geq z_1$, $\Delta z_2^* < \Delta z_1^*$ and $B_2 \leq B_1$.

**Characterizing Bunching**

In characterizing bunching when moving from $K_1$ to $K_2$, individuals may be separated into several groups based on their optimal level of earnings $z_0$ in the absence of a kink. First, there are individuals with $z_0 < z^*$. They will locate to the left of the kink under both $K_1$
and $K_2$.

Second, we have individuals with $z^* < z_0 \leq z_1$ (area $i$ in Figure 9). These individuals would optimize in the presence of $K_1$ by moving to $z^*$, were it not for the adjustment cost. Now, with a smaller kink $K_2$, these individuals continue to remain at the initial earnings level $z_0 > z^*$, as the utility gain to reoptimizing to $z^*$ is even smaller than it was under $K_1$.

Third, we have those with $z_1 < z_0 \leq z_2$ (area $ii$ in Figure 9). When moving from no kink to $K_1$, these individuals locate at the kink, $z^*$. If the budget set had hypothetically transitioned from no kink to $K_2$, these individuals would have chosen to remain at $z_0$, due to the fixed adjustment cost. However, when moving from $K_1$ to $K_2$, these agents remain at the kink $z^*$. The reason is that the frictionless optimum under $K_2$ is $z^*$ for everyone initially earning in the range $[z^*, z^* + \Delta z_2]$.

Fourth, we have agents with $z_2 < z_0 \leq z^* + \Delta z_2$ (area $iii$ in Figure 9). These individuals bunch at $z^*$ when moving from no kink to either $K_1$ or $K_2$. Thus, they remain bunching at $z^*$ when moving from $K_1$ to $K_2$.

Fifth, we have agents with $z^* + \Delta z_2 < z_0 \leq z^* + \Delta z_1$ (areas $iv$ and $v$ in Figure 9). When starting from a budget set with no kink, these agents bunch under $K_1$, but not under $K_2$. Starting instead from $K_1$, they must choose between remaining at the kink $z^*$ or moving to the frictionless optimum under $K_2$, $z_2 > z^*$. We know that at least some of these individuals will remain bunching. To see this, consider an individual with earnings under no kink $z_0 = z^* + \Delta z_2 + \delta_0$. For small enough $\delta_0$ optimal earnings under $K_1$ is $z^*$, and optimal earnings under $K_2$ tends to $z^*$ as $\delta_0$ tends to zero. Likewise, the net utility gain from relocating from $z^*$ to $z_2$ under $K_2$ tends to zero as $\delta_0$ tends to zero. However, the fixed adjustment cost remains strictly positive. Therefore, this individual will remain at $z^*$ when moving from $K_1$ to $K_2$ for small enough $\delta_0$. In Figure 9, area $iv$ shows those with initial earnings $z^* + \Delta z_2 < z_0 < \bar{z}_0$ who remain bunching at the kink when transitioning from $K_1$ to $K_2$. Area $v$ shows those with initial earnings $\bar{z}_0 < z_0 < z^* + \Delta z_1$, who "debunch" from the kink.
when moving from $K_1$ to $K_2$.

When reoptimizing is beneficial for at least some agents in this final group, we will have a reduction in bunching when transitioning from $K_1$ to $K_2$.\textsuperscript{17} Empirically, we observe such a reduction over time, so this is the case relevant to our setting. In this case, the marginal “de-buncher” will be defined by the following conditions:

\[
\begin{align*}
\frac{-u_z(c_2, \bar{z}_2; \bar{n}_2)}{u_c(c_2, \bar{z}_2; \bar{n}_2)} &= (1 - \tau_2) \\
- \frac{u_z(c_0, \bar{z}_0; \bar{n}_2)}{u_c(c_0, \bar{z}_0; \bar{n}_2)} &= (1 - \tau_0) \\
\bar{z}_0 &\leq z^* + \Delta z^*_1
\end{align*}
\]

In words, the first line indicates that $\bar{z}_2 > z^*$ is the optimal, frictionless level of earnings chosen by the top buncher in the presence of $K_2$. The second line requires that when facing $K_2$, this agent is indifferent between remaining at $z^*$ and moving to $\bar{z}_2$ through paying the adjustment cost. The third line defines $\bar{z}_0$ as the initial level of earnings that this individual chooses when facing a constant marginal tax rate of $\tau_0$ and no kink. The fourth line requires that this individual is initially bunching at $z^*$ in response to $K_1$. If this last inequality is binding, then when moving from $K_1$ to $K_2$, none of the bunchers “debunch” and the fraction bunching is unchanged. In that case, we have no variation available to identify both $\varepsilon$ and $\phi$. As noted above, empirically we do observe that excess bunching falls around 1990 when the BRR falls from 50 percent to 33.33 percent (as well as in the other cases we examine empirically, in which the BRR falls from a positive level to zero). Thus, we restrict attention to the case in which $\bar{z}_0 < z^* + \Delta z^*_1$.

\textsuperscript{17}In certain cases, it is possible that reoptimizing away from the kink is not optimal for anyone in the initial earnings range $[z^* + \Delta z_2, z^* + \Delta z_1]$. In that case, there is no change in bunching when moving from $K_1$ to $K_2$. 

