NBER WORKING PAPER SERIES

MANAGING INTELLIGENCE:
SKILLED EXPERTS AND AI IN MARKETS FOR COMPLEX PRODUCTS

Jonathan Gruber
Benjamin R. Handel
Samuel H. Kina
Jonathan T. Kolstad

Working Paper 27038
http://www.nber.org/papers/w27038

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020

We thank Picwell for providing the data for this study. Ned Augenblick and Filip Matejka provided excellent comments. We thank participants at MIT, Northwestern, University of Arizona, BU/Harvard/MIT health seminar, University of Rochester, National University of Singapore, Western Economic Assoc. Conference, University of North Carolina, Haas, BYU and the NBER workshops on Economics of AI and Machine Learning in Health Care Markets for feedback. All authors hold ownership stakes in Picwell. The findings represent our own views and all errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w27038.ack

Managing Intelligence: Skilled Experts and AI in Markets for Complex Products
Jonathan Gruber, Benjamin R. Handel, Samuel H. Kina, and Jonathan T. Kolstad
NBER Working Paper No. 27038
April 2020
JEL No. I13,J24,L15

ABSTRACT

In numerous high stakes markets skilled experts play a key role in facilitating consumer choice of complex products. New artificial intelligence (AI) technologies are increasingly being used to augment expert decisions. We study the role of technology and expertise in the market for health insurance, where consumer choices are widely known to be sub-optimal. Our analysis leverages the large-scale implementation of an AI-based decision support tool in a private Medicare exchange where consumers are randomized to skilled agents over time. We find that, prior to AI-based technology, skilled experts in this market exhibit the same type of inconsistent behavior found in previous studies of individual choices, costing consumers \$1260 on average. The addition of AI-based decision support improves outcomes by \$278 on average and substantially reduces heterogeneity in broker performance. Experts efficiently synthesize private information, incorporating AI-based recommendations along dimensions that are well suited to AI (e.g. total expected patient costs), but overruling AI-based recommendations along dimensions for which humans are better suited (e.g. specifics of doctor networks). As a result, switching plans, an ex-post measure of plan satisfaction, is meaningfully lower for agents making AI-based recommendations. While AI is a complement to skill on average, we find that it is a substitute across the skill distribution; lower quality agents provide better recommendations with AI than the top agents did without it. Overall productivity rises, with the introduction of decision support associated with a 21% reduction in call time for enrollment.

Jonathan Gruber
Department of Economics, E52-434
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
gruberj@mit.edu

Benjamin R. Handel
Department of Economics
University of California, Berkeley
508-1 Evans Hall #3880
Berkeley, CA 94720
and NBER
handel@berkeley.edu

Samuel H. Kina
Picwell Inc.
1900 Market St, Suite 120
Philadelphia, PA 19103
sam@picwell.com

Jonathan T. Kolstad
Haas School of Business
University of California, Berkeley
Berkeley, CA 94720
and NBER
jkolstad@berkeley.edu

1 Introduction

Skilled experts play a key role in assisting consumers in decision making in a variety of markets, from doctors advising patients to financial planners assisting investors, and beyond. The centrality of expertise in market function has long attracted the attention of economists. Since at least Arrow (1963), economists have focused on the potential role of skilled experts in improving choice quality and market function. The rise of artificial intelligence (AI) along with increasingly rich data offers an alternative to human expertise.[1] AI is distinguished by the potential to perform tasks typically reserved for human expertise as well as to exceed human performance on some domains involving complex computation.

Despite this promise, current incarnations of AI tools generally perform very well only on a subset of tasks required of a human expert. Thus, AI has generally been brought to market alongside human experts.[2] Understanding AI, therefore, requires not simply asking the aggregate questions of substitution between humans and machines but also of the detailed interaction between humans and AI tools where different aspects of the production function can be either augmented or replaced by technology. Whether and how technology complements or substitutes for skilled expertise will shape both product markets where expert recommenders play a role and the associated labor markets for that expertise (Acemoglu and Restrepo (2019); Athey et al. (2020)).

We study the role of AI in a specific setting — the market for health insurance — that lends itself to asking general questions about the interaction of AI and expertise in determining market outcomes. Using rich administrative data, randomization of clients to enrollment agents, and temporal variation in AI availability, we are able to recover the underlying parameters governing expert recommendations without AI as well as when AI is completely embedded in the choice process.

Beyond the general questions of AI-human interaction, market outcomes and welfare in health insurance markets are of intrinsic interest. Many health care systems and policies rely on the private provision of insurance to cover consumer health risks, from Medicare Advantage, which offers private plans that seniors can purchase in lieu of enrolling in original Medicare, to the private insurance options offered on state exchanges under the Affordable Care Act. In principal, competing private plans discipline the market, driving down premiums conditional on product quality and allowing the most innovative and efficient insurance products to succeed. Whether or not this occurs, however, depends on consumers making sound, well-informed choices in a competitive environment.

There is now a substantial body of evidence showing that, in practice, consumers face difficulties in making choices in insurance markets (see Chandra et al. (2018) and Handel and Kolstad

[1] We will use the term Artificial Intelligence broadly and interchangeably with machine learning. There are important differences but they depend on myriad definitions of each. We view the semantic distinction as beyond the scope of our paper and instead use the term AI which has been widely adopted by firms and policy makers considering the degree to which prediction can facilitate decision making (see e.g. Agrawal et al. (2018)). Our setting, if anything, extends beyond machine learning alone insofar as the tool incorporates individual preferences and seeks to mimic the behavior of an informed individual, at least up to the specification of utility.

[2] AI is likely to be implemented alongside humans for the longer run in settings in which prediction is not amenable to portions of the task or the state space of the problem is too large. Even in settings in which AI could eventually replace experts entirely we expect to see such tools alongside humans as data is collected to inform the AI system.

(2015a) for overviews of this literature). Consumers leave large amounts of money on the table in their insurance choices, due to a combination of factors including search costs, switching costs, insurance literacy, inattention, and limited information. Structural models show systematic choice inconsistencies in consumer decisions over Medicare Part D prescription drug plans (Abaluck and Gruber (2011, 2016, 2017), Ketcham et al. (2012), Heiss et al. (2010), Polyakova (2016), Ericson (2014), Ho et al. (2017), and Ketcham et al. (2016)), Medigap (Fang et al. (2008)) and employer sponsored insurance (Bhargava et al. (2017), Handel (2013), Handel and Kolstad (2015b)). We note, however, that there is no work that we are aware of on choice quality in Medicare Advantage, despite the prominence of the program.

A market answer to the issue could be to rely on skilled agents — insurance brokers and enrollment agents. Despite the important theoretical role in market function, there is little empirical evidence on the role played by agents in choices in health insurance markets nor on the degree to which reliance on human expertise addresses choice errors.[3] Evidence from other settings suggest that skilled agents do not ameliorate market failures stemming from a lack of consumer information. E.g. financial advisers maximize their own fees not client results (Mullainathan et al. (2012), Egan et al. (2019), Egan (2019), Gambacorta et al. (2017)) and auctioneers have dramatic differences in the prices they obtain for homogeneous products in relatively simple auction formats (Lacetera et al. (2016)).

As technological capabilities improve and detailed data become available, decision aids — AI based assessment of choices — offer an additional option to improve consumer decisions in insurance markets. Research in this area is, however, still nascent, focusing on either rudimentary forms of information provision (e.g Kling et al. (2012)), lab experiments (e.g. Bhargava et al. (2017)) or tools that are poorly designed (e.g. Abaluck and Gruber (2016)). The limited evidence to date finds that consumers are unwilling to engage with decision support even when it is a default (Abaluck and Gruber (2017); Bundorf et al. (2019)). This raises the question of whether decision support can be more effective when paired with skilled experts.

In this paper we study the roles of skilled expert intermediaries and sophisticated AI-based decision support technology in health insurance markets. We focus on a retiree health insurance exchange, one of the largest private Medicare exchanges in the United States (hereafter The Exchange). The Exchange offers products from across the Medicare universe including Medicare Advantage (MA), Medicare Part D (prescription drug coverage) and Medigap (supplemental financial coverage). Firms contract with the Exchange to offer an insurance options to their retiree population. We focus our analysis on the MA market, which in 2019 enrolled 22 million seniors nationwide (34% of all seniors in Medicare). Over the time period 2015-2017 we study the behavior of the approximately 800 enrollment agents advising seniors on MA choices. Agents expend considerable effort in terms of time spent advising enrollees; in 2015, prior to the adoption of AI-based decision support, the mean agent-consumer call time in our study sample was 59.4 minutes.[4]

[3] see Karaca-Mandic et al. (2018) for a discussion of the role of brokers.

[4] Labor cost associated with calls are one of the primary the costs of operating an insurance market place and essentially the total marginal cost of enrollment. Thus, our measure of productivity which is based on call time

During our study period, the Exchange partnered with Picwell, a technology firm, to implement AI-based decision support for its agents who were tasked with assisting consumers in their plan choices. Beginning in 2017, Picwell was fully integrated into the enrollment software used by agents providing them by default with individual-specific information on the quality of different plan options. Throughout our study period, potential enrollees were randomly assigned to agents, allowing us to avoid issues related to endogenous matching between agents and consumers. Combining random assignment to agents with the integration of Picwell into the agent enrollment software allows us to investigate the effectiveness of skilled agents in aiding consumer insurance choices and how that effectiveness changes when their production function is disrupted by new technology.

We begin our analysis with a simple model of the joint agent-consumer decision, with a focus on (i) how agent advice shapes consumers decisions and (ii) how improved agent information influences consumer decisions. We pay particular attention to the fact that the AI tool provides sophisticated information on some dimensions (notably expected plan financial impacts) but not on other dimensions (e.g. plan networks and consumer preferences for those networks). The model allows for a range of scenarios, including, e.g., the possibility that agents will overly rely on the AI relative to information that agents observe but is excluded from the AI tool. The model motivates our empirical work in which we model how agents weight (i) information incorporated into the algorithm and (ii) information that is excluded from the algorithm but is likely observed by the agent.

Our empirical analysis first assesses joint agent-consumer choice quality in the absence of decision support. We find that the average plan enrollment led to a \$1,260 financial loss for consumers, relative to the best financial option available to them. These foregone savings amount to 30% of total costs, which is comparable to findings from previous studies of consumer choice of Part D drug plans and consumer choice of employer-provided insurance. We next estimate a structural model of plan choice as a function of plan characteristics. We show that in 2015, before decision support was fully integrated, agent-consumer choices display a number of choice errors. Choices are between 6 and 7 times more sensitive to plan premiums than to expected out-of-pocket spending. Choices are sensitive to broad plan characteristics such as deductibles even after controlling for individual-specific out-of-pocket spending risk. Moreover, choices continue to value brands heavily, despite, in some cases, a lack of obvious reasons to do so. Thus, we find clear evidence that skilled agents alone do not overcome the choice errors made by consumers.[5]

We next turn to the impact of AI-based decision support on choice quality. We find that implementation of decision support has a meaningful impact on choices. We use our structural model to control for changes in the nature of choice set. We find that the 2017 choices made under 2015 decision-making weights are \$278 worse on average than 2017 choices made under 2017 decision-making. Notably, large mistakes are substantially more likely under 2015 decision-making

accords closely with the objectives of the firm.

[5]We address the validity of these findings in two key additional ways. First, we show in a series of simulations that measurement error of projected out-of-pocket spending by the broker / consumer pair cannot reasonably account for our findings. Second, we make the case that consumer risk-aversion is highly unlikely to justify the expected financial losses we document.

than under 2017 decision-making.

Our model estimates shed light on factors contributing to these improved decisions. In contrast to 2015, 2017 decision-making places almost identical weights on premiums and expected out-of-pocket spending — the weights expected for a fully informed, rational decision maker. Furthermore, 2017 decision-making placed much lower weights on plan characteristics such as deductible and out-of-pocket maximum when those characteristics are included in a model that also includes expected out-of-pocket spending.[6]

In addition to studying the impact of AI on the financial implications of choices, we also study a key element of plan choice that is not included in the algorithm: provider network breadth. Network breadth and preferences for that breadth are a feature where human agents might have a comparative advantage in assessment. Since this information is an "unused observable" (Finkelstein and Poterba (2014)) in the algorithm we can test for asymmetric information in recommendations and the degree to which following algorithmic recommendations crowds-out or distorts recommendation components observed by decision makers but not the algorithm. Our estimates of the importance of network breadth on decision-making remain essentially unchanged from 2015 to 2017. This suggests the skilled agents retain some value in eliciting private preferences that may not be included in, and could be distorted by, algorithmic recommendations.

Continuing in this vein, we study the impact of AI on brand preferences. We find that once AI-based decision support is made available, most insurer brand effects fall by an order or magnitude and have little impact on demand. An exception is Kaiser Permanente, which is the only brand that reflects vertically integrated health care delivery and is widely believed to offer a different form of coverage and care than traditional private insurers. Kaiser continues to be chosen in 2017 even when there is a significant expected cost of doing so relative to other brands – but those choices change in a systematic way that trades off expected cost with brand preferences. Those who were incurring large financial losses in Kaiser, relative to other options, do not choose it after AI-based decisions support but those with smaller losses remain. These results demonstrate agents' ability to synthesize information that trades-off marginal effects, rather than more blunt heuristics such as blanket brand preferences or blindly following the AI recommendations.

Though our evidence clearly points to improved choices after the introduction of AI, we provide additional evidence that this is true by examining an ex-post measure of choice satisfaction: subsequent switching out of the chosen plan. We find that people who enrolled in plans that the algorithm did not recommend in 2017 were more than twice as likely to switch plans the following year, compared to people who enrolled in a recommended plan. We augment this analysis by instrumenting for plan choice using a judges design with agent fixed effects. We find a causal impact of higher AI-based recommendations/score on subsequent enrollee experience utility, measured by turnover. For an agent who is one standard deviation better (3 plan score points) in terms of predicted plan score, consumers are 7 percentage points less likely to switch plans the following

[6] Our measurement error robustness analysis suggests that improvements in the precise measurement of out-of-pocket spending by agents are not the reason for these across the board performance improvements. Instead, other factors such as the integration of the various determinants of plan values into a final decision likely underlie these marked improvements.

year. This effect is equivalent to nearly the full propensity to switch plans in the overall sample.

We extend our core findings in three ways. First, we assess the impact of decision support across the skill distribution of agents. We estimate a model of plan choice quality in 2015 that includes agent fixed effects as a measure of average agent skill. Measured agent skill heterogeneity is large, with foregone savings that are twice as large in the worst quintile of agents as in the best quintile. Mandating AI-based decision support dramatically compresses this distribution, with little impact on the top quintile and dramatic improvements for the bottom quintile. We show that these results are very likely not the result of statistical mean reversion but, instead, the result of the compression of skill towards the top once AI is implemented.

We also analyze the impacts of decision support for the production efficiency of health insurance recommendations. We find that the introduction of decision support lowered call times by roughly 20%, while improving recommendations along ex-ante and ex-post measures. Moreover, the reduced time investment is constant throughout the distribution of agent skill, so that the return to the tool is much larger for the least skilled agents.

Finally, we consider the inherent link between choice adequacy and adverse selection, as discussed in, e.g., Handel (2013): it is possible that better individual consumer choices can facilitate more acute sorting of sicker consumers to generous plans (and vice-versa), leading to a greater degree of adverse selection. While a number of papers have discussed this concern, no paper that we are aware of shows how reduced choice errors actually impact adverse selection in a given empirical context. We demonstrate that improved choices do lead to more acute sorting, implying the potential for greater adverse selection if tools like the one we study are rolled out to all market participants.

The rest of the paper proceeds as follows. Section 2 presents the data and setting. Section 3 develops our model and empirical approach. Section 4 presents results and Section 5 concludes.

2 Data and Setting

2.1 Medicare Advantage

Medicare provides universal government-sponsored health insurance for the elderly and disabled in the U.S. Enrollees can access coverage through various channels. Medicare-eligible individuals are automatically enrolled in the Medicare Part A program, which covers inpatient hospital expenses. Eligible individuals can elect to enroll in Medicare Part B to cover outpatient expenses and choose among privately provided Medicare Part D plans to cover prescription drug expenses. Beyond this, privately offered Medigap plans that cover out-of-pocket costs under Medicare Parts A and B (Original Medicare) are also available. A combination of these options constitutes an enrollment in Original Medicare.

Alternatively, an enrollee can opt out of Original Medicare by choosing among a set of competing private Medicare Advantage plan. MA plans can be offered as a stand-alone plan only covering medical care or a product that combines this coverage with prescription drug insurance: MA-PD.[7] MA plans can offer additional benefits above and beyond those provided by Original Medicare and may charge additional premiums. MA plans are offered on a county-by-county basis, and their plans offer services through managed care networks. Nationwide, approximately 22 million people - or about 34% of those eligible for Medicare - enrolled in an MA plan in 2019. In addition to the patient premium, MA plans receive reimbursements from CMS based on bids that they submit, costs relative to local original Medicare costs, and risk-adjustments to account for differences in the enrolled population (Geruso and Layton (2015) ;Brown et al. (2014) ; Newhouse et al. (2012)).

We focus on MA-PD plans in our analysis. This allows us to study choices of insurance across the bulk of health care utilization (i.e. prescription drugs, inpatient and outpatient medical care). In this way, MA-PD plans more closely resemble the kinds of health benefits chosen by those outside of Medicare (e.g. employer-based) and in other settings (e.g. Medicaid and outside of the U.S.). The complexity of this more general setting may also make expert recommendations more beneficial than in simpler settings where consumer choice quality is studied (e.g. Medicare Part D coverage for prescription drugs).

People who choose to enroll in MA-PD plans cannot also enroll in Medigap or Part D plans. That is, the Medicare enrollment decisions that people make can be described as a decision tree where, at the first level they choose whether to build there coverage around Original Medicare or MA. Those who choose Original Medicare must then select a Part D plan and they must decide whether they want additional Medigap insurance. Those who choose an MA-PD plan must choose from among the plans available, but they only need to make one plan choice. This simplifies our analysis and makes it likely that the choices we observe reflect the near entirety of consumers' health benefits.

2.2 Data

We study health plan choices for individuals enrolling in MA-PD plans through a private health insurance exchange ("the Exchange") for the 2015 and 2017 Open Enrollment Periods.[8] We observe detailed information on approximately 59,000 MA-PD enrollees, their agents, and their enrollment options in both 2015 and 2017. At the enrollee level, we observe age, sex, zip code, county of residence and a list of prescription drugs that they take. At the agent level, we observe the identity of each customer, the number and duration of calls for that customer, the plan that customer enrolls in, and the number of years of experience each agent has working on the Exchange. For each choice made we observe the set of Medicare plans available to each enrollee and detailed information about those plans including premiums, deductibles, out-of-pocket maximums, coinsurance rates for

[7] See Starc and Town (2015) for a discussion of the impacts of bundling coverage components together on benefit design in the MA market.

[8] The volume of new enrollments on the Exchange varies from year-to-year, depending on the sizes of retiree groups entering the Exchange. In 2016, the second year that decision support was available, the number of new enrollments and the number of agents on the Exchange were substantially lower, compared to both 2015 and 2017, due primarily to lower participation levels by large employers. We exclude 2016 from our analysis because the low volume of agents and the low volume of customers per agent do not allow us to draw any meaningful conclusions about the relationship between decision support and plan selection.

various categories of coverage, actuarial value, plan type (e.g. PPO, HMO, POS), brand name and network breadth. We also observe both the mean and variance of out-of-pocket for each individual in each plan and a plan score that enrollment agents could use to compare plans. This combination of enrollee, agent, and choice set characteristics allows us to observe and evaluate the joint enrollment decisions that Medicare enrollees make with their enrollment agent.

Table 1 also presents summary statistics for the enrollee population in 2015 and 2017. We restrict our study population to people between the ages of 64 and 90 at the time of enrollment and only consider enrollment decisions made by people who ultimately enrolled in a MA plan that also covers prescription drugs (MA-PD). After these restrictions, the remaining populations for 2015 and 2017 were 31,090 and 27,739, respectively. In both years, approximately 55% of of enrollees were female. The study population was slightly older in 2017, with a mean age of 72.7 compared to 71.2, but 2017 enrollees took slightly fewer prescriptions, with a mean number of 3.4 prescriptions per enrollee in 2017 compared to 3.7 in 2015. Overall, the population in the MA-PD market stayed remarkably consistent based on observables between 2015 and 2017. This suggests that the introduction of AI-based decision support did not systematically alter the extensive margin decision to select MA versus original Medicare.[9]

We observe enrollment appointment information for 835 agents in 2015 and 732 agents in 2017. Table 1 presents summary statistics on the agent population. 305 worked for the Exchange in both years, but the agent population was less experienced, on average, in 2017 compared to 2015. The average years of experience for agents in 2017 was 2.98 compared to 4.02 in 2015. More than half of the agents in 2017 had no prior experience selling Medicare plans on the Exchange, reflecting the fact that a large part of the agent workforce is comprised of seasonal hires. Average appointment duration was 59 minutes in 2015 and average appointment duration was 48 minutes in 2017. Conditional on ultimately purchasing a plan, 78% of consumers needed only one call to finalize the purchase while the remaining 22% needed more than one call to do so.[10]

Table 1 presents summary statistics for the MA-PD plans available to enrollees in 2015 and 2017. In both 2015 and 2017, average premiums were similar, with a mean monthly premium of \$49.21 in 2015 and \$55.77 in 2017 (in addition to the base Part B premium), and the range in premiums available in a choice set was similar, with a mean difference between the highest and lowest monthly premium available of \$156.47 in 2015 and \$154.25 in 2017. In both years, HMO plans made up a similar share of plans offered. In 2015, slightly more regional PPOs and slightly fewer local PPOs were available. In both years, cost plans and private Fee for Service plans were rare.

In both periods there were a large number of plan choices available. The mean number of

[9] These empirical results are also consistent with our understanding of the marketplace where the process for selecting a particular type of coverage did not change following the introduction of Picwell. AI was available once an enrollee selects a particular type of coverage to select among options but not in choosing MA versus product combinations under Original Medicare. AI was available in all product markets including MA, MA-PD, Medigap and Part D, suggesting that there was no reason to differentially steer an enrollee between different market options after the introduction of Picwell.

[10] We do not observe calls that were made by consumers who did not end up purchasing a plan. Our understanding is that a vast majority of consumers purchase a plan once they engage with an agent in this market.

Table 1: Summary Statistics

	Enrollee			Agent			MA Plans	
	2015	**2017**		**2015**	**2017**		**2015**	**2017**
Enrollees	31,090	27,739	Agents	835	732			
Age			Years of experience			# Plans in Choice Set		
Mean	71.15	72.73	Mean	4.02	2.98	Mean	12.43	12.47
p25	67	67	p25	3	1	p25	7	8
p50	70	71	p50	4	1	p50	13	12
p75	74	77	p75	5	6	p75	17	17
p95	82	86				p95	23	22
Drugs			Total customers			Monthly Premium		
Mean	3.68	3.37	Mean	37.32	37.10	Mean	49.21	55.77
p25	1	1	p25	15	77	p25	0.00	0.00
p50	3	3	p50	29	33	p50	26.00	37.00
p75	5	5	p75	55	54	p75	80.00	86.00
p95	9	9				p95	188.00	179.00
Female	55.3%	55.5%	Call time			Plan Type		
			Mean	59.4	47.8	Cost	0.7	0.7
			p25	29	11	HMO	53.2	54.3
			p50	50	40	HMO-POS	8.1	7.4
			p75	79	69	PFFS	3.0	2.9
			p75	79	69	PPO	21.8	24.3
						Regional PPO	13.2	10.4
						Premium Range	156.47	154.25

options is 12.5 with a 95th percentile of 23 options in 2015 and remains very similar in 2017.

2.3 The Exchange

The Exchange is one of the largest private Medicare exchanges in the United States. Medicare eligible individuals can select plans on the Exchange throughout the year as they turn 65 or have another qualifying event (e.g., losing employer-provided coverage) that triggers a special enrollment period. Most enrollment in the Exchange occurs between October 15 and December 31 each year during the Open Enrollment period. During the 2017 Open Enrollment period (which occurred between October 15, 2016 and December 31, 2016), 87,691 Medicare eligibles enrolled in plans through the Exchange. 41,563 of these people enrolled in a Medigap plan, 44,883 enrolled in a Part D plan, and 34,616 enrolled in a MA plan.

The Exchange employs enrollment agents who help customers understand their Medicare options and enroll in a plan. Each open enrollment period, the Exchange schedules appointments for customers to review their options and enroll in a plan with an agent. All agents are licensed to sell Medicare policies in multiple states, and agents are randomly assigned to appointments with customers.

To confirm that assignment is random we divide agents by baseline skill level. We return to our definition of skill below but demonstrate here that assignment is orthogonal to agent skill. This allows us to test for random assignment that is a primary threat to our analysis: that particular types of enrollees (e.g. more complex) might be assigned to particularly agents (e.g. more skilled at handling particular levels of complexity or specific conditions).

Table 2 presents gender, age and number of prescriptions by quintiles of agents skill. We can reject differences across agents in any of these key demographics that determine both the value of a specific product and the potential complexity of identifying the right plan. We, therefore, proceed with the assumption that agent-enrollee assignment is random.

Table 2: Summary of 2015 customer characteristics by agent quality quintile

Agent quality	**Female share of customers**		**Customer age**		**Rx per customer**	
quintile	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	0.55	0.50	71.2	5.3	4.4	3.1
2	0.55	0.50	71.0	5.1	4.5	3.0
3	0.56	0.50	71.0	5.2	4.5	3.1
4	0.56	0.50	71.2	5.3	4.5	3.0
5	0.56	0.50	71.8	5.7	4.6	3.0

Agents were paid hourly and received additional bonuses when customers enrolled in a plan. Enrollment bonuses did not vary by plan, so an agent would receive the same enrollment bonus whether a customer enrolled in an MA-PD plan or a Medigap plan and a Part D plan. The bonus for customers enrolling in just a Part D plan was less than half of the bonus for enrolling in both medical and drug coverage, so, while agents did not face financial incentives to direct customers towards one particular health plan, they did face incentives to enroll agents in a plan or set of plans that would cover both medical and prescription drug costs.[11]

In 2015 agents used web-based enrollment software that did not include decision support for MA plans. Decision support was available in 2015 but only a few weeks prior to open enrollment and to use the tool an agent had to leave their existing software platform. Accordingly, few agents consistently used Picwell decision support for a variety of potential reasons including because they were unaware of it, they were not familiar with how to use it and how to explain the resulting recommendations to customers, or they did not fully trust the recommendations.

The Exchange embedded decision support in the software used for all of their customers in 2017. This change was accompanied with training on how Picwell scores and cost estimates are generated and how to use the decision support tool to help Exchange customers choose Medicare plans.

To receive a Picwell recommendation, agents would walk customers through a user intake process that required entering in personal information across several steps. In the first page, agents would enter in the customer's age, sex, zip code and county. Agents would then ask customers about any prescription drugs that they routinely take and they would ask them to list any health care providers that they routinely see. After entering this information, agents would request recommendations, and Picwell would return sets of recommendations for all Medicare Advantage, Medigap and Part D plans sold on the Exchange in the customer's county. Within each type of Medicare plan, recommendations would be returned, sorted by the Picwell Score. Picwell Scores were correlated with one of three color tiers - Green, Yellow, and Red - that indicate, in descending order, how well plans match customer preferences, based on expected utility calculations. The expected utility calculation underlying the Picwell Score uses estimated variance in OOP costs to account for risk aversion among Medicare enrollees, we discuss this approach in more depth below. In addition to the Picwell Score and color tier, Picwell decision support predicts OOP costs for each plan available to a customer and combines this with premiums to return estimates of each plan's total expected cost, which is characterized as a plan's "RealCost"" on the Exchange. In 2017, in addition to seeing a point estimate of the "RealCost", Picwell users could also see a range that indicated the 20th and 80th percentiles of predicted OOP costs for each person and plan.

2.4 Decision Support Background

Agents on the Exchange were able to use decision support technology to evaluate and compare Medicare plans available to Exchange customers. The decision support evaluates all options within a specific type of Medicare plan type (e.g. all MA-PD plans), but it does not make comparisons across plan types.[12]

[11] This level of alignment between agents and enrollees is not representative of many insurance transactions in the U.S. Frequently, insurance brokers are paid commissions that vary by carrier and need to be disclosed to the enrollee. Thus, we expect our results on agent quality to be an upper bound for the quality of agency.

[12] This technology is available. However, Medicare marketing rules limit the ability to make recommendations across bundles of plans.

The plan evaluations include (i) predicted "RealCost" which combines annual premiums with mean estimated OOP, (ii) a Picwell plan Score that rates plans on a 100 point scale, and (iii) a color tier that is simply a mapping of plan score to one of 3 color tiers with scores of 90 or greater assigned to the "Green" tier, scores of 75 to 89 assigned to the "Yellow" tier, and scores of 74 and lower assigned to the "Red" tier. The Picwell Score identifies the "utility maximizing" plan within each choice set and assigns the highest score to this plan. Scores for all other plans identify how close the expected utility of each plan is to the plan with the highest expected utility. Agents were instructed to interpret the Picwell Score as an identifier for how well each plan matches a customer's preferences and to treat any plan on the "Green" color tier as a good match.

The process of generating a set of plan recommendations can be divided into three distinct steps. The first step estimates total medical spending for each individual k, the second step translates predicted spending for each individual k into OOP costs for each plan j available to individual k, and the third step translates the OOP for each plan j available to individual k into a utility that is than converted to a 100 point scale.

In the first step of this process, a machine learning model predicts annual medical spending for individual k. The prediction model uses a database with claims for MA enrollees that includes 2 years of continuous enrollment and claims for approximately 1.2 million individuals. Model features are defined based on observable characteristics in the first year (t) of the 2 year claim period, and the prediction target is the total allowed costs incurred in year $t + 1$. This generates a mapping from individual characteristics μ_k (including age, sex, and a list of prescription drugs) to a "risk group" of individuals K in the claims data with a distribution of allowed costs $f(ALLOWED|\mu_K)$.