187
Bunching at $K_2$ following $K_1$, when $\bar{z}_0 < z^* + \Delta z_1$, can therefore be expressed as:

$$\tilde{B}_2 = \int_{\bar{z}_1}^{\bar{z}_0} h(\zeta) d\zeta$$

We can again solve this system of equations for $\phi^*$ and $\varepsilon$. Note that $\varepsilon$ is still identified by the potential adjustment of the top-most buncher:

$$\varepsilon = \bar{z}_0 - \bar{z}_2 \left(1 - \tau_0\right)$$

Note that when moving from $K_1$ to $K_2$, the change in bunching is smaller than it would be if we had started with steady state bunching at $K_1$ following no kink ($B_1$) and then moved to steady state bunching $K_2$ following no kink ($B_2$). That is:

$$B_1 - \tilde{B}_2 = \int_{\bar{z}_1}^{z^* + \Delta z_1^*} h(\zeta) d\zeta - \int_{\bar{z}_1}^{\bar{z}_0} h(\zeta) d\zeta$$

$$\leq \int_{\bar{z}_1}^{z^* + \Delta z_1^*} h(\zeta) d\zeta - \int_{\bar{z}_2}^{z^* + \Delta z_2^*} h(\zeta) d\zeta$$

$$= B_1 - B_2,$$

where the second line follows from the fact that $\bar{z}_0 \geq z^* + \Delta z_2^*$ and $\bar{z}_1 \leq \bar{z}_2$.

**Simplified Approximation**

We can again build intuition for this result by simplifying the formula for bunching in the second period. Assuming the density is constant over the range $[z^*, z^* + \Delta z_1^*]$, we have:

$$\tilde{b}_2 = \bar{z}_0 - \bar{z}_1$$

$$= (\bar{z}_0 - \bar{z}_2) + \bar{z}_2 - \bar{z}_1$$

$$= \varepsilon \left(\bar{z}_2 \frac{d\tau_2}{1 - \tau_0}\right) + \bar{z}_2 - \bar{z}_1$$

$$= \varepsilon \left(\bar{z}_2 \frac{d\tau_2}{1 - \tau_0}\right) + (\bar{z}_2 - z^*) - \frac{\phi}{d\tau_1}$$
where $\tilde{b}_2 \equiv \tilde{B}_2/h(z^*)$. On the third line, we have used the definition of the elasticity and on the fourth line, we used a first-order approximation to solve for $\tilde{z}_1$ as before in Section (A.11.6). Thus, we can see why bunching when moving from a larger to smaller kink is greater than would be predicted by the Comparative Static method. First, the term multiplying $\varepsilon$ has a $\tilde{z}_2$ instead of a $z^*$ and there is an additional term $\tilde{z}_2 - z^*$ – both of which increase bunching, since $\tilde{z}_2 > z^*$. Both of these capture of the excess bunching from above, due to inertia. Finally, the third term has a $d\tau_1$ in the denominator instead of a $d\tau_2$. The larger denominator increases bunching – $d\tau_1 > d\tau_2$ – and captures the fact that there is less attenuation in bunching from below, also due to inertia.

**Elasticities Under Frictionless (Saez 2010) Formula**

We investigate the results when applying the Saez (2010) formula for estimating elasticities—applicable to a frictionless setting—in a setting in which there are in fact adjustment costs. In other words, we answer the question: if there are adjustment costs and we mis-specify our estimate of the elasticity by assuming that we are in a frictionless setting, in what way do we mis-estimate the elasticity? As we show, if we face an adjustment cost as in the Sharp Change model, but we estimate the elasticity using the Saez (2010) formula applicable to a frictionless setting, we will see the elasticity estimate increase when we move from a larger kink to a smaller kink (as in our empirical application, and as we observe empirically in Figure 34). Thus, an increase in the Saez (2010) estimate of the elasticity—of the sort that we observe empirically—is a telltale sign that we are operating under a Sharp Change model.

We first present the formula for the elasticity in a frictionless model, as in Saez (2010). Next, we present the formulas for the elasticities we would estimate if we mis-specified the model as the frictionless (Saez 2010) model, even though we in fact face the Sharp Change model.
Saez (2010) model  We assume that tax changes are relatively small and that we can therefore treat the density is constant. (We derive analogous results if we instead use exact formulas under quasilinearity.) Assume that we begin with a more pronounced kink $K_1$ and then move to a less pronounced kink $K_2$, by lowering the jump in marginal tax rates at exempt amount from $d\tau_1$ to $d\tau_2$. Assume in each year, we can estimate normalized bunching: $b \equiv B/h(z^*)$. In a frictionless (Saez 2010) model, we have:

\[
b_1 = \Delta z_1^* = \varepsilon z^* \frac{d\tau_1}{1 - \tau_0} \\
b_2 = \Delta z_2^* = \varepsilon z^* \frac{d\tau_2}{1 - \tau_0}
\]

where we have used the fact that: $\Delta z^* = \varepsilon z^* d\tau / (1 - \tau_0)$. A natural estimator of the elasticity is the Saez estimator $e^S$:

\[
e^S = \frac{b}{z^*} \frac{(1 - \tau_0)}{d\tau} = \frac{b}{a \cdot d\tau}
\]

where $a \equiv z^* / (1 - \tau_0)$.

In each period (denoted by the subscript), we have the following for the Saez estimator when there are no frictions:

\[
e_1^S = \frac{b_1}{a \cdot d\tau_1} = \varepsilon \\
e_2^S = \frac{b_2}{a \cdot d\tau_2} = \varepsilon
\]

Thus, $e_1^S = e_2^S$. Here $e_1^S$ denotes the Saez (2010) estimate of the elasticity in period 1 (under $K_1$), and $e_2^S$ denotes this elasticity in period 2 (under $K_2$).