The performance of the machine learning model compares favorably to other risk rating models in terms of out-of-sample prediction. The particular version of the model that was deployed in this setting, which predicts costs using age, sex and prescriptions as model inputs, is able to explain 3.5 times more variation in medical cost for new Medicare enrollees than the risk rating model that CMS uses for new enrollees, and the version of the model that the Exchange currently implements is able to match the performance of the CMS risk-rating model for continuing Medicare enrollees, which includes diagnosis codes. In other words, using less, and easier to obtain information, the machine learning approach is able to achieve the same performance benchmarks of other models that make use of more detailed information.

In the second step, the decision support applies benefit calculators for each plan j to year $t + 1$ claims for each of the individual in risk group K to generate a distribution of OOP cost for each plan $f(OOP|\mu_K, \psi_j)$. The benefit calculators account for detailed plan information (ψ_j) including deductibles, OOP maximums, formularies and coverage and cost sharing rules for every benefit category in each plan. The decision support calculates $E(OOP_{kj})$ based on $f(OOP_{kj})$, and uses this to return the RealCost, or expected total cost, for each person-plan pair, where $RealCost_{kj} = Premium_{kj} + E(OOP_{kj})$.

In the third step, the decision support calculates utility and a score from $f(OOP)$. Utility is calculated as a function of personal and plan attributes using a constant absolute risk aversion model.[13] U_{kj} is translated to dollars by calculating a certainty equivalent CEQ_{kj}. This allows us to estimate a risk penalty, r_{kj}, where

$$r_{kj} = CEQ_{kj} - (P_{kj} + E(OOP_{kj})) \tag{1}$$

The risk penalty can be interpreted as the additional annual premium that an individual with risk aversion γ and individual characteristics μ_k would be willing to pay to eliminate all variance around $E(OOP_{jk})$. In other words, it represents a marginal willingness to pay to eliminate risk that represents both individual preferences toward risk and exposure to risk. Finally, the decision support applies a score function that translates each CEQ_{kj} in individual ks choice set into a score between 0 and 100.

In 2017, the third year that decision support was available, agents were required to generate recommendation requests for all customers. Furthermore, prior to the 2017 Open Enrollment period agents received training in the use of the decision support technology. Combined, this led to near universal adoption of the tool and, potentially, enhanced trust in the recommendations.[14]

3 Choice Model and Empirical Specification

3.1 Insurance Demand

We begin with a baseline model of an expected utility maximizing enrollee choosing among insurance products. Each individual, indexed by $k \in K$ faces a distribution of cost outcomes for Medicare Advantage plan options $j \in J$ $F_{k,j}()$. Enrollees get latent utility from choosing a plan j according to the following von Neumann-Morgenstern (vNM) expected utility:

$$U_{kj} = \int_0^{\infty} f_{kj}(s) u_k(W_k, x_{kj}(P_{kj}, s)) ds \tag{2}$$

Here u_k is a vNM utility index and s is a realization of out-of-pocket cost from the distribution $F_{k,j}()$. Individual specific wealth is captured by W_k and P_j is the premium for plan j. x_{kj} is an individual's level of consumption based on their realized health shock, s. We model this as:

$$x_{kj} = W_k - P_j - s + \epsilon_{kj} \tag{3}$$

where ϵ_{kj} is a mean zero individual specific shock. We follow the literature and assume that families have constant absolute risk aversion (CARA) preferences implying that, for a given *ex post*

[13] In typical implementations, the decision support technology assigns risk aversion parameters to customers based on responses to survey questions about attitudes towards risk, but in this particular application, such survey questions were not permitted, so all customers were assigned a risk aversion parameter of $\gamma = 4.0 * 10^{-4}$, which is similar to estimates from Handel (2013). As shown in the online appendix in Abaluck and Gruber (2011), the distinction between CRRA and CARA risk preferences in our context is very unlikely to matter materially for our empirical results.

[14] We do not study the training aspect directly. However, the cost of training agents is important to operating the exchange. AI-based decisions support has the potential to reduce training time, holding quality fixed.

consumption level x:[15]

$$u_k(x) = \frac{1}{\gamma_k} e^{-\gamma_k x} \tag{4}$$

This specification constitutes the baseline choice model; reflecting the basic role that insurance plays as a product to mitigate financial risk. However, we extend the model to allow for three specific additional features that are likely to impact choices: (i) heuristic/behavioral decision making (ii) the role health insurance plays in allowing access to different health care providers and (iii) brand preferences for insurers. For (i), we include a set of salient plan features λ_j including the deductible and out-of-pocket maximum for plan j. The elements of λ_j are observable but, as equation 2 shows, do not affect realized utility conditional on the realized spending level s. Even though these financial characteristics of a plan only affect utility through their impact on the distribution of risk consumers may still heuristically place weight on them when making choices (see, e.g., Abaluck and Gruber (2011)). For (ii) we also allow the network of available providers in plan j to enter utility captured by n_j. For (iii), we incorporate brand effects where a given insurer's brand κ is constant across all plans j that that insurer sells. Preferences for the insurer brand could capture pure non-welfare-relevant brand effects or welfare-relevant effects such as, e.g., differences in administrative support or online tools. Incorporating these features we express the utility of each state s as:

$$x_{kj} = W_k - P_j - s + \lambda_j + n_j + \kappa_j + \epsilon_{kj} \tag{5}$$

To implement the model empirically we parameterize utility as:

$$u_{kj} = \rho P_k - \delta E(OOP_{kj}|\mu_k, \psi_j) + \phi R_k(\gamma, s) + \omega \lambda_k + \beta n_j + \alpha_j \kappa_j + \epsilon_{kj} \tag{6}$$

where P_j is the premium for plan j, $E(OOP_{kj}|\mu_k, \psi_j)$ is a measure of expected out-of-pocket spend for individual k in plan j based on the predictive model and R_j is a function reflecting the risk protective value of plan j. We assume a uniform risk aversion value in the population ($\gamma = 4.0 * 10^{-4}$) and compute R for each customer and plan following equation 1. Finally, ϵ_{kj} is a type-I extreme value error term.

3.2 Agency and AI

To this point our model has simply specified an enrollee utility from insurance. The central question of this paper, however, is the role played by both agents and AI in choices. We model enrollees who rely on agents to help them choose an insurance plan that maximizes their utility. Assume that true enrollee utility can be expressed as:

$$E(u_{kj}) = \Lambda X_{kj} + \epsilon_{kj} \tag{7}$$

[15] As shown in the online appendix in Abaluck and Gruber (2011), the distinction between CRRA and CARA risk preferences in our context is very unlikely to matter materially for our empirical results.

where Λ captures a vector of weights on plan attributes that map to realized plan utility.[16] Utility, with full information, is maximized over an expectation based on unbiased observations of plan attributes available to the enrollee and accords with a full information version of equation 7.

We assume agents have an (unobserved) information set/signal Ω^b that consists of information from enrollee k and knowledge of plan attributes for plan j. Latent agent skill as well as effort affect the realization of this signal. Agents then develop a set of weights for plan attributes that scale true enrollee utility captured by the vector Θ. Agents and enrollees then make enrollment decisions by maximizing expected utility according to:

$$E(u_{kj}) = (\Theta|\Omega^b)\Lambda X_{kj} + \epsilon_{kj} \tag{8}$$

Agents' weights (Θ) reflect the degree to which they are able to capture the true, latent preferences of enrollees (Λ). For example, elements of the vector Θ equal to 1 represent perfect agency; the agent puts the same weight on an attribute that an enrollee would.

AI-based decision support enters as an additional source of information that an agent includes in their information set, captured in the model as a new, post-AI, information set:

$$\Omega' = \alpha\Omega^{AI} + (1-\alpha)\Omega^b \tag{9}$$

where α captures the relative weight on AI-based information versus prior beliefs used to form the new information set.[17]

We assume that the AI-based signal provides information on only a subset of the attributes that enter choice. This accords with our specific setting, where the AI-based tool focused primarily on financial/cost components of the choice. It also captures a general phenomenon in which AI-based tools are particularly good at accounting for quantifiable aspects of decisions but rarely account for the universe of welfare relevant aspects of a decision.

Attributes are partitioned into X^i_{kj} and X^g_{kj} where i indexes attributes observed by agents and included in AI and g indexes attributes observed by agents but not included in AI. After AI-based decision support is introduced enrollment decisions are made to maximize expected utility according to:

$$E(u_{kj}) = (\Theta'_i|\Omega')\Lambda X^i_{kj} + (\Theta'_g|\Omega')\Lambda X^g_{kj} + \epsilon_{kj} \tag{10}$$

Instead of specifying the micro-foundations either for information acquisition in equation 9 or the resulting weights in recommendations in equation 10, we allow for a flexible model in which AI can concurrently change the information available and weights on different attributes — those included in AI and those excluded. This allows us to flexibly capture a variety of ways in which

[16] Without loss of generality we express utility here as a linear function of plan attributes. One could alternatively specify utility as $E(u_{kj}) = \Lambda f(X_{kj}) + \epsilon_{kj}$ to capture a flexible function of observables. We follow our empirical implementation in equation 6 in which we express CARA risk preferences as a linear term because we implement the model of agency and AI empirically using the same specification.

[17] We express information acquisition as a linear combination of beliefs. One could specify the updating process more generally. Since we do not empirically model learning itself we simplify exposition with linear weights.

agents might incorporate information. It nests models in which agents efficiently (in a Bayesian sense) integrate information (e.g. Diamond (1971), Dranove and Satterthwaite (1992)) as well as a broader class of models in which agents are rationally inattentive in selecting attributes, the associated weights and in making recommendations with and without AI (see Mackowiak et al. (2018) for a review). The structure also allows for more behavioral models in which attention is heuristically allocated in ways that need not be optimal and may reflect a variety of biases (see e.g. Handel and Schwartzstein (2018)).[18]

Following equation 10 we expect the introduction of AI-based decision support to alter the weight placed on attributes included in the AI-based tool when agents i) incorporate the AI-based recommendations and ii) AI provides new information that changes beliefs. The aggregate change in attribute weights therefore depends on the combination of agent updating (captured by Ω') as well as the covariance of the attributes across choice set options.

For example, when an agent is able to access a predicted measure of out-of-pocket cost, this may make out-of-pocket costs more salient and provide more precise individual-plan-specific information on those costs. Both of these factors could contribute to an increased weight on expected out-of-pocket spending in observed plan choices.

Equation 10 also conditions the weights on X^g_{kj} — attributes not directly changed by AI — on the new information set Ω'. Weights on attributes not included in the AI still change because we expect relative attribute weights to change as the overall information set changes. For example, were agents to rely on heuristics to deal with the complex prediction problem of estimating cost they might have put a high weight on plan premium or simple measures of generosity such as the level of deductible or out-of-pocket maximum prior to AI. If they gain new information on total cost we expect the weights on those measures to decline relative to the out-of-pocket cost prediction that can be generated with AI.

Our model does not assume that AI improves choices, in the sense of better approximating true preferences. AI-based decision support might make some attributes of a plan more salient at the expense of harder to observe but nevertheless valuable components of a plan. For example, particular insurance brands like Kaiser Permanente that enrollees prefer do not have their non-financial features accounted for in the AI (e.g. the breadth of physician network or specific health care delivery models). As a result, weights on these excluded attributes may decline relative to the weights on included attributes. Alternatively, if AI provides valuable new information and agents are rationally inattentive (i.e. they efficiently integrate information to optimize recommendations) we expect recommendations to (weakly) better approximate enrollee preferences. Which model better reflects behavior is an empirical question.

[18]With sufficient assumptions on both the model of learning and the covariance of the elements of X_{kj} we could recover estimates for α. Rather than undertake this structural approach (for which one might want to implement a more flexible parameterization of this model) we focus on a measure of how attributes are updated that accord with measures of enrollee outcome utility to evaluate how agents update and what the positive and normative impacts of that updating are.

3.3 Identification

Our empirical analysis relies on two key sources of identifying variation in our data. First, we use the fact that customers are randomly assigned to agents to deal with any issues related to agent-customer matching. Second, we leverage the change from 2015 to 2017 when the marketplace moved from little/no AI-based decision support to integrating the AI-based recommendation into the enrollment software used by all agents for all enrollments.

Relying on intertemporal variation alone presents an obvious challenge: what other features changed over time that might affect plan choice quality? The summary statistics demonstrate consistency over time in key statistics for both enrollees and agents, suggesting that we don't need to be concerned about large-scale changes to those participating in the exchange. We are concerned, however, about changes in the set of plans available to enrollees.

To address this issue we estimate a variant of equation 6 separately for 2015 and 2017. The associated model parameters capture the weights placed on plan attributes in each of those years. Based on estimated choice model parameters, we simulate choices in 2017 *holding fixed* the set of plans available. We compute:

1. 2017 choices based on 2017 demand parameter estimates

2. 2017 choices based on 2015 demand parameter estimates

We use this approach to compare how choices change moving from 2015 choice parameters to 2017 choice parameters for the set of plans available in 2017. Put differently, using our structural plan model we estimate the choices that would have been in made in 2017 by the same agents had they behaved as they did in 2015.

4 Results

Before moving to our primary results, using our choice model estimates, we discuss observed money left on the table in 2015. Figure 1 plots the money left on the table for a consumer's actual choice in 2015, relative to the best possible financial choice in their choice set.

Clearly, there are substantial sums of money left on the table in 2015 when looking just at the financial dimension of consumers' choices. More than half of consumers leave $1,000 on the table while a meaningful proportion leave over $2,000 on the table. The fact that consumers leave this much money on the table descriptively is in line with the basic facts put forth in other studies in the insurance choice literature including Abaluck and Gruber (2011), Handel (2013), Handel and Kolstad (2015b), Bhargava et al. (2017) and others, as summarized in Chandra et al. (2018).

Though this evidence is just suggestive of meaningful choice errors, there are quite a few reasons why this kind of histogram could be consistent with good choices. First, plans could be differentiated by networks of providers and brand effects (some portion of which may reflect substantive differentiation). Second, this is only an expected value for consumers. Consumers who value risk reduction (one of the primary purposes of insurance) will want to give up some expected value in

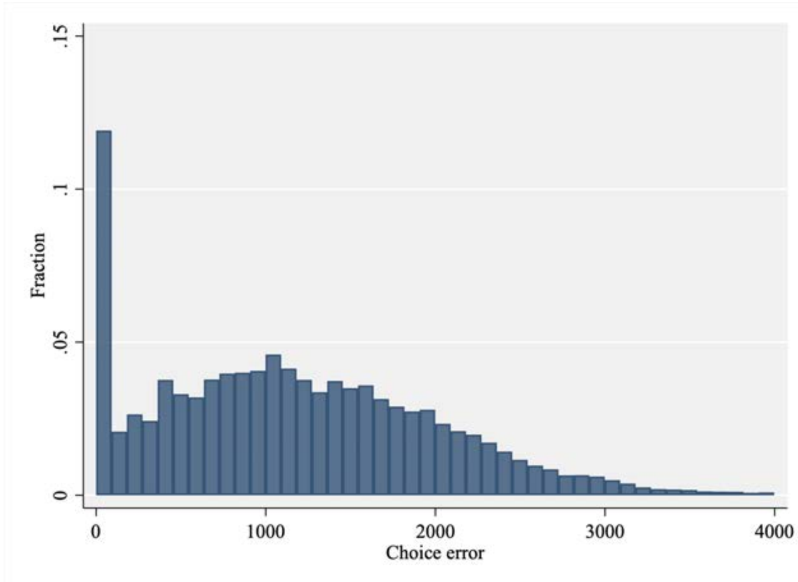

Figure 1: Histogram of observed 2015 money left on the table (financial choice error).

return for lower cost variance. We now turn to our primary choice model estimates, which account for these potentially important additional plan aspects.