Sharp Change Model  By contrast, in the Sharp Change model, we start at kink $K_1$ and then move straight to $K_2$, once again estimating normalized bunching. We have the
following results here:

\[ b_1 = \Delta z_1^* + z^* - \bar{z}_1 \]
\[ b_2 = \bar{z}_0 - \bar{z}_1 \]

Rewrite \( b_1 \) and \( b_2 \) as follows:

\[ b_1 = \Delta z_1^* - (\bar{z}_1 - z^*) \]
\[ b_2 = \Delta z_2^* + (\bar{z}_0 - \Delta z_2^* - \bar{z}_1) \]

The Saez estimators now return:

\[ e_{S1}^S = \frac{b_1}{a \cdot d\tau_1} = \frac{\varepsilon - (\bar{z}_1 - z^*)}{a \cdot d\tau_1} \]
\[ e_{S2}^S = \frac{b_2}{a \cdot d\tau_2} = \varepsilon + \frac{(\bar{z}_0 - \Delta z_2^* - \bar{z}_1)}{a \cdot d\tau_2} \]

The change in Saez estimators is:

\[ e_{S2}^S - e_{S1}^S = \frac{(\bar{z}_0 - \Delta z_2^* - \bar{z}_1)}{a \cdot d\tau_2} + \frac{(\bar{z}_1 - z^*)}{a \cdot d\tau_1} \]
\[ \geq \frac{(\bar{z}_0 - \Delta z_2^* - \bar{z}_1)}{a \cdot d\tau_1} + \frac{(\bar{z}_1 - z^*)}{a \cdot d\tau_1} \]
\[ = \frac{\bar{z}_0 - \Delta z_2^* - \bar{z}_1 + \bar{z}_1 - z^*}{a \cdot d\tau_1} \]
\[ = \frac{\bar{z}_0 - \Delta z_2^* - z^*}{a \cdot d\tau_1} \]
\[ \geq \frac{\Delta z_2^* + z^* - \Delta z_2^* - z^*}{a \cdot d\tau_1} \]
\[ = 0 \]

where in the second line, we use the fact that \( d\tau_1 > d\tau_2 \) and in the second-to-last line, we use the fact that \( \bar{z}_0 \geq \Delta z_2^* + z^* \).

Thus, \( e_{S2}^S - e_{S1}^S \) is weakly greater than zero: the Saez (2010) frictionless elasticity estimate
weakly increases (as we observe empirically). The pattern we observe empirically—an upward jump in the Saez (2010) estimate of the elasticity when the policy change occurs in 1990, as shown in Figure 34—is a telltale sign that we are operating in the Sharp Change model.

**Sharp Change Approach with Linear Adjustment Cost**

Note that under the Sharp Change approach, with a linear adjustment cost we can derive an approximation for bunching (analogously to Section A.11.6) to show that:

$$\tilde{b}_2 = \varepsilon \left( \frac{d\tau_2 - 2\phi^L}{1 - \tau_0 - \phi^L} \right) - (z_2 - z^*) - \frac{\phi^F}{d\tau_1 - \phi^L}.$$

In our empirical application, we apply the Sharp Change approach in estimating elasticities and adjustment costs using data on individuals in different years (in the baseline specification, 1989 and 1990). Our empirical approach is applicable in the case in which individuals make year-by-year static earnings decisions; in the baseline, this effectively assumes that individuals weigh the cost of adjustment against the benefits in 1990. If instead individuals compare the costs of adjustment in 1990 to the benefits of adjustment in 1990 and subsequent years, then the benefits and therefore the estimated cost of adjustment would likely be larger. In this case, our estimated cost of adjustment could be considered a lower bound. Our estimates demonstrate the applicability of the methodology, including in settings in which the benefits may be realized over more years (as these discounted benefits would then be weighed against the costs). As we discuss, we view our static approach as a natural first step toward estimating elasticities and adjustment costs, but developing a dynamic model of adjustment frictions represents an important next step.

**A.11.8. Estimating the Elasticity and Adjustment Cost**

In this section, we describe in more detail how we use data on the amount of bunching to estimate the elasticity and adjustment cost. Let $b = (b_1, b_2, \ldots, b_K)$ be a vector of
(estimated) bunching amounts normalized by the density at the kink, using the method described in Section (2.3). Let $\tau = (\tau_1, \ldots, \tau_K)$ be the tax schedule at each kink. The triplet $\tau_k = (\tau_{k0}, \tau_{k1}, \tau_k^2)$ denotes the tax rate below ($\tau_{k0}$) and above ($\tau_{k1}$) the kink $k$; when using the Sharp Change method, $\tau_{k1}^2$ denotes the ex post marginal tax rate above the kink after it has been reduced, as in Section (A.11.7). Let $z^* = (z_1^*, \ldots, z_K^*)$ be the earnings levels associated with each kink. To estimate $(\epsilon, \phi)$, we seek the values of the parameters that make predicted bunching $\hat{b}$ and actual (estimated) bunching $b$ as close as possible on average.