4.1 Impact of Decision Support: Financial Value

Table 3 presents the estimates from our choice model, described in equation 4. The first two columns of the table present estimates for 2015, prior to the widespread use of algorithmic decision support. The first specification is a simplified version that includes both key inputs into the algorithm (annual premium and predicted OOP) and potentially important choice factors excluded from the algorithm (network coverage, plan type dummies, brand dummies). The second specification, our primary specification, also includes risk aversion and plan financial characteristics whose value should be fully subsumed by the predicted OOP variable but may not be due to customer / agent use of heuristics.

For our primary specification, in 2015, joint agent/consumer decisions place substantially more weight on plan premium than they do on expected plan out-of-pocket. For a rational, informed consumer — 'homo economicus' — these attributes should be valued identically. In practice, consumers choosing plans in 2015 weight premiums 6.5 times more than expected plan out-of-pocket spending.

A number of other results in the 2015 specification are at odds with standard economic models of choice. Even holding constant the individual's own out of pocket risk, individuals have a strong distaste for higher deductibles and higher maximum out-of-pocket spending levels. Once those distastes are factored in, consumers then (i) have a preference for plans with lower actuarial values and (ii) are willing to pay more for plans with higher risk premia, both inconsistent with typical "homo economics" choice models.

How do consumer choices, and associated preference estimates, change when algorithmic decision support is introduced in 2017? The third and fourth columns of Table 3 present estimates for the same choice model specifications estimated for 2017.

A number of important results emerge. First, the implementation of algorithmic decision support entirely removes the large bias weighting premiums more heavily than out-of-pocket spending: the ratio of these coefficients in 2017 is approximately 1 to 1, as opposed to 6.5 to 1 in 2015. This is a substantial change, especially given that this ratio is shown in the literature to be robustly different than one across many choice settings.

It is important to note that this change is not 'mechanical,' in the sense that it has to follow from the integration of decision support. As discussed in the model in Section 3.2, agents were not required to accept recommendations. They can consider both the algorithm's recommendation, and whether or not to take it, and the non-financial plan dimensions not included in the algorithm. In our coming analysis, we show that agents / consumers continue to value non-financial dimensions of plans, even while making choices that are more consistent with a 'homo economics' model in terms of financial plan dimensions.

The results in 2017 are also much more consistent with the standard economic model in a variety of ways. The coefficients on the deductible and maximum OOP are greatly reduced, as should be the case given the inclusion of individual out of pocket spending in the recommendation algorithm. The coefficient on actuarial value is right signed and remains small. The risk penalty coefficient becomes negative, consistent with individuals preferring plans that are less risky all else equal. Taken together, these results show clearly that decisions improve greatly when focusing on the dimensions that the recommendation algorithm incorporates.

Figure 2 illustrates the magnitude of the monetary improvement in choices from 2015 to 2017. It plots three distributions. First, it plots the actual distribution of money left on the table due to 2017 plan choices. Next, it plots the predicted distribution of money left on the table in 2017 choices using estimates from the choice model estimated on 2017 choices. These two lines are essentially on top of each other, showing strong model fit. In addition, both lines show that there are still large sums of money left on the table in 2017 choices, though this could be because of, e.g., preferences for Kaiser (discussed momentarily), as well as because of poor choices.

The third line plots the distribution of money left on the table for predicted 2017 choices using estimates from the 2015 choice model. This line reflects how well agents/consumers choose in 2017 if they act like they did in 2015. The figure clearly shows that 2015 choice model parameters lead to substantively worse outcomes. The share of enrollments in plans with zero or near zero choice error is substantially lower. In 2015, 9.8% of people enrolled in the plan with the lowest expected cost compared to 18.0% in 2017, and only 24.2% of people enrolled in a plan that was within \$500 of the lowest expected cost available in 2015 compared to 47.4% in 2017.

Table 4 presents key related statistics. First, the table shows that the average actual money left on the table in 2015 was \$1,261, as compared to \$895 in 2017. The money left on the table thus went down by \$365 after decision support was in widespread use. To make this comparison 'apples to apples' we compare mean 2017 money left on the table under actual choices to counterfactual

	2015		2017	
	(1)	**(2)**	**(1)**	**(2)**
Annual Premium ($100)	-0.0746***	-0.0984***	-0.0746***	-0.0633***
	(0.00117)	(0.00132)	(0.00128)	(0.00212)
Predicted OOP ($100)	-0.0110***	-0.0151***	-0.0214***	-0.0721***
	(0.001)	(0.00146)	(0.00108)	(0.00251)
Deductible ($100)		-0.347***		-0.0428***
		(0.00704)		(0.00606)
Max OOP ($100)		-0.0384***		-0.0129***
		(0.000702)		(0.00114)
Risk Penalty ($100)		0.204***		-0.0579***
		(0.00497)		(0.00246)
Actuarial Value		-0.0118***		0.0130***
		(0.00252)		(0.00296)
Network Coverage	0.0133***	0.0192***	0.0262***	0.0301***
	(0.00072)	(0.000722)	(0.000713)	(0.000754)
Plan Type				
HMO	-	-	-	-
PPO	0.974***	1.181***	0.832***	1.200***
	(0.0206)	(0.0225)	(0.0179)	(0.0211)
Other	-1.672***	-3.070***	-0.863***	-0.742***
	(0.114)	(0.0758)	(0.0753)	(0.0957)
Brand				
Regional carrier	-	-	-	-
Aetna	0.792***	0.343***	0.130***	0.240***
	(0.031)	(0.033)	(0.0238)	(0.0257)
Blue	1.129***	0.980***	-0.138***	0.0609*
	(0.0245)	(0.0251)	(0.0269)	(0.0276)
Humana	0.708***	0.985***	-0.829***	-0.317***
	(0.0271)	(0.0285)	(0.0294)	(0.0321)
Kaiser Permanente	3.184***	3.170***	1.664***	2.058***
	(0.0327)	(0.0388)	(0.0443)	(0.0474)
United	0.551***	1.037***	-0.364***	-0.362***
	(0.0305)	(0.0325)	(0.027)	(0.0263)
Pseudo R-squared	0.135	0.171	0.098	0.13
Observations	385,883	385,883	337,198	337,198

Standard errors in parentheses, * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Table 3: This table presents the estimates from our main structural choice models.

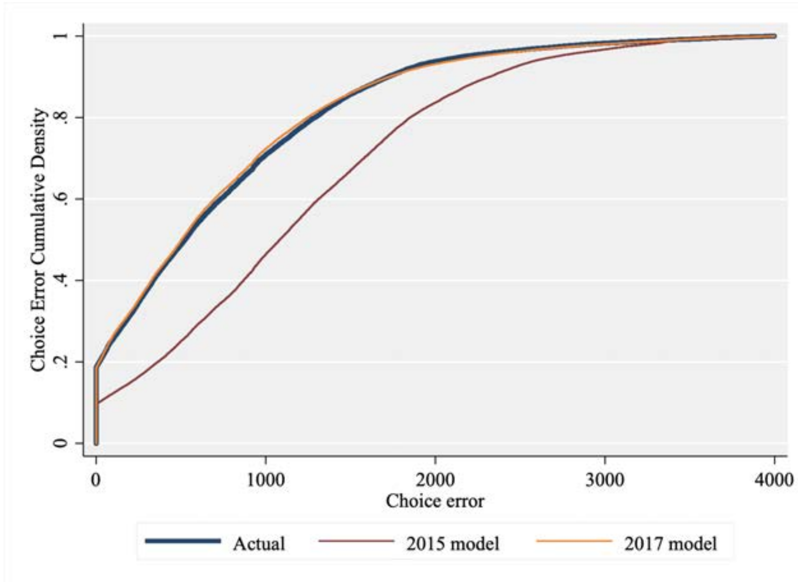

Figure 2: Observed 2017 choice error compared to simulated choice error

choices based on 2015 choice model parameter estimates. We find that now the average gain in terms of money left on the table from AI-based decision support is $278. Table 4 also presents distributional statistics that for these key quantities showing meaningful variation in the differential money left of on the table pre and post decision support.

Table 4: Actual and counterfactual choice error, full sample

		Percentile				
	Mean	10	25	50	75	90
Full Sample						
2015 Actual	$1,261	$66	$550	$1,124	$1,762	$2,342
2017 Actual	$895	$0	$101	$549	$1,190	$1,878
2017 Error w/ 2015 Sim. demand	$1,173	$15	$500	$1,083	$1,712	$2,331
2017 Error w/ 2017 Sim. demand	$895	$0	$90	$528	$1,150	$1,937
2015 Act. - 2017 Act.	$365	$66	$449	$575	$571	$464
2015 Sim. - 2017 Act.	$278	-$1,009	-$68	$389	$1,104	$1,738
2017 Sim. - 2017 Act.	-$1	-$1,134	-$434	$0	$416	$1,099

* Error differences >0 indicate reductions in cost error.

4.1.1 Robustness: Measurement Error

One potential concern with our analysis in this section is that our 2015 estimates are the result of brokers measuring consumer plan-specific out-of-pocket spending with error, rather than a jointly inconsistent decisions between brokers and customers. It seems clear that our results are not caused solely by econometric measurement error in plan-specific out-of-pocket spending. If that kind of measurement error were important, we would not find a major improvement in the weighting of premiums and out-of-pocket spending after the widespread decision support was introduced in 2017. Instead, we would find the same over-weighting of premiums in 2017.

However, our estimated coefficients could reflect a combination of (i)agent/consumer measurement error in estimating out-of-pocket spending and (ii) jointly inconsistent decisions in valuing a dollar of premiums versus out-of-pockets costs. In this case, the removal of the premium versus out-of-pocket coefficient wedge between 2015 and 2017 cannot be interpreted solely as improving choices through reducing inconsistency – it may instead simply reflect a more mechanical reduction in measurement error in out-of-pocket estimation. To assess the importance of these mechanisms, we perform a series of exercises.

We run a series of simulations that assume that agents have the normatively correct weights for premiums, out-of-pocket spending, and plan financial characteristics. We assume that preferences for things besides these financial components are as estimated in our primary specification. We then simulate 2015 choices under the following scenarios for agent beliefs about out-of-pocket spending:

1. **Baseline:** algorithm predicted individual-pan specific out-of-pocket used in primary model
2. **Coarse rounding:** assume that agents round consumer-plan-specific out-of-pocket to the nearest \$500 increment. We also implement this for the nearest \$1,000 increment.
3. **Normal Noise:** assume that agents have normally distributed mean 0 noise around the algorithmic projection. We use individual-plan-specification normal distributions with standard deviations equal to 200,500, and 1,000, 2,000, and 3,000 in five different implementations.

After we simulate choices in these scenarios, we estimate our primary structural model. Table 13 in the appendix reports the estimates for these specifications. We find that estimates based on the simulation with baseline out-of-pocket predictions yield estimates where agents value premiums and out-of-pocket spending similarly and don't place any additional weight on financial characteristics, both as expected. When we move to the simulations with coarse rounding for predictions of out-of-pocket spending, we find that measurement error from those predictions do not meaningfully alter the estimates from the baseline scenario (i.e. premium and out-of-pocket equally weighted and no emphasis on additional financial characteristics).

For the specifications which add normally distributed noise with σ of 200,500 or 1,000 (truncated to 0 from below) the coefficients remain similar to the homo economics parameters that the underlying simulations are based on. We also consider extreme noise increases to 2000 and 3000. Even in these extreme cases, while premiums are weighted more heavily than OOP costs, the ratio never rises above 2 to 1, well below our 2015 estimate. Moreover, the coefficients on other plan characteristics never rise to more than a small fraction of their 2015 values shown in Table 3. These results confirm that the initial wedge in the respective weights for premiums vs. expected out-of-pocket spending in 2015 is due to behavioral foundations that are different than agent measurement error of individual-plan-specific out-of-pocket spending.

4.2 Non-Financial Dimensions

Overall, the results of Section 4.1 suggest that consumers and their agents make much more financially sensible choices in 2017 than in 2015. But insurance choices are not just about financial aspects. There are a variety of other attributes of insurance plans that matter to individuals in making these decisions. Indeed, a key concern with algorithmic decision-support tools is that they will lead agents and consumers to over-emphasize the plan attributes included in the algorithm but under-emphasize the welfare-relevant aspects excluded from the algorithm. In the notation of our model in Section 3.2, financial plan dimensions are the variables X^i_{kj} (observed by agents and included in algorithm) while non-financial dimensions such as brand and network are the variables X^g_{kj} (observed by the agents but not included in the algorithm).

Our results in Table 3, however, show that this does not appear to be the case in the setting we study. Non-financial aspects of preferences continue to be valued - and appear to be more appropriately valued - in 2017. First, the coefficient on network breadth remains similar, if somewhat larger, in 2017 to what it is in 2015, suggesting that this important attribute continues to be weighted heavily despite not being included in the recommendation algorithm. If anything, the increase in magnitude suggests the weight on network is more aligned with enrollee preferences for a broader network, all else equal. This implies that the weights agents and consumers place on unused observables (Θ_g) are clearly non-zero on this important domain.

Second, brand preferences for insurers who generally offer similar broad network PPO products are (i) lower in magnitude in 2017 relative to 2015 and (ii) are more similar to each other in 2017 relative to 2015. In particular, the 'branded' fee-for-service carriers, Blue Cross Blue Shield and United, are no longer much preferred to substantively similar but potentially less well known regional carriers in 2017, even though they were strongly preferred in 2015. It is, of course, possible that consumers are over-weighting financial characteristics relative to brand under decision-support, if one thinks that the brands provide substantive value relative to one another. However, under the hypothesis that similarly structured plans provide similar value, it seems clear that algorithmic decision-support reduces the emphasis on brands and increases the emphasis on financial value.

Another test of whether the algorithm moves consumers away from valuable brands is how preferences for Kaiser Permanente change with the introduction of AI-based decision support. Unlike the brands mentioned above, Kaiser has a separate network and an integrated care delivery model that differentiate its care substantively from other insurers.[19]

In 2015 there are high brand preferences estimated for Kaiser plans, an order of magnitude higher than for other brands. Once decision support is introduced in 2017, Kaiser brand preferences

[19] KP offers MA-PD plans that are available in a limited set of markets where KP operates. KP is a vertically integrated, closed network managed care plan. Despite the limited choice, KP has been demonstrated to provide high quality health care (see, e.g., McHugh et al. (2016)). These preferences, however, were not available in the algorithm and due to the financial structure of many of the plan offerings KP typically had a relatively high expected OOP cost as well as a low plan score.

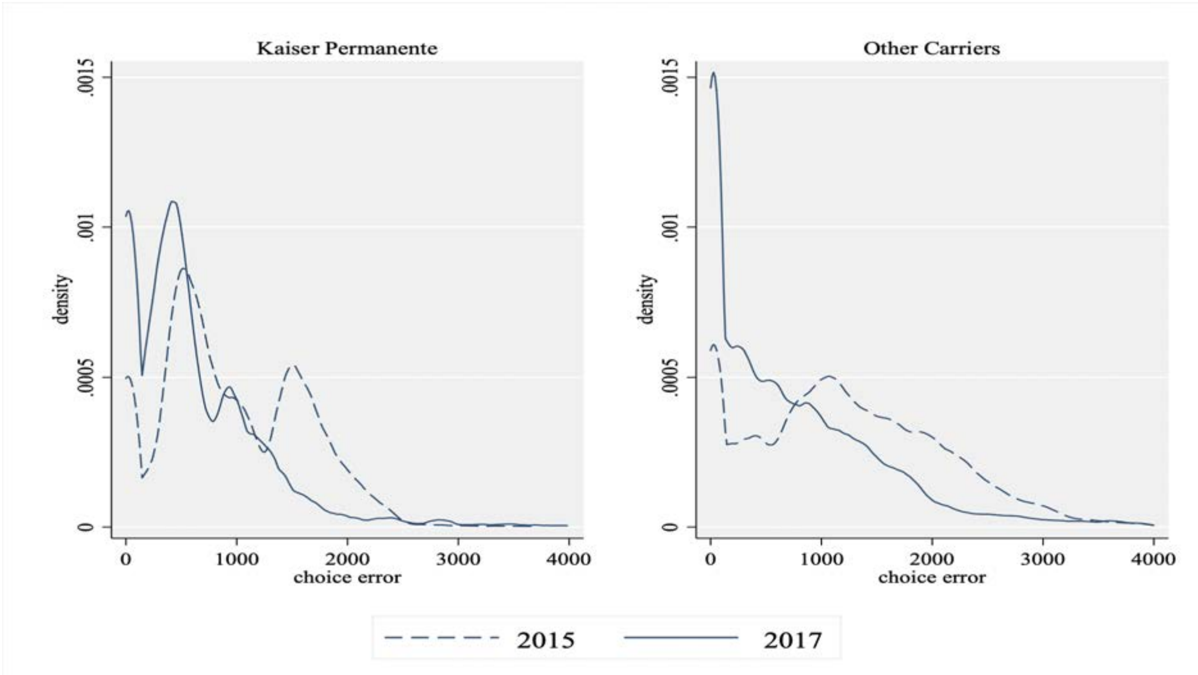

Figure 3: Choice Error by Plan Enrollment (Kaiser Permanente vs. Other) in 2015 and 2017

are reduced, but still high in value, reflecting the fact that agents and consumers do not blindly follow the decision support and pick one of the top recommended options if the agent/consumer has strong preferences for the non-financial aspects of Kaiser plans.

Without decision support in 2015, estimated brand preferences could reflect 'true' preferences that are above and beyond the expected financial outcomes under each plan. Alternatively, they could reflect agent and/or consumer biases and heuristics that can be overcome with sophisticated decision support. We investigate this in more depth by assessing whether or not the marginal switchers away from Kaiser are at the bottom end of the distribution for Kaiser plan financial value, as one would expect if agents are combining the algorithmic recommendations and information about consumer preferences for Kaiser in a sophisticated manner.

The left panel of Figure 3 presents the distribution of monetary loss for those enrollees choosing Kaiser in 2015 and those choosing Kaiser in 2017. The right panel of the figure does the same for all other plans offered. In 2015 we see that both Kaiser enrollees and those in the PPO plans leave meaningful sums on the table. However, the Kaiser distribution is different in that it has two masses at both \$750 and again at \$1600.

When decision support is integrated in 2017, the monetary loss for PPO plans uniformly shifts to the left, reducing the foregone savings associated with choosing these relatively homogeneous plans. For Kaiser, on the other hand, the large mass at \$1600 has disappeared, but there remains a sizeable mass at a valuation of around \$500. That is, many individuals are willing to forgo significant savings to choose Kaiser, but those who were leaving the most on the table have been dissuaded. Indeed, among those who have a monetary loss of more than \$1000 from choosing Kaiser, 61% do so in 2015 - but only 25% do so in 2017.

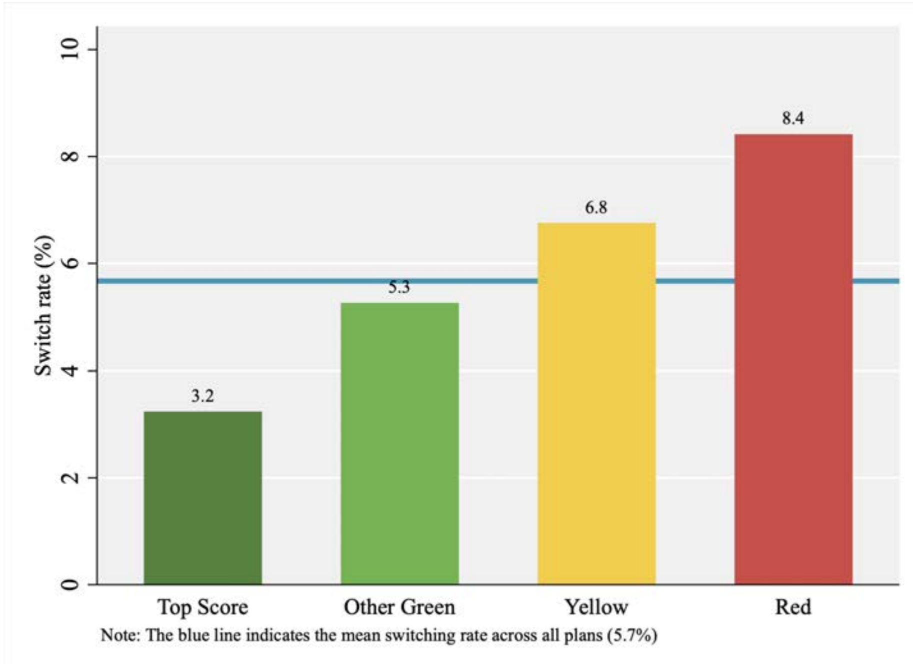

Figure 4: 2018 switching rates by 2017 plan score

Taken together, this evidence suggests that (i) an agent/consumer pair is willing to overrule the algorithm if there is a meaningful consumer preference for Kaiser and (ii) the willingness of an agent/consumer pair to overrule the algorithm's recommendation depends on the magnitude of the loss — the cost of overruling. This is consistent with agents who integrate their information excluded from the algorithm (Kaiser brand preferences) with AI-based recommendations (cost error) efficiently by trading off at the margin.

Overall, the choice model estimates suggest that agents / consumers continue to value non-financial plan attributes that are likely welfare relevant, such as network breadth and the Kaiser delivery model, but do not continue to value non-financial aspects that are likely not welfare-relevant, e.g. the brands of relatively similar broad network PPO carriers.

4.3 Impact on Enrollee Experience

While our choice model estimates are strongly suggestive of the positive impacts that AI-based decision-support has on plan choices, they are still ex ante measures of revealed preference rather than ex post measures of experienced utility. In other contexts where we have both ex-ante and ex-post enrollee spending, we find that our ex-ante model of mistakes is strongly correlated with ex-post outcomes.[20] Nevertheless, in this section we focus on the impact of decision-support on subsequent plan turnover, a key ex post measure of enrollee experience.

If decision support provides people with valuable information that allows them to make better plan choices, we expect to see lower turnover rates among people who enrolled in recommended

[20] See discussion of the model performance in section A. We show that in a non-Medicare population over 55 Picwell's cost predictions perform very well.

plans. The effect is apparent in Figure 4, which shows 2018 switching rates for people who enrolled in MAPD plans in 2017, based on the Score of their 2017 plan. Only 3.2% of people who enrolled in the top scoring plan in 2017 switched plans in 2018, compared to 6.8% and 8.4% of people who enrolled in Yellow (scores between 75 and 89) and Red (scores below 75) plans, respectively. If we look at switching among those who enrolled in a Green (scores of 90 or higher) plan we see switch rates of 4.0% compared to 7.1% among enrollees in lower color tier plans. Taken together, those following the AI recommendations were meaningfully less likely to switch than those who chose poorly rated plans.

One concern in interpreting these results is the endogeneity of the decision to take the AI-based recommendation. To address this issue we develop an IV strategy that takes advantage of random assignment of enrollees to agents. To do this, we estimate heterogeneity in predicted agent plan score and then investigate whether plan turnover directly relates to agent-specific fixed effects. This approach is comparable to a 'judges design' (see, e.g., Kleinberg et al. (2017)), and allows us to isolate the impact of a plan that would be scored higher by the AI-based tool solely due to being randomly assigned to a particular agent.

For our first-stage of this analysis we use the following fixed effects specification:

$$PlanScore_{kj} = Age_k + Female_k + Cost_k + Agent_b + e_{kj} \tag{11}$$

Recall from our discussion in Section 2 that plan score is a measure of plan financial value ranging from 0 to 100, with 100 being the top end of the scale and 0 being the low end. Age_k and $Female_k$ indicate the age at the time of enrollment and whether individual k is female. We also include $Cost_k$, which assigns individuals to one of five quintiles based on predicted costs, in order to account for the possibility that higher cost individuals have more complicated cases. We recover the fixed effects for each agent indexed by b. [21]

Figure 5 presents the results of this first-stage, showing some meaningful variation in predicted 2017 Picwell score, which has a mean in this sample of approximately 84 and a standard deviation of approximately 3 (which is reasonably large since most scores are concentrated near the top end of the range 0 to 100). Thus, being assigned to agents one standard deviation apart from each other means that, on average, your expected plan score falls by 3 percentage points.

We then instrument for the enrolled plan score using the agent fixed effect estimate. Table 5 presents IV estimates for the impact of a higher quality plan according to the AI tool on subsequent turnover. For an agent who is one standard deviation better (3 plan score points) than another in terms of predicted plan score, consumers are 7 percentage points less likely to switch plans the following year. This large effect is equivalent to nearly the full propensity to switch plans in the overall sample. This shows that customers who are randomly assigned to agents who recommend higher ranking plans are more likely to stick with those plans for multiple years, a strong indication that they are better off in those plans and that the tool is leading to better ex-post outcomes.

[21] We perform this analysis only for agents with 20+ customers in 2017, to reduce concerns about statistical noise with the estimated fixed effects. See Section 5 for greater detail on our analysis of agent heterogeneity and see Appendix C for greater detail on our statistical mean reversion robustness analysis.

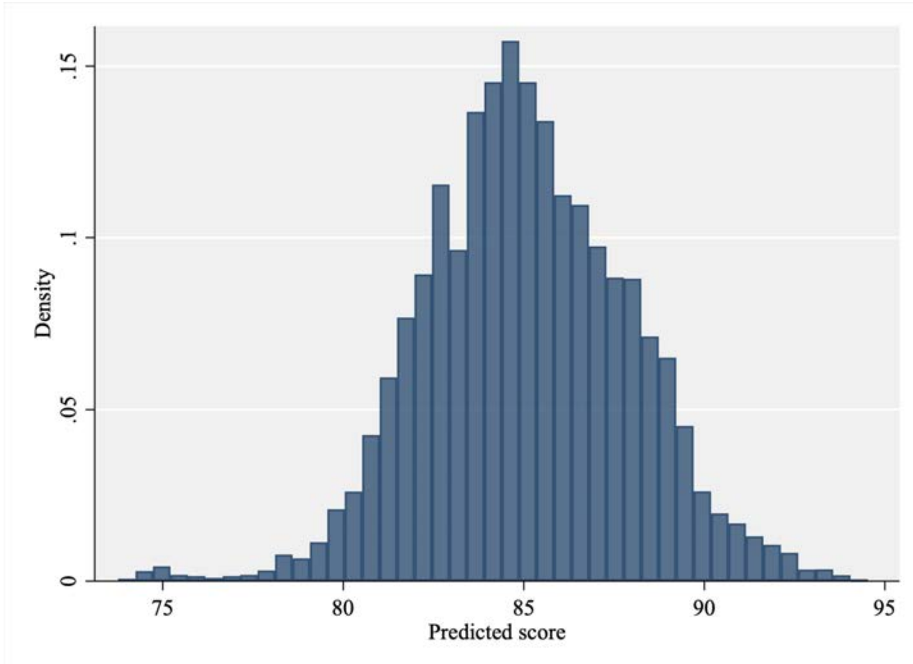

Figure 5: Distribution of predicted plan scores for 2017, across agents who work with over 20+ customers.

The table also presents some additional correlates of switching. Consumers who are 65 at the time of plan choice are more likely to switch the next year (relative to older consumers), as are consumers who chose an Aetna, United or Humana plan (relative to those who chose a regional carrier). Consumers who chose Kaiser were less likely to switch in 2018.

One concern with this analysis is that the agents who are most "tool compliant" in 2017 may be better for other, unobserved, reasons - so that it is not adherence to the tool which is lowering switching, but other aspects of agent behavior. Ideally we would address this by using our IV to predict switching rates for 2015 choices, as revealed in 2016, as a function of the instrument values from 2017. Unfortunately, the weak data available for 2016 means we can't rely on this specification test.[22] Instead, we turn back to our ex-ante measures of foregone savings in Appendix Table 12.

To do so, we restrict ourselves to the set of agents who are in our data for both 2015 and 2017. We then estimate these same fixed effects model for this restricted set of agents in 2017, and use those fixed effects to instrument for ex-ante savings in both 2015 and 2017. Unsurprisingly, we find that those enrollees randomly assigned to agents more likely to follow the tool recommendations in 2017 have lower foregone savings in 2017. But we also show that enrollees using the same 2017-compliant agents when making their choices in 2015 have no differential foregone savings in 2015. If these agents were systematically "better", we would expect it to show up in the ex-ante 2015 measure. Thus, it appears that tool compliance in 2017, and not underlying agent quality, is driving the increased enrollee satisfaction that we see in Table 5.[23]

[22]Although, when we do run the analysis we find no impact of the 2017 IV on switching rates for 2016.

[23]We also perform an analysis, presented in the appendix, where we investigate how the worst agents in 2015

Table 5: Switch IV

	2017 to 2018 Switching (1)
Agent Level Score	-0.0221***
	(-0.0066)
Age Group	
<=65	-
66-70	-0.157***
	(-0.0439)
71-75	-0.210***
	(-0.0476)
76+	-0.206***
	(-0.0443)
Brand	
Regional carrier	-
Aetna	0.837***
	(-0.0569)
Blue	0.430***
	(-0.0622)
Humana	0.183**
	(-0.07)
Kaiser Permanente	-0.272**
	(-0.102)
United	0.225***
	(-0.0672)
Constant	0.175
	(-0.569)
Observations	20,147

Standard errors in parentheses,
* $p<0.05$ ** $p<0.01$ *** $p<0.001$

5 Extensions and Applications

5.1 Agent Heterogeneity

Our results thus far show that (i) skilled agents alone are not sufficient to address choice errors (ii) algorithmic decision support meaningfully improves decisions, especially on the financial dimensions included in the algorithm and (iii) that agents/consumers still consider factors excluded from the algorithm when making recommendations after decision support. In this section we investigate the heterogeneous treatment effects of decision support, with a focus on effects by baseline agent quality. The correlation between these treatment effects and baseline quality bring direct evidence to bear on whether decision support is a complement or substitute for human skill.