Letting $\hat{b}(\epsilon, \phi) = (\hat{b}(\tau_1, z_1^*, \epsilon, \phi), \ldots, \hat{b}(\tau_K, z_K^*, \epsilon, \phi))$, our estimator is:

$$\left(\hat{\epsilon}, \hat{\phi}\right) = \arg \min_{(\epsilon, \phi)} \left(\hat{b}(\epsilon, \phi) - b\right) W \left(\hat{b}(\epsilon, \phi) - b\right),$$

(A.7)

where $W$ is a $K \times K$ diagonal matrix whose diagonal entries are the inverse of the variances of the estimates of the $b_k$.

We obtain our estimates by minimizing equation (A.7) numerically. Solving this problem requires evaluating $\hat{b}$ at each trial guess of $(\epsilon, \phi)$.

Recall that in general bunching takes the form:

$$B_k(\tau_k, z_k^*; \epsilon, \phi^*) = \int_{z_k^{lb}}^{z_k^{ub}} h(\zeta) d\zeta,$$

where $(z_k^{lb}, z_k^{ub})$ are the ex-ante earnings levels of the lowest and highest earning bunchers, in the presence of linear tax at the lower tax rate, $\tau_{k0}$. Define $z_k^* + \Delta z_k^1$ as the ex ante earnings level for the highest earning buncher – in the absence of frictions – when the size of the kink is $d\tau_1^k = \tau_{k1}^1 - \tau_{k0}^1$. As in the main text, we continue to assume that $h(\cdot)$ is uniform in $[z_k^*, z^* + \Delta z_1^k]$, so that

$$b_k(\tau_k, z_k^*; \epsilon, \phi^*) = z_k^{ub} - z_k^{lb},$$

where $b = B/h(z_k^*)$. The definitions of $(z_k^{lb}, z_k^{ub})$ vary depending on the setting and are

---

In solving problem (A.7), we impose that $\phi \geq 0$. When $\phi < 0$, every individual adjusts her earnings by at least some arbitrarily small amount, regardless of the size of $\phi$. This implies that $\phi$ is not identified if it is less than zero.
defined as follows. In the frictionless case (Saez 2010), we have:

\[
\begin{align*}
z_k^{lb} &= z_k^* \\
z_k^{ub} &= z_k^* + \Delta z_k^1.
\end{align*}
\]

In the presence of a fixed adjustment cost, we have under the Comparative Static approach:

\[
\begin{align*}
z_k^{lb} &= \hat{z}_1^k \\
z_k^{ub} &= z_k^* + \Delta z_k^1,
\end{align*}
\]

where \( \hat{z}_1^k \) is the ex-ante earnings of the marginal buncher from below given a fixed cost of adjustment. This is defined in the indifference condition above in equation (A.3). Finally, under the Sharp Change method, we have:

\[
\begin{align*}
z_k^{lb} &= \hat{z}_1^k \\
z_k^{ub} &= \bar{z}_0^k,
\end{align*}
\]

where \( \hat{z}_1^k \) is similarly the ex ante earnings of the marginal buncher from below (calculated using a kink with \( d\tau_k^1 = \tau_k^1 - \tau_k^0 \)). The ex ante earnings of the marginal buncher from above, \( \bar{z}_0^k \), is defined in Section (A.11.7) where \( d\tau_k^1 = \tau_k^1 - \tau_k^0 \) and \( d\tau_k^2 = \tau_k^2 - \tau_k^0 \).

Our estimator assumes a quasilinear utility function, \( u(c, z; n) = c - \frac{n}{1+1/\varepsilon} \left( \frac{z}{n} \right)^{1+1/\varepsilon} \), following Saez (2010), Chetty et al. (2011) and Kleven and Waseem (forthcoming). (In order to relax this assumption empirically, we would have to observe wealth, which is not available in the data.) Note that because we have assumed quasilinearity, \( \phi^* = \phi \), \( \Delta z_k^1 = z_k^* \left( \left( \frac{1-\tau_k^1}{1-\tau_k^0} \right)^\varepsilon - 1 \right) \) and \( n = z(\tau) / (1 - \tau)^\varepsilon \), where \( z(\tau) \) are the optimal, interior earnings under a linear tax of \( \tau \). However, there typically is not a closed form solution for the \( (z_k^{lb}, z_k^{ub}) \) in other cases. Instead, given \( \varepsilon \) and \( \phi \), we find \( (z_k^{lb}, z_k^{ub}) \) numerically as the
solution to relevant indifference condition. For example, $z^k_1$ is defined implicitly by:

$$u((1 - \tau^k_1)z^*_k + R^k_1, z^*_k; z^1_1/(1 - \tau^k_0)) - u((1 - \tau^k_1)\xi^k_1 + R^k_1, \xi^k_1; \xi^1_1/(1 - \tau^k_0)) = \phi,$$

The equation is continuously differentiable and has a unique solution for $z^k_1$. As such, Newton-type solvers are able to find $z^k_1$ accurately. Note that some combinations of $\tau_k, z^*_k, \varepsilon$, and $\phi$ imply $z^l_k > z^u_k$. In this case, the lowest-earning adjuster does not adjust to the kink, and whenever this happens we set $\hat{b}_k = 0$. The predicted amount of bunching is therefore:

$$\hat{b}_k(\tau_k, z^*_k; \varepsilon, \phi) = \max(z^u_k - z^l_k, 0).$$

We have also shown a robustness check in our Tables in which we assume that the earnings distribution is lognormal, rather than assuming that $h(\cdot)$ is uniform in $[z^*_k, z^*_k + \Delta z^k_1]$. Specifically, we use the distribution of earnings at age 61 over 1986-1988 and 1992-1994 to estimate the parameters of a lognormal earnings distribution, $(\mu_z, \sigma_z)$, using maximum likelihood. (Individuals age 61 are not subject to the AET but are not far removed in age from those at retirement age, making this a reasonable counterfactual earnings density for those subject to the AET). We then solve for bunching using:

$$B_k(\tau_k, z^*_k; \varepsilon, \phi^k) = \Phi \left( \frac{\log z^l_k - \mu_z}{\sigma_z} \right) - \Phi \left( \frac{\log z^u_k - \mu_z}{\sigma_z} \right)$$

where $\Phi (\cdot)$ is the standard normal CDF.