We start by presenting estimates on heterogeneity in the quality of agents' recommendations prior to the introduction of AI-based decision support. We estimate the following simple model in 2015, prior to the introduction of decision support:

$$ChoiceError_{kj} = Age_k + Female_k + Cost_k + Agent_b + e_{kj} \tag{12}$$

This model is very similar to the fixed effects specification estimated in equation 11 in Section 4.3: the one difference is the dependent variable is now $ChoiceError_{kj}$, defined as the difference in expected total cost ($premium + E(OOP)$) for individual k who enrolls in plan j relative to the plan in their choice set with predicted lowest cost for that person. Under the assumption of random assignment to enrollees these fixed effects represent the causal "value add" associated with each agent for choice quality. Figure 6 presents the distribution of fixed effect estimates.

The estimates in Figure 6 demonstrate that heterogeneity in agent skill has a material impact on the quality of health insurance choices. We see a mass of average agents but there are also many high performing and low performing agents in terms of their ability to match a particular enrollee to a plan. Moving from the 25th percentile of the distribution to the 75th percentile improves choice quality by \$350 per enrollee per year, on average.

To better understand the nature of this heterogeneity in recommendation quality, we divide agents into quintiles of choice error based on the estimated fixed effects that we show in Figure 6. Table 2 presents agent and customer characteristics by agent quality. As we would expect to see when assignment is random, we do not see significant differences in the mean characteristics of customers across the different agent quality levels. The percent or female customers, age and number of prescriptions are all similar.

Table 6 presents the average choice error and call times for each quintile. Despite the balanced characteristics of consumers across quality quintiles, we see a systematic difference in choice quality

perform in terms of switching rates for 2018. Figure 13 shows that the worst quintiles of agents in 2015 in terms of foregone savings are more likely to have chosen plan with low algorithm plan scores in that year, by a large magnitude. We have shown in this section that from 2017 to 2018, there is much more turnover in the lower score plans. Figures 14 and 15 in the appendix show that (i) the worst performing 2015 agents choose plans of similar scores in 2017 to the best performing 2015 agents and (ii) that conditional on those plan scores, 2018 turnover is similar across the distribution of 2015 quintiles. This analysis also suggests that our switcher IV analysis results do not stem from persistent unobserved heterogeneity in broker performance.

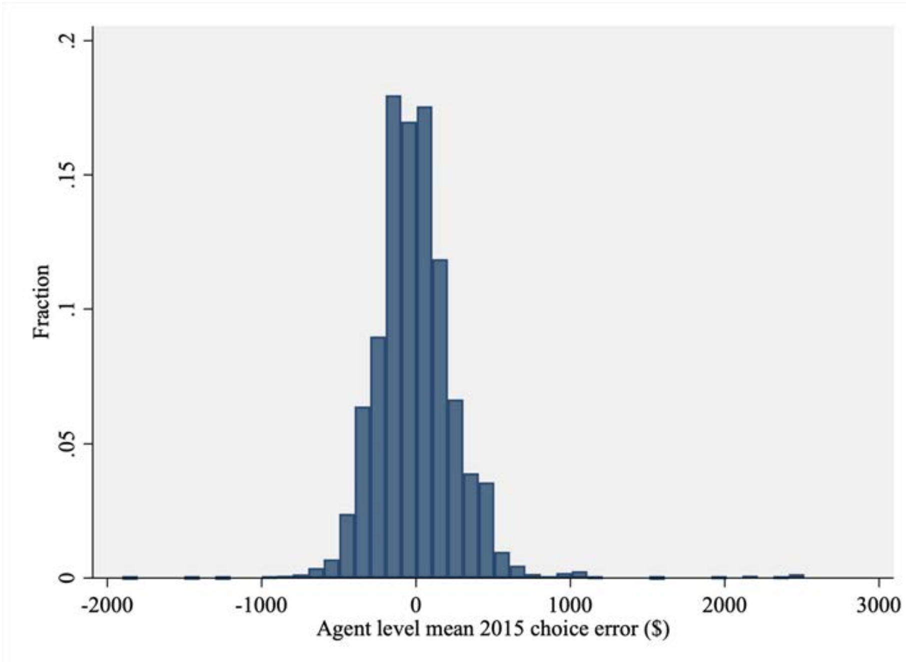

Figure 6: Agent fixed effects in 2015

Table 6: Summary of 2015 agent level choice error, call time and tenure by quality quintile

Agent quality quintile	Choice error Mean	Choice error SD	Call time Mean	Call time SD	Yrs. Experience Mean	Yrs. Experience SD
1	$893	$786	54.3	33.4	4.2	0.80
2	$1,086	$848	53.2	35.3	4.1	0.79
3	$1,213	$1,000	48.6	35.6	4.2	0.83
4	$1,369	$1,237	53.6	35.2	4.1	0.81
5	$1,734	$2,452	56.7	38.1	4.0	0.79

across the quintile groups in 2015. For agents in the top quality quintile, consumers lose an average of $893 per year in their chosen plan relative to the best plan for them in terms of expected financial outcome. The analogous amount for the lowest quintile of agents is $1,734. Thus, the difference in being randomly assigned to a low quality agent as opposed to a high quality agent is almost $1,000 per year in expected spending. Figure 7 presents the kernel density plot of choice error for each quintile group. For all groups we see substantial variance in the choice errors made, but that higher quality agents have choice error distributions that stochastically dominate those of lower quality agents, at each quality level.

Table 6 also shows that (i) agents have the same average tenure across the quality quintiles and that (ii) agents have the same mean call times with customers across the quality quintiles. The former suggests that learning from experience is not a key determinant of quality while the latter shows that agent/consumer call time, one granular measure of agent effort, is not a strong predictor

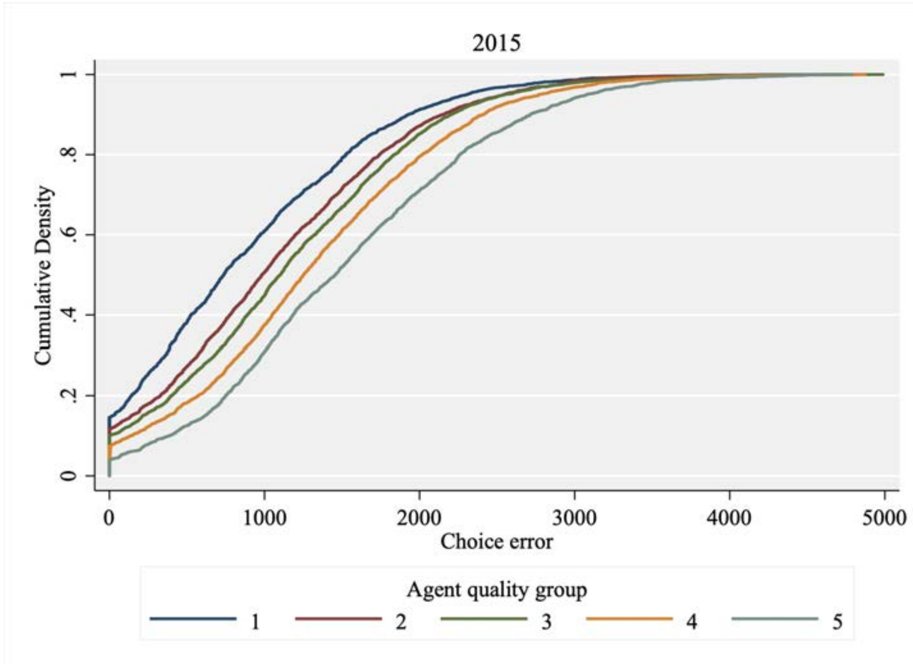

Figure 7: Distribution of 2015 choice error by agent quality quintile

of agent quality.

We use estimates from our structural choice model to assess changes in agent quality over time. To this end, we stratify agents into baseline quintiles using fixed effect estimates from (12) and use the choice model parameters to simulate changes in money left on the table. We simulate 2017 choices using 2015 choice model estimates (including 2015 fixed effects) and compare to 2017 choices using 2017 choice model estimates. Recall that this approach controls for choice set changes over time, which can meaningfully impact the results.

Table 7: Choice error changes by agent quality

agent quality	**metric**	**mean**	**p10**	**p25**	**p50**	**p75**	**p90**
1	2015 Sim. - 2017 Act.	-$55	-$1,543	-$388	$0	$826	$1,547
2	2015 Sim. - 2017 Act.	$118	-$1,210	-$293	$96	$916	$1,598
3	2015 Sim. - 2017 Act.	$133	-$1,232	-$262	$163	$978	$1,639
4	2015 Sim. - 2017 Act.	$305	-$1,008	-$29	$406	$1,128	$1,768
5	2015 Sim. - 2017 Act.	$354	-$996	$0	$501	$1,255	$1,898

Table 7 presents the results from this exercise, including mean and quantile effects by baseline agent quality. The results show that the best agents in 2015 perform roughly the same in 2017, while the worst agents in 2015 perform much better in 2017 (saving an average of $354 per enrollee through improved recommendations). These results reveal (i) meaningful improvement on average

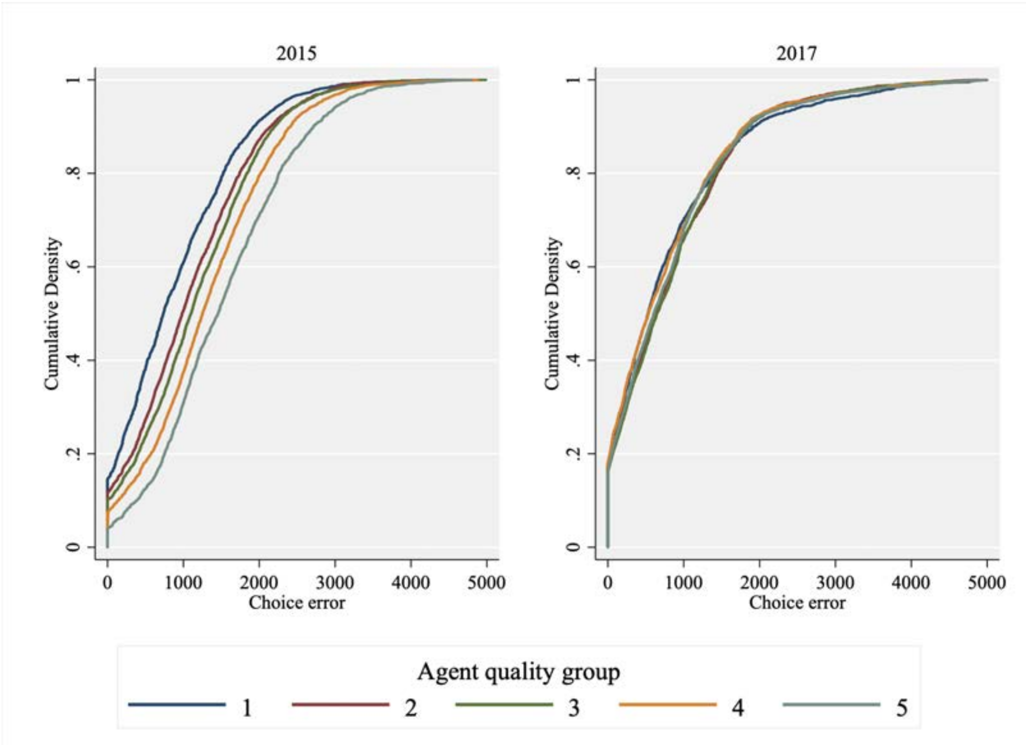

Figure 8: Choice error by agent quality in 2015 and 2017

after decision support and (ii) large improvements at the low end of the quality distribution but no changes at the top of this distribution.[24]

We also present a set of kernel density plots for choices in 2015 and 2017 in Figure 8 to show that the net impact of these differential changes in agent quality is that we see both higher and more homogeneous agent quality in 2017 compared to 2015. The left hand panel replicates Figure 7 above for comparison and the right hand panel presents the same plot for plan choices after the widespread adoption of AI in 2017.

The shift in choice quality is striking. All distributions shift to the left — moving choice error towards zero. There does, however, remain a long tail of observed recommendation errors potentially reflecting underlying enrollee tastes or private information. The similarity of decision quality across the different skill levels in 2017 is also striking when compared to 2015. The entire distribution sits on top of one another for all groups in 2017 but is clearly distinct in 2015.

We dive deeper into the heterogeneous impacts of decision support by investigating how some of the micro-foundations estimated in our choice model change as a function of baseline agent quality. Recall that we found systematic mis-weighting of premium relative to $E(OOP)$ in the demand estimates for 2015 (*See* Table 3). This kind of heuristic decision making was previously attributed, at least implicitly, to consumers choosing and enrolling in plans [see, e.g., Abaluck and

[24] One concern is that agents simply improve over time due to the extra experience incurred between 2015 and 2017. This is unlikely to explain the differential changes to quality. First, agents already had an average of 4 years of prior experience in 2015. Furthermore, we find no significant differences in experience by quality quintile (*See*, Table 6). Thus, as of 2015, there appears to be no relationship between experience and the quality of recommendations.

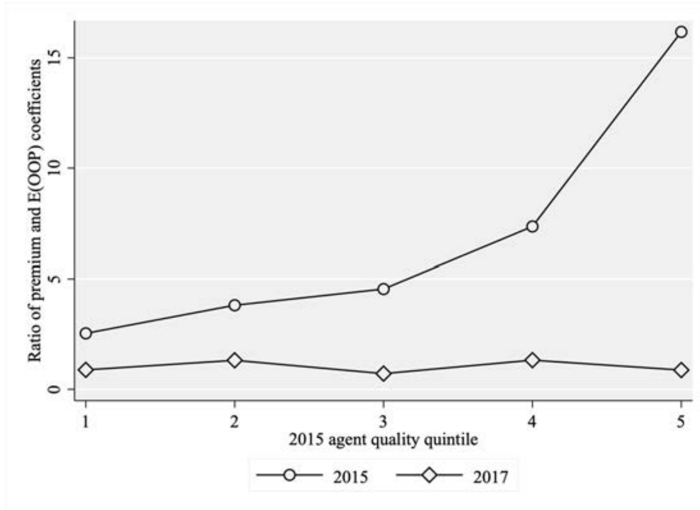

Figure 9: Ratio of coefficient estimates for premium and E(OOP) by year

Gruber (2016) and Abaluck and Gruber (2011)]. Our analysis thus far shows that, even under the guidance of skilled agents, this mis-weighting remains. Here, we ask (i) to what degree does mis-weighting vary by baseline skill and (ii) does offering AI-based decision support correct this mis-weighting to make lower skilled agents look more like their more highly skilled counterparts?

Figure 9 presents the ratio of weights on premium to $E(OOP)$ (rows 1 and 2 in Table 3) by baseline skill level. In 2015 we see a strong relationship between baseline recommendation quality and relative weights on premium and $E(OOP)$. The worst quintile of agent skill in 2015 has an average ratio of 16:1 while the best has a ratio of approximately 2:1. This ratio declines monotonically in baseline skill — agents who mis-weight premium relative to out-of-pocket cost less make better recommendations prior to decision support.

Figure 9 shows that, on the financial dimensions, offering athe AI-based tool harmonizes the recommendations of ex ante different agents. In 2017, all quintiles have ratios near 1:1, reflecting a correct weighting of these financial factors. Introducing AI makes the financial recommendations from ex ante low quality agents similar to the financial recommendations from ex ante high quality agents. Taken together with our earlier results, we find that the use of the tool improves quality across the board but is also a clear substitute for ex ante expertise.

One potential concern with our primary findings about the heterogeneous effects of decision support is mean reversion induced by statistical noise. To address this we re-estimate our primary models only for a subset of agents who have substantially more customers in both years and show the results are unchanged. We also note that mean reversion would suggest symmetry in impact between those who appear to perform well (who should get worse) and those who appear to perform poorly (who should improve). We do not find such symmetry in practice. These results are presented in depth in Appendix C.

Our empirical strategy focuses on brand as an "unused observable" measure of agent weights on enrollee preferences. We have already shown how average brand preferences change as a result

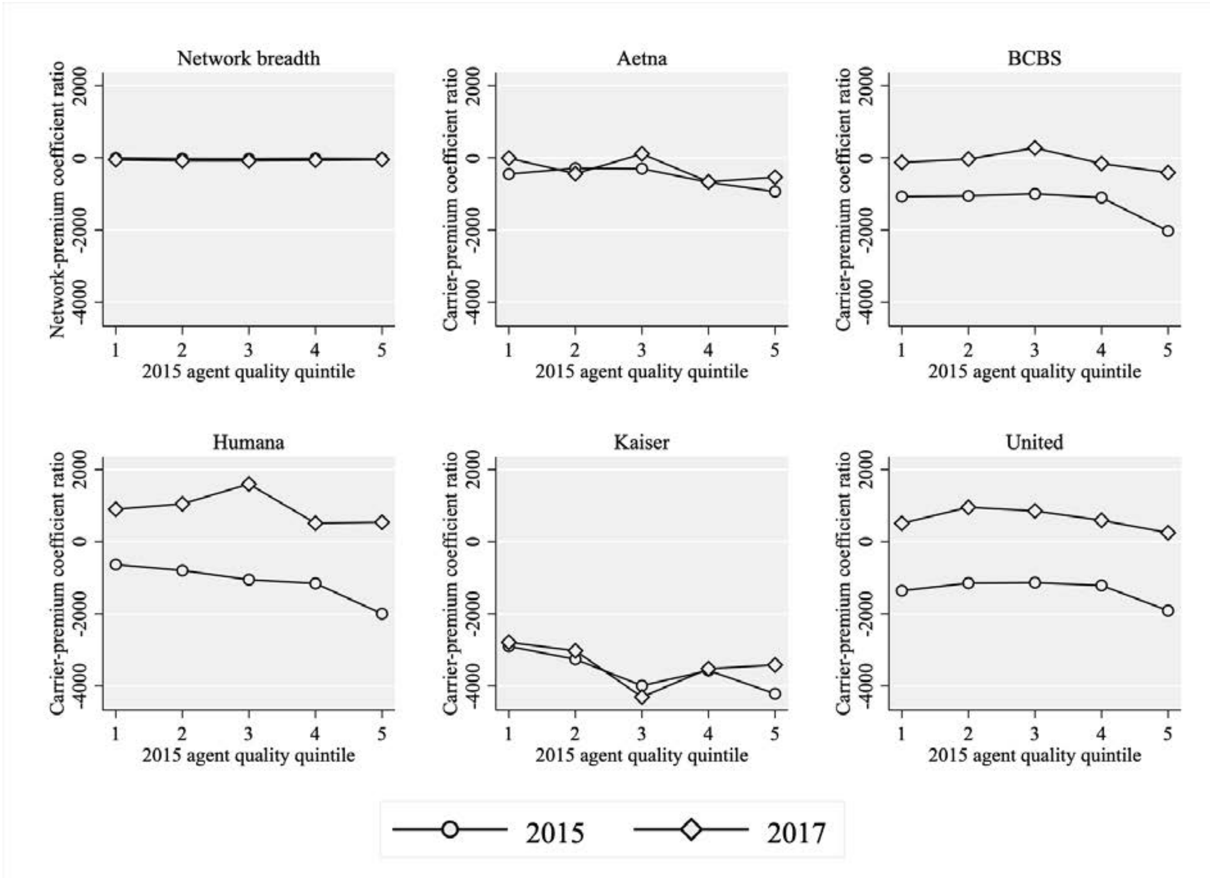

Figure 10: Ratio of coefficient estimates for network, brand preference and premium by year

of decision support. We now turn to how agents heterogeneously emphasize different brands to consumers over time. This is useful both (i) to understand heterogeneous steering and (ii) as a specification check on our primary model, which estimates only a mean brand preference coefficient for each brand.

Figure 10 presents the ratio of the network breadth and different brand preference coefficients to the premium coefficient, as a function of the baseline quality quintile. This figure is similar in spirit to Figure 9 but focuses on the ratio of network and brand to premium preferences, as opposed to the ratio of premium preferences to expected out of pocket cost preferences. The first panel reveals that not only did the average weight on network change little once AI was introduced, this effect was similar across the distribution of agents by baseline quality. Comparing this panel to Figure 9 shows that the agents who improved by the most on the financial aspects of plan choice — the lower quality quintiles — did not differentially change the weight they placed on network breadth.

The remaining 5 panels in Figure 10, each focused on a specific brand demonstrate that, though average brand preferences change after decision support, these changes are not markedly different across baseline quality quintiles, for any of the brands studied. For example, the Kaiser brand preference decreases a little bit relative to the omitted category (regional carries), close to uniformly across the quality quintiles. United, BCBS, and Humana all lose essentially their entire brand advantage relative to the regional carries, across all the quality quintiles. This reveals that, on average, the lower quality agents at baseline were not lower quality due to some clear ex ante bias in favor of certain brands.

Importantly, the fact that shifts in brand choices/preferences occur, for each brand, across the entire distribution of agents suggests that the mean brand preference effects captured in Table 3 do a good job of capturing the change in brand preferences resulting from decision-support roll-out.[25]

5.2 Productivity

We also investigate agent productivity by using agent call times as a measure of agent effort. While call times could reflect a range of underlying agent-driven or consumer-driven inputs, and the correlation between call time and agent quality is ex ante ambiguous, we think it is still instructive to assess (i) how call times change with the introduction of decision support and (ii) how call times correlate with the quality of choices pre- and post-decision support. Call times are also of particular interest as they represent the majority of the marginal cost of enrollment. Therefore, from a producer productivity perspective, call time relative to the quality of the plan chosen is a primary consideration.

Table 8: Average call time by agent quality level

Agent Quality	**2015**	**2017**
Average	53.27	41.90
1	54.27	38.85
2	53.18	43.64
3	48.59	42.68
4	53.63	41.61
5	56.69	42.74

Table 8 shows mean call times in 2015 and 2017 in the population as a whole and by 2015 quality level quintiles. Across the distribution of agent skill, call times are similar within both years. The average 2017 call time for agents in the top quality quintile are 1 to 3 minutes shorter than the average call times for agents in the lower quality quintiles, but, on the whole, call times are quite similar across the quintiles. This pattern was similar in 2015, prior to AI. Quality is not reflected in call time before or after the introduction of AI.

The main impact of AI was to reduce call times uniformly across the distribution. Average call time fell from 53 to 42 minutes, a reduction of 21% compared to the pre-AI average in 2015. This effect comes alongside an overall improvement in the quality of a call — measured by recommendations quality — and a convergence in agent quality after AI is available.

Combining results, AI-based decision support allows the lowest skilled agents to make higher

[25] We also include an exercise in Appendix D that plots the actual brand shares chosen by each agent and compares this with the amount they would have chosen that brand if they randomly chose across all options in their customers' choice sets. This exercise shows that, after decision-support is fully integrated in 2017, the shift in brand shares changes across the entire distribution (high vs. low share of a given brand), i.e. the distribution has a level shift not a shape shift. See Figure 17 in Appendix D for more details on this exercise.

quality recommendations than the ex ante highest skilled agents at significantly lower cost in terms of call time. Put differently, in the cross-section AI appears to be a substitute for skill.

5.3 Adverse Selection

As a last extension, we consider the inherent link between choice adequacy and adverse selection. Adverse selection can significantly reduce welfare in insurance markets, and as emphasized by Handel (2013), reducing choice inconsistencies could potentially worsen adverse selection. Whether the selection effect dominates gains from improved choices depends on a variety of setting specific factors and empirical estimates of the impact vary (e.g. Handel (2013), Polyakova (2016)). Handel et al. (2019) demonstrate the equilibrium welfare impact of choice improvements depend on a set of underlying micro-foundations that include consumer (i) costs, (ii) risk aversion and (iii) choice frictions. However, none of the existing papers actually show how reduced choice errors impact adverse selection in practice.

To understand the impact of AI-based decision support on adverse selection we implement a simple "correlation" test (see, e.g., Chiappori and Salanie (2000) or Einav et al. (2010)). We ask whether improved choices in 2017 lead to an increased correlation between risk and chosen plan generosity. We capture plan generosity in three ways. First, we rank plans by premium. A low premium rank implies a plan that provides less financial protection against out of pocket spending, and vice versa. Second, we consider plans whose premiums are $0, a common benefit design used to attract customers to plans with less generous coverage. Finally, we rank plans by the actuarial value of the plan overall.

The top panel in Figure 11 shows the correlation between health decile and the premium rank of the enrolled plan. We see an increase in the relationship between plan generosity and choice from 2015 to 2017. The average premium rank chosen is similar across health deciles in 2015. In 2017 it is upward sloping, implying a greater correlation between chosen plan generosity and expected health spending. Particularly striking are the results for the lowest cost decile: these healthiest enrollees were actually more likely to choose a high premium plan in 2015, and are much less likely in 2017.

The second panel in the figure shows the correlation between expected health risk and % of consumers who enroll in a $0 premium plan. The third panel shows the correlation between expected health risk and the actuarial value rank of an enrolled plan. Both show evidence of increasing correlation between generosity and the health risk of enrollees choosing a plan. Despite the change in sign of the correlation, the relationship between risk and actuarial value rank is still relatively flat in 2017.

These results present evidence of an increase in adverse selection. However, we do not estimate the welfare impact of adverse selection in this marketplace in this paper for a number of reasons. First, the MA program includes robust risk adjustment, an important counter to adverse selection and complement to friction reducing policies such as AI-based decision support (Handel et al. (2019)). Second, the exchange is only small share of the MA market overall and in any particular

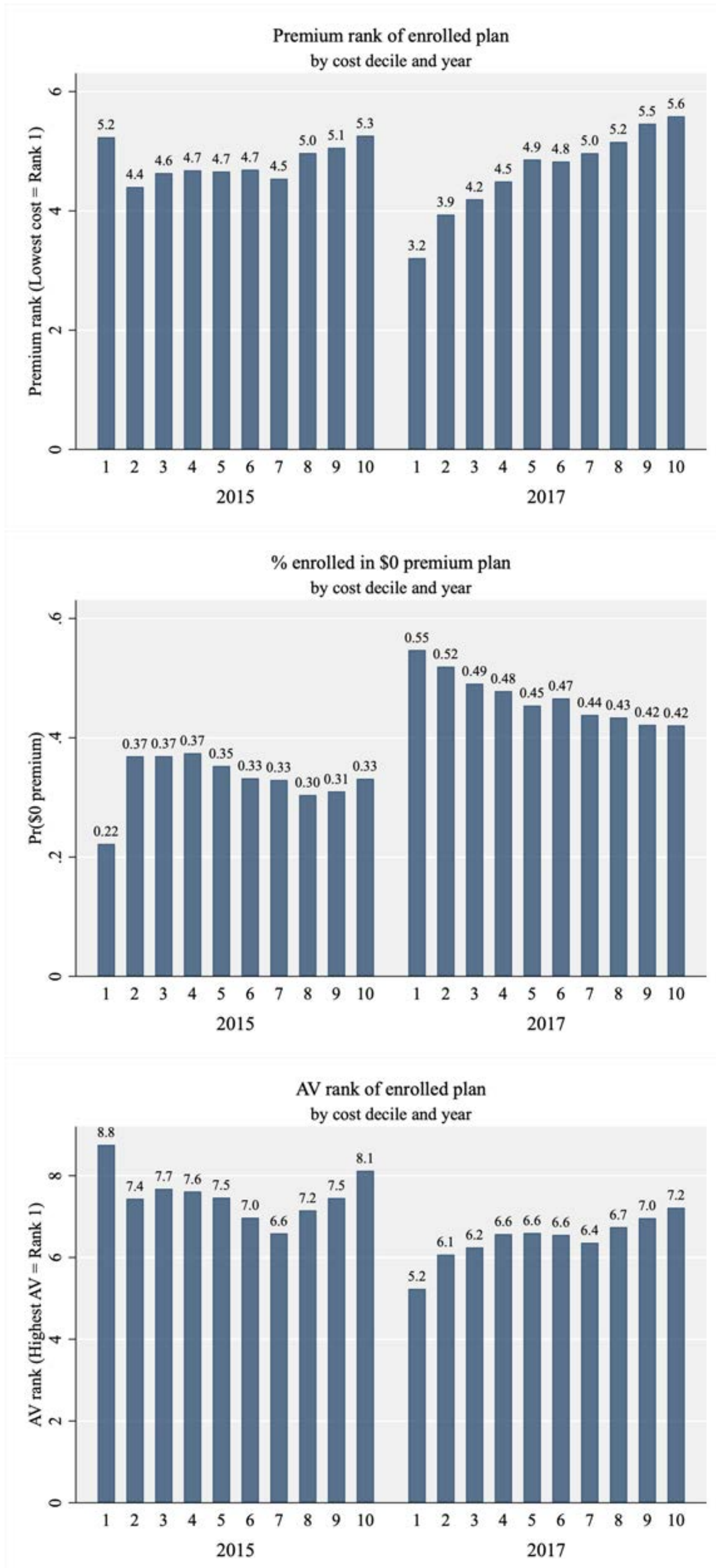

Figure 11: Relationship between choice and cost in 2015 and 2017.

county, the market-level at which prices are set. Even for pricing of plans to the exchange operator specifically, insurers offering insurance product to the operating firm do so across a variety of markets (e.g. Medicare and commercial). Therefore, we do not expect dramatic shifts in plan offerings or pricing due to adverse selection in the MA marketplace alone nor can we easily identify the direct effects in our setting. Third, evaluating the welfare impact of improved choices in a marketplace requires moving beyond static evaluation of adverse selection in which current enrollee marginal cost and willingness-to-pay determine the optimum (e.g. Einav et al. (2010), Hackmann et al. (2015), Handel (2013)). In offering a marketplace to employers who want to cover retirees through the remainder of their lives, the exchange operator needs to provide insurance not only in a current year but against becoming sick: reclassification risk (Handel et al. (2015)). A model of such risk and the associated welfare impacts of information provision and adverse selection therein, while interesting, is beyond the scope of what we study. It does, however, represent an important future avenue for policy relevant research insurance market, particularly as AI-based tools become available.

6 Conclusion

We study insurance choice on a Medicare Advantage exchange platform where (i) consumers receive advice from agents, (ii) consumers are randomized to agents advising them and (iii) the platform fully integrated an agent-facing decision support tool over time. At baseline, we found that jointly made agent-consumer choices leave a lot of money on the table and exhibit many of the same biases found more broadly in the insurance choice literature, e.g. an emphasis on premiums at the expense of the more complicated to evaluate expected out-of-pocket spending. We find that the introduction of AI-based decision support improves decisions greatly on financial dimensions that the AI is well suited to address, essentially removing the financial biases found at baseline. Importantly, we also show that (i) agents continue to integrate non-financial dimensions that are excluded from the algorithm into choices in a sophisticated manner and (ii) that consumers have better plan experiences after the introduction of AI-based decision support, as evidenced by reduced plan switching rates for enrollees randomly assigned to more tool-compliant agents and, therefore, choosing plans predicted by the AI tool to be a better match.