We estimate bootstrapped standard errors. Observe that the estimated vector of parameters $(\hat{\varepsilon}, \hat{\phi})$ is a function of the estimated amount of bunching; call this function $\theta(b)$. To compute bootstrapped standard errors, we use the bootstrap procedure of Chetty et al. (2011) to obtain 200 bootstrap samples of $b$. For each bootstrap sample, we compute $\hat{\varepsilon}$ and $\hat{\phi}$ as the solution to (A.7). The standard deviation of $\hat{\varepsilon}$ and $\hat{\phi}$ across bootstrap samples is the bootstrap standard error, and we compute confidence intervals analogously. We estimate
whether an estimate is significantly different from zero by assessing how frequently the constraint $\phi \geq 0$ binds in our estimation. Given this constraint, p-values are from a one-sided test of equality with zero. We have also estimated the standard errors using the delta method and obtained similar results.

A.12. Appendix: Additional Figures
Note: When we move from a linear budget constraint (Panel A) to a convex kink (Panel B), individuals with initial earnings between $z^*$ and $z^* + \Delta z^*$ relocate to the kink. As we move from a linear budget constraint (Panel C) to a convex kink (Panel D), a spike in the earnings density appears at the kink, corresponding to the density that was initially located between $z^*$ and $z^* + \Delta z^*$. The spike is spread out in the vicinity of the kink in Panel D; this may result from several factors discussed in Saez (2010), such as inability to control earnings precisely.
Note: Panel A decomposes the ex-post earnings distribution shown in Appendix Figure 14 Panel D into two groups. The bunchers, group $X$, are those who bunch at the kink in the presence of the higher marginal tax rate $\tau + d\tau$ but not at the lower marginal tax rate $\tau$. The non-bunchers, group $Y$, are comprised of those who locate to the left of the kink under the initial lower marginal tax rate $\tau$, and those who locate to the right of the kink under the higher marginal tax rate $\tau + d\tau$. Panel B demonstrates how the distribution of earnings in the absence of the kink is estimated to recover the share of bunchers, by excluding data in a neighborhood of $z^*$.
Figure 16: Mean Percentage Change in Earnings from Age 70 to 71, by Earnings at 70, 1990-1998

Note: The figure shows the mean percentage change in earnings from age 70 to age 71 (y-axis), against earnings at age 70 (x-axis). Earnings are measured relative to the kink, shown at zero on the x-axis. The data are a 20 percent random sample of 70-year-olds in the LEHD in 1990-1998. We exclude 1999 as a base year in this and similar graphs because the AET is eliminated for those over NRA in 2000. Higher earnings growth far below the kink reflects mean reversion visible in this part of the earnings distribution at all ages. We also find a spike in mean earnings growth from age 70 to age 71 among those born in January.
Figure 17: Adjustment Across Ages: Histograms of Earnings and Normalized Excess Mass, 59-73-year-old OASI Claimants, 1990-1999

Panel A: Earnings histograms, by age

Panel B: Normalized excess mass, by age

See notes to Figures 5 and 6. This figure differs from Figures 5 and 6 only because the sample in year $t$ consists only of people who have claimed OASI in year $t$ or before (whereas in Figures 5 and 6 it consists of all those who claimed by age 65).
Figure 18: Adjustment Across Ages: Histograms of Earnings and Normalized Excess Mass, 59-73-year-olds Claiming OASI by Age 65, 1972-1982

Panel A: Earnings histograms, by age

Panel B: Normalized excess mass, by age

See notes to Figures 5 and 6. This figure differs from Figures 5 and 6 only because the years examined are 1972-1982 (whereas in Figures 5 and 6 the years examined are 1990-1999).
Figure 19: Adjustment Across Ages: Histograms of Earnings and Normalized Excess Mass, 59-73-year-olds Claiming OASI by Age 65, 1983-1989

Panel A: Earnings histograms, by age

Panel B: Normalized excess mass, by age

See notes to Figures 5 and 6. This figure differs from Figures 5 and 6 only because the years examined are 1983-1989 (whereas in Figures 5 and 6 the years examined are 1990-1999).
Figure 20: Adjustment Across Ages: Histograms of Earnings and Normalized Excess Mass, 59-73-year-olds Claiming OASI by Age 65, 2000-2006

Panel A: Earnings histograms, by age

Panel B: Normalized excess mass, by age

See notes to Figures 5 and 6. This figure differs from Figures 5 and 6 only because the years examined are 2000-2006 (whereas in Figures 5 and 6 the years examined are 1990-1999). As explained in the main text, the NRA slowly rose from 65 for cohorts that reached age 62 during this period; the results are extremely similar when the sample is restricted to those who claimed by 66, instead of 65. In the year of attaining NRA, the AET applies for months prior to such attainment.
Figure 21: Comparison of Normalized Excess Bunching Among 62-64 Year-Olds and 66-69 Year-Olds, 1982-2004

Note: the figure shows excess normalized bunching among 62-64 year-olds and 66-69 year-olds in each year from 1982 to 2004. Note the caveat that the 62-64 year-old group faces a notch at the exempt amount, as opposed to the kink faced by those 66-69.
Figure 22: Adjustment Across Ages: Histograms of Earnings, 66-69 Year-Olds, 1999-2001

Panel A: Earnings histogram, 66-69 year-olds, 1999

Panel B: Earnings histogram, 66-69 year-olds, 2000

Panel C: Earnings histogram, 66-69 year-olds, 2001

Note: the figure shows a histogram of earnings in 1999, 2000, and 2001, using LEHD data on 66-69 year-olds. Earnings are measured relative to the kink, shown at zero.
Figure 23: Adjustment Across Ages: Histograms of Earnings and Normalized Excess Mass, 59-73-year-olds Claiming OASI by Age 65, 1990-1999

Panel A: Earnings histograms, by age

Panel B: Normalized excess mass, by age

See notes to Figures 5 and 6. This figure differs from Figures 5 and 6 only because the bandwidth is $500 (whereas in Figures 5 and 6 it is $800).
Figure 24: Robustness to Polynomial Degree: Normalized Excess Mass by Age and Year, OASI Claimants by 65


Panel B: Normalized Excess Mass by Year, Ages 66-69

Notes: The figure shows the difference in estimates of normalized excess bunching as we vary the degree of the polynomial used. For additional notes on the samples see Figure 6 for Panel A and Figure 7 for Panel B.
Figure 25: Robustness to the Excluded Region: Normalized Excess Mass by Age and Year, OASI Claimants by 65

**Panel A: Normalized Excess Mass by Age, 1990-1999**

**Panel B: Normalized Excess Mass by Year, Ages 66-69**

Notes: The figure shows the difference in estimates of normalized excess bunching as we vary the region about the kink that is "dummied out" in the polynomial estimation. For additional notes on the samples see Figure 6 for Panel A and Figure 7 for Panel B.
Figure 26: Adjustment by Sex: Histograms of Earnings, 59-73-year-olds Claiming OASI by Age 65, 1990-1999

See notes to Figure 5. The sample examined is the same as in Figure 5 but examines men and women separately.
Figure 27: Adjustment Across Ages: Histograms of Earnings, 59-73-year-olds Claiming OASI by Age 65 with Self-Employment Income, 1990-1999

See notes to Figure 5. The figure differs from Figure 5 only because the sample consists of those with positive self-employment income (whereas in Figure 5 those with positive self-employment income are excluded).
Figure 28: Mean Percentage Change in Earnings from Age 69 to 70, by Earnings at 69, 1990-1998

Note: The figure shows the mean percentage change in earnings from age 69 to age 70 (y-axis), against earnings at age 69 (x-axis). Earnings are measured relative to the kink, shown at zero on the x-axis. The data are a 20 percent random sample of 69-year-olds in the LEHD in 1990-1998. We exclude 1999 as a base year in this and similar graphs because the AET is eliminated for those over NRA in 2000.
Figure 29: Probability that Earnings Move with Kink, 1990-1998

Note: The figure shows the probability that individual earnings move with the kink from year to year (i.e. the probability that an individual locates at the kink in year $t + 1$, conditional on locating at the kink in year $t$), for age groups 58 to 60, 62 to 63, and 65 to 68. (Each of these ages refers to age in year $t$.) Results are similar when considering similar age bins. The kink is defined as the region within $2800$ of the exempt amount. We exclude the year 1999 because the AET is eliminated for those over NRA in 2000. See other notes to Figure 5.
Figure 30: Earnings Distributions by Age, OASI Claimants by Age 65, 1990-1999

Panel A: Earnings Distributions by Age, 60-62

Panel B: Earnings Distributions by Age, 69-71

Notes: The figure shows earnings distributions at ages 60, 61, and 62 (Panel A) and at ages 69, 70, and 71 (Panel B).
Figure 31: Mortality Analysis: Mean Age at Death, 62-69-year-olds Claiming OASI by Age 65, 1966-1971 and 1990-1999

Note: The figure shows mean age at death from a one percent random sample of SSA administrative data on individuals aged 59-73, claiming OASI by age 65, between 1966 and 1971 (inclusive) in the top panel, and between 1990 and 1999 (inclusive) in the bottom panel. The figure shows no clearly noticeable patterns at the kink that are different from those away from the kink. This holds true for those 62-64 and 66-69, in a period prior to the introduction of the Delayed Retirement Credit (i.e. 1966-1971) and subsequent to its introduction (i.e. 1990-1999). Results are similar for other time periods.
Figure 32: Normalized Excess Mass at Kink by Age, 1966-1971

Note: The figure shows normalized excess bunching, ages 18 to 75, 1966-1971. We group ages into three-year bins. See other notes to Figure 6.
Figure 33: Fraction of Workers Changing Employers from Age $t$ to Age $t+1$, by Age $t$ Earnings, 1990-1998

Note: The figure shows the fraction of workers who change employers from age $t$ to age $t+1$ (y-axis), plotted against earnings at age 69 (x-axis). For example, 0.16 on the y-axis implies that 16 percent of workers change employers from age $t$ to age $t+1$. Earnings are measured relative to the kink, shown at zero on the x-axis. The data are a 20 percent random sample of the LEHD in 1990-1998. The bin width is $800. Solid (dotted) lines show point estimates (95 percent confidence intervals).
Note: The figure shows elasticities estimated using the Saez (2010) method, by year from 1982 to 1994, among 67-68 year-old OASI claimants. We use our methods for estimating normalized excess bunching but use Saez’ (2010) formula to calculate elasticities, under a constant density. This method yields the following formula:

\[ \varepsilon = \left[ \log \left( \frac{b}{g} + 1 \right) \right] \div \left[ \log \left( \frac{1 - \tau_0}{1 - \tau_1} \right) \right] \]
Figure 35: Probability of claiming OASI in year $t+1$ among 61-68 year-olds in year $t$ who are not claiming, 1990-1998

Note: The figure shows the probability that an individual claims OASI in year $t+1$, conditional on not claiming OASI in year $t$, for those ages 61-68 in year $t$ from 1990-1998.
Table 37: Estimates of Elasticity and Adjustment Cost Using Sharp Change Method and Disappearance of Kink at Age 70

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\varepsilon)</td>
<td>(\phi)</td>
<td>(\varepsilon \cdot \phi = 0, \text{Age 69})</td>
</tr>
<tr>
<td>Basic</td>
<td>0.28</td>
<td>89.93</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>[0.25, 0.34]***</td>
<td>[25.51, 240.00]***</td>
<td>[0.22, 0.30]***</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.30</td>
<td>87.19</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>[0.26, 0.36]***</td>
<td>[31.88, 256.39]***</td>
<td>[0.23, 0.32]***</td>
</tr>
<tr>
<td>Benefit Enhancement</td>
<td>0.42</td>
<td>58.01</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>[0.37, 0.50]***</td>
<td>[16.70, 146.86]***</td>
<td>[0.33, 0.45]***</td>
</tr>
<tr>
<td>Excluding FICA</td>
<td>0.36</td>
<td>83.01</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>[0.31, 0.43]***</td>
<td>[23.79, 213.17]***</td>
<td>[0.28, 0.38]***</td>
</tr>
<tr>
<td>Bandwidth = $500</td>
<td>0.20</td>
<td>17.05</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>[0.22, 0.32]***</td>
<td>[0.69, 76.00]***</td>
<td>[0.16, 0.23]***</td>
</tr>
<tr>
<td>68-70 year-olds</td>
<td>0.30</td>
<td>79.10</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>[0.27, 0.35]***</td>
<td>[24.61, 189.34]***</td>
<td>[0.25, 0.32]***</td>
</tr>
</tbody>
</table>

Note: The table estimates elasticities and adjustment costs using the removal of the AET at age 70, using data on 69-71 year-olds. We cannot estimate the constrained elasticity using only data on age 70 because the benefit reduction rate is zero at that age. We show these results in the Appendix, rather than the main text, because the estimates of excess bunching at age 70 are potentially affected by the coarse measure of age that we use, as explained above in the Appendix. To address this issue, we use both age 70 and age 71 in estimating these results. Using only age 70—or alternatively using only age 71—both show very similar results, which is unsurprising because Figure 6 shows that normalized excess bunching is similar at ages 70 and 71. The row labeled “68-70 year-olds” uses data from ages within this range. See also notes to Table 11.

A.13. Appendix: Additional Estimates of \(\varepsilon\) and \(\phi\)
Table 38: Estimates of Elasticity and Adjustment Cost Using Sharp Change Method and 1990 Policy Change, Assuming no Pre-Period Bunching Attenuation

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon$</td>
<td>$\phi$</td>
<td>$\varepsilon</td>
<td>\phi = 0$</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.22</td>
<td>$139.53$</td>
<td>0.39</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>[0.18, 0.26]***</td>
<td>[46.25, 359.40]***</td>
<td>[0.33, 0.48]***</td>
<td>[0.18, 0.26]***</td>
</tr>
<tr>
<td>Lognormal density</td>
<td>0.24</td>
<td>$149.87$</td>
<td>0.43</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>[0.19, 0.29]***</td>
<td>[42.55, 426.32]***</td>
<td>[0.35, 0.54]***</td>
<td>[0.19, 0.29]***</td>
</tr>
<tr>
<td>Benefit Enhancement</td>
<td>0.37</td>
<td>$76.24$</td>
<td>0.59</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>[0.30, 0.43]***</td>
<td>[17.28, 237.63]***</td>
<td>[0.49, 0.73]***</td>
<td>[0.30, 0.43]***</td>
</tr>
<tr>
<td>Excluding FICA</td>
<td>0.30</td>
<td>$119.84$</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>[0.25, 0.35]***</td>
<td>[32.33, 344.45]***</td>
<td>[0.42, 0.62]***</td>
<td>[0.25, 0.35]***</td>
</tr>
<tr>
<td>Bandwidth = $500</td>
<td>0.24</td>
<td>$83.80$</td>
<td>0.38</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>[0.19, 0.30]***</td>
<td>[5.86, 296.54]***</td>
<td>[0.30, 0.50]***</td>
<td>[0.19, 0.30]***</td>
</tr>
<tr>
<td>68-70 year-olds</td>
<td>0.31</td>
<td>$106.32$</td>
<td>0.50</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>[0.27, 0.34]***</td>
<td>[25.21, 291.94]***</td>
<td>[0.42, 0.62]***</td>
<td>[0.27, 0.34]***</td>
</tr>
</tbody>
</table>

Note: The table applies the Sharp Change method to the 1990 policy change, using data from 1989 and 1990, but assumes that bunching in 1989 is not attenuated by adjustment frictions. The constrained estimate of bunching using data only from 1989 is mechanically the same as the unconstrained estimate, as both rely on the Saez (2010) formula for bunching. See also notes to Appendix Table 37.
Table 39: Estimates of Elasticity and Adjustment Cost Using Sharp Change Method and Elimination of Earnings Test in 2000 for 66-69 Year-Olds

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon$</td>
<td>$\phi$</td>
<td>$\varepsilon</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.10</td>
<td>$23.23$</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>[0.03, 0.19]$^{***}$</td>
<td>[0.00, 331.52]$^*$</td>
<td>[0.04, 0.14]$^{***}$</td>
</tr>
<tr>
<td>Lognormal density</td>
<td>0.10</td>
<td>$20.17$</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>[0.03, 0.21]$^{***}$</td>
<td>[0.00, 346.90]$^*$</td>
<td>[0.04, 0.14]$^{***}$</td>
</tr>
<tr>
<td>Benefit Enhancement</td>
<td>0.15</td>
<td>$14.93$</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>[0.04, 0.36]$^{***}$</td>
<td>[0.00, 350.46]$^*$</td>
<td>[0.07, 0.21]$^{***}$</td>
</tr>
<tr>
<td>Excluding FICA</td>
<td>0.13</td>
<td>$21.31$</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>[0.03, 0.28]$^{***}$</td>
<td>[0.00, 402.69]$^*$</td>
<td>[0.06, 0.18]$^{***}$</td>
</tr>
<tr>
<td>Bandwidth = $500</td>
<td>0.10</td>
<td>$56.48$</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>[0.07, 0.21]$^{***}$</td>
<td>[0.03, 367.05]$^{**}$</td>
<td>[0.05, 0.13]$^{***}$</td>
</tr>
<tr>
<td>1998 to 2000</td>
<td>0.12</td>
<td>$23.09$</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>[0.08, 0.55]$^{***}$</td>
<td>[0.00, 957.13]$^*$</td>
<td>[0.07, 0]$^{***}$</td>
</tr>
</tbody>
</table>

Note: The table applies the Sharp Change method to the 2000 policy change, using data on 1999 and 2000. It is not possible to perform the constrained estimate using a cross-section from the year 2000 because the marginal tax rate is zero in this year, implying that the elasticity is undefined. The final row uses data from 1998 and 2000, rather than 1999 and 2000. Note that the left endpoint of the confidence interval is sometimes at 0.00 even though the p-value indicates significance at the 5 percent level; this is because a few of the bootstrap replications show point estimates that are positive and very small so round to 0.00. See also notes to Appendix Table 37.
Table 40: Estimates of Elasticity and Adjustment Cost Using Comparative Static Method and 1990 Policy Change

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ε</td>
<td>0.21</td>
<td>$0.00</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>φ</td>
<td>$0.00</td>
<td>0.00</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td>ε</td>
<td>φ = 0</td>
<td>1993</td>
<td>1989</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.18</td>
<td>[0.18, 0.34]***</td>
<td>0.14</td>
<td>[0.17, 0.26]***</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
<td>[0.00, 902.56]</td>
<td>0.31</td>
<td>[0.14, 0.31]***</td>
</tr>
<tr>
<td>Lognormal density</td>
<td>0.24</td>
<td>$33.35</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>[0.20, 0.36]***</td>
<td>[0.00, 577.01]</td>
<td>0.15</td>
<td>[0.19, 0.29]***</td>
</tr>
<tr>
<td>Benefit Enhancement</td>
<td>0.37</td>
<td>$78.09</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>[0.30, 0.50]***</td>
<td>[0.00, 352.03]</td>
<td>0.21</td>
<td>[0.29, 0.43]***</td>
</tr>
<tr>
<td>Excluding FICA</td>
<td>0.30</td>
<td>$58.54</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>[0.25, 0.42]***</td>
<td>[0.00, 492.47]</td>
<td>0.18</td>
<td>[0.24, 0.35]***</td>
</tr>
<tr>
<td>Bandwidth = $500</td>
<td>0.25</td>
<td>$0.00</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>[0.21, 0.80]***</td>
<td>[0.00, 4188.55]</td>
<td>0.20</td>
<td>[0.19, 0.32]***</td>
</tr>
<tr>
<td>1988 and 1993</td>
<td>0.26</td>
<td>$146.33</td>
<td>0.21</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>[0.23, 0.33]***</td>
<td>[0.00, 408.86]</td>
<td>0.14</td>
<td>[0.22, 0.28]***</td>
</tr>
<tr>
<td>Linear Approximation</td>
<td>0.46</td>
<td>$379.10</td>
<td>0.30</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>[0.58, 0.60]***</td>
<td>[351.27, 406.93]***</td>
<td>0.19</td>
<td>[0.31, 0.47]***</td>
</tr>
</tbody>
</table>

Note: The table applies the Comparative Static method to the 1990 policy change, using data from 1989 and 1993. "Linear approximation" refers to estimates based on the linearized formulas presented in the text. See also notes to Appendix Table 37.
BIBLIOGRAPHY


M. J. Dickstein. Patient vs. physician incentives in prescription drug choice. Unpublished


228


C. M. Turra and O. S. Mitchell. The impact of health status and out-of-pocket medical