We also investigate agent heterogeneity, and find that the top performing agents at baseline are helped a little bit by the algorithmic support but that the poorest performing agents are helped substantially, bringing their performance up to, and even slightly above, the level of the top agents at baseline.

We take several key lessons away from these results. First, in the health insurance literature specifically, quite a few studies have investigated consumer choices and interventions to improve those choices (Chandra et al. (2018)). Yet, there has been surprisingly little evidence of interventions that markedly improve choice quality. In our study, choice quality does improve meaningfully. One key difference between our study and prior studies is that we combine two interventions (i) expert advising and (ii) sophisticated algorithmic decision support. One implication may be that multiple simultaneous interventions are needed to help consumers make better choices in this kind of market.

Second, our results show that incorporating sophisticated AI into expert advising greatly improves advising on the dimensions included in the decision support while experts continue to incorporate dimensions not included in the decision support that are valued by consumers. It is reassuring that joint agent-consumer decisions do not blindly follow the decision support, but instead integrate it in a fairly nuanced way with information that is excluded from the algorithm.

Third, our results show that sophisticated decision support is a substitute for agent quality. We document substantial baseline agent heterogeneity but show that, once sophisticated decision support is heavily used, this heterogeneity decreases substantially and the ex ante poorer performing agents look a lot like, and even a little better than, the ex ante best performing agents did at baseline. Ultimately, though this is only one example, this is consistent with the hypothesis that greater artificial intelligence capital can substitute for labor expertise and, potentially, lead to a lower wage premium for skilled labor.

Moving forward, there are many interesting angles to assess. What relationships between sophisticated decision support and actual decisions would we find in other markets where advising is omnipresent, such as the financial or investment advice? What features of a choice environment make algorithmic support more or less effective relative to the human advising element? Are decision support and expert advice generally complements, in the sense that decision support improves choices on the dimensions included, but doesn't preclude use of excluded dimensions? It will also be interesting to assess the broader equilibrium effects of sophisticated decision support. We briefly discussed how decision support seems to slightly increase adverse selection in our setting. Moreover, one could be concerned that, if decision support became especially prominent, that firms would respond to that by offering products that look best to the algorithm, at the expense of excluded dimensions. The ability of algorithms and advising to jointly integrate heterogeneous preferences and dimensions excluded from decision support are likely crucial determinants of whether such support helps markets function more fluidly or, instead, leads to market capture by firms that game the algorithm.

References

Abaluck, Jason and Jon Gruber, "Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicare Part D Program," *American Economic Review*, 2011, *101* (4), 1180–1210.

_ **and Jonathan Gruber**, "Improving the Quality of Choices in Health Insurance Markets," December 2016. NBER Working Paper No. 22917.

_ **and** _ , "Evolving choice inconsistencies in choice of prescription drug insurance," *American Economic Review*, 2017, *106* (8), 2145–84.

Acemoglu, Daron and Pascual Restrepo, "Automation and new tasks: how technology displaces and reinstates labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press, 2018.

Athey, Susan, Kevin Bryan, and Joshua S Gans, "The Allocation of Decision Authority to Human and Artificial Intelligence," *Available at SSRN*, 2020.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor, "Choose to lose: Health plan choices from a menu with dominated option," *The Quarterly Journal of Economics*, 2017, *132* (3), 1319–1372.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston, "How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program," *American Economic Review*, 2014, *104* (10), 3335–64.

Bundorf, Kate, Maria Polyakova, and Ming Tai-Seale, "How do Humans Interact with Algorithms? Experimental Evidence from Health Insurance," Technical Report, National Bureau of Economic Research 2019.

Chandra, Amitabh, Benjamin Handel, and Josh Schwartzstein, "Behavioral Economics and Health Care Markets," August 2018. UC Berkeley working paper.

Chiappori, Pierre and Bernard Salanie, "Testing for Asymmetric Information in Insurance Markets," *Journal of Political Economy*, 2000, *108*, 56–78.

Diamond, Peter A, "A model of price adjustment," *Journal of economic theory*, 1971, *3* (2), 156–168.

Dranove, David and Mark A Satterthwaite, "Monopolistic competition when price and quality are imperfectly observable," *The RAND Journal of Economics*, 1992, pp. 518–534.

Egan, Mark, "Brokers versus Retail Investors: Conflicting Interests and Dominated Products," *The Journal of Finance*, 2019, *74* (3), 1217–1260.

_ , Gregor Matvos, and Amit Seru, "The Market for Financial Adviser Misconduct," *Journal of Political Economy*, 2019, *127* (1), 233–295.

Einav, Liran, Amy Finkelstein, and Jon Levin, "Beyond Testing: Empirical Models of Insurance Markets," *Annual Review of Economics*, 2010, *2*, 311–336.

Ericson, Keith, "Market Design When Firms Interact with Inertial Consumers: Evidence from Medicare Part D," *American Economic Journal: Economic Policy*, 2014, *6* (1), 38–64.

Fang, Hanming, Michael Keane, and Dan Silverman, "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market," *Journal of Political Economy*, 2008, *116* (2), 303–350.

Finkelstein, Amy and James Poterba, "Testing for asymmetric information using "unused observables" in insurance markets: Evidence from the UK annuity market," *Journal of Risk and Insurance*, 2014, *81* (4), 709–734.

Gambacorta, Leonardo, Luigi Guiso, Paolo Mistrulli, Andrea Pozzi, and Anton Tsoy, "The Cost of Distorted Financial Advice - Evidence from the Mortgage Market," EIEF Working Papers Series 1713, Einaudi Institute for Economics and Finance (EIEF) 2017.

Geruso, Michael and Timothy Layton, "Upcoding or selection? Evidence from Medicare on squishy risk adjustment," *NBER Working Paper*, 2015, *21222.*

Hackmann, Martin, Jonathan Kolstad, and Amanda Kowalski, "Adverse Selection and an Individual Mandate: When Theory Meets Practice," *American Economic Review*, 2015, *105* (3), 1030–1060.

Handel, Benjamin, "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 2013, *103* (7), 2643–2682.

_ and Jonathan Kolstad, "Getting the Most from Marketplaces: Smart Policies on Health Insurance Choice," *Hamilton Project Discussion Paper*, 2015.

_ and _ , "Health Insurance For Humans: Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 2015, *105* (8), 2449–2500.

_ and Joshua Schwartzstein, "Frictions or mental gaps: what's behind the information we (don't) use and when do we care?," *Journal of Economic Perspectives*, 2018, *32* (1), 155–78.

_ , Igal Hendel, and Michael Whinston, "Equilibria in Health Exchanges: Adverse Selection vs. Reclassification Risk," *Econometrica*, 2015, *83* (4), 1261–1313.

_ , Jonathan Kolstad, and Johannes Spinnewijn, "Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets," *Review of Economics and Statistics*, 2019, *2* (101), 326–340.

Heiss, Florian, Daniel McFadden, Joachim Winter, Amelie Wupperman, and Bo Zhou, "Mind the Gap! Consumer Perceptions and Choices of Medicare Part D Prescription Drug Plans," *Research Findings in the Economics of Aging*, 2010, pp. 413–481.

Ho, Kate, Joseph Hogan, and Fiona Scott Morton, "The Impact of Consumer Inattention on Pricing in the Medicare Part D Program," *RAND Journal of Economics*, 2017, *48* (4), 877–905.

Karaca-Mandic, Pinar, Roger Feldman, and Peter Graven, "The role of agents and brokers in the market for health insurance," *Journal of Risk and Insurance*, 2018, *85* (1), 7–34.

Ketcham, Jonathan, Claudio Lucarelli, Eugenio Miravete, and Christopher Roebuck, "Sinking, Swimming, or Learning to Swim in Medicare Part D?," *The American Economic Review*, 2012, *102* (6), 2639–2673.

— , **Nicholas Kuminoff, and Christopher A. Powers**, "Etimating the Heterogeneous Welfare Effects of Choice Architecture: An Application to the Medicare Prescription Drug Insurance Market," October 2016. NBER Working Paper No. 22732.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*, 08 2017, *133* (1), 237–293.

Kling, Jeffrey, Sendhil Mullainathan, Eldar Shafir, Lee Vermeulen, and Marian Wrobel, "Comparison Friction: Experimental Evidence from Medicare Drug Plans," *Quarterly Journal of Economics*, 2012, *127* (1), 199–235.

Lacetera, Nicola, Bradley Larsen, Devin Pope, and Justin Sydnor, "Bid Takers or Market Makers? The Effect of Auctioneers on Auction Outcomes," *American Economic Journal: Microeconomics*, 2016, *8* (4), 195–229.

Mackowiak, Bartosz, Filip Matejka, and Mirko Wiederholt, "Rational inattention: A disciplined behavioral model," *Goethe University Frankfurt mimeo*, 2018.

McHugh, M, L Aiken, M Eckenhoff, and L Burns, "Achieving Kaiser Permanente Quality," *Health Care Manage Review*, 2016, *3* (41), 178–188.

Mullainathan, Sendhil, Markus Noeth, and Antoinette Schoar, "The market for financial advice: An audit study," Technical Report, National Bureau of Economic Research 2012.

Newhouse, Joseph P, Mary Price, Jie Huang, J Michael McWilliams, and John Hsu, "Steps to reduce favorable risk selection in Medicare Advantage largely succeeded, boding well for health insurance exchanges," *Health Affairs*, 2012, *31* (12), 2618–2628.

Polyakova, Maria, "Regulation of Insurance with Adverse Selection and Switching Costs: Evidence from Medicare Part D," *American Economic Journal: Applied Economics*, 2016, *8* (3), 165–195.

Starc, Amanda and Robert J Town, "Externalities and benefit design in health insurance," Technical Report, National Bureau of Economic Research 2015.

A Decision Support Performance

The Medicare decision support technology studied here can generate cost predictions based on two different levels of customer data. The base level, predicts cost based on age, sex and prescriptions, and another version of the model predicts cost based on these base characteristics plus responses to utilization survey question that ask users to indicate the number of primary care visits, specialist visits and hospital admissions that they experienced in the preceding 12 months. The base version of the decision support model is what was used to generate recommendations during our study period, but we present out-of-sample (OOS) R^2 (generated using 5-fold cross validation) for both the base and enhanced version to demonstrate how performance changes as more information is available to the model and to allow comparisons with other models that include additional information.

Table 9: R^2 of AI-Based prediction models and CMS HCC models for Medicare non-drug spending

AI Prediction Model	
Age+Sex+Drugs	0.069
Age+Sex+Drugs+Utilization survey	0.105
CMS HCC model V21	
New enrollees: age, sex, disability, Medicaid enrollment	0.019
Non-institutional continuing enrollees: age, sex, disability, Medicaid enrollment, ICD-10 codes/HCCs	0.125

In Table 9, we compare the R^2 of both versions of these Medicare prediction models to different versions of the CMS HCC model. In both cases, we consider $OOSR^2$ for models that predict inpatient and outpatient costs, and exclude prescription drug costs. The base decision support model R^2 is about 3.5 times larger than the CMS HCC model for new enrollees, and the decision support model with additional survey questions explains a similar level of variance to the CMS HCC model for continuing enrollees. These values are low in absolute terms, reflecting the high level of variance in medical spending among the Medicare population, but the performance of the Medicare decision support tool relative to the CMS HCC models indicates that a machine learning based model can use limited and easy to collect information to provide useful information to consumers of health insurance (or their agents).

In an alternative setting — recommendations for employer sponsored insurance where prior claims data are available — the same decision support technology predicts total allowed costs, including inpatient, outpatient and prescription drug costs, with an R^2 of 0.32, which exceeds R^2 values of other models that utilize similar inputs in a recent Society of Actuaries review of risk-scoring models (cite Society of Actuaries, 2016).

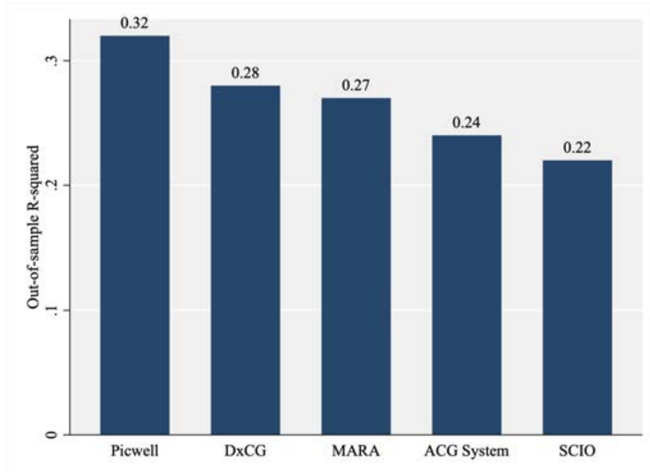

Figure 12: R^2 of health care cost prediction models that use prior claims data

B Additional Specifications

Tables 10 and 11 provide estimates for a version of our primary choice model that censors the top 5% of consumers with the largest estimated choice errors.The results in this specification are similar to those in our primary implementation discussed in the main text. Table 12 presents the specifications for how foregone savings in 2015 and 2017 respectively are associated with 2017 mean broker plan score. This table is discussed in the main text and is consistent with the 2017 instrument for plan quality being associated with 2017 foregone savings but not 2015 foregone savings, implying that 2017 variation in mean plan score, after the widespread adoption of decision support, is not predictive of 2015 performance but is predictive of 2017 performance. This suggests that the plan switch IV regressions in the main text are not the result of unobservable broker heterogeneity that is correlated with their mean plan score.

Figure 13 shows that the worst quintiles of agents in 2015 in terms of foregone savings are more likely to have chosen plan with low algorithm plan scores in that year. We have shown in the main text that from 2017 to 2018, there is much more turnover in the lower score plans. Figures 14 and 15 in show that (i) the worst performing 2015 agents choose plans of similar scores in 2017 to the best performing 2015 agents and (ii) that conditional on those plan scores, 2018 turnover is similar across the distribution of 2015 quintiles. This analysis also suggests that our switcher IV analysis results do not stem from persistent unobserved heterogeneity in broker performance.

Table 10: Choice Model (censored sample)

	(1)		(2)		(3)	
	2015	2017	2015	2017	2015	2017
Premium ($100)	-0.0876***	-0.0903***	-0.0874***	-0.0565***	-0.112***	-0.0838***
	(72.70)	(67.38)	(70.55)	(38.30)	(82.84)	(37.68)
Predicted OOP ($100)	-0.0308***	-0.0873***	-0.0312***	-0.140***	-0.0487***	-0.109***
	(38.93)	(53.12)	(34.79)	(77.24)	(40.12)	(41.80)
Risk Penalty ($100)			0.00354	-0.0777***	0.244***	-0.0455***
			(1.11)	(50.04)	(50.44)	(18.38)
Actuarial Value					-0.0310***	0.00778*
					(12.01)	(2.57)
Deductible ($100)					-0.347***	-0.0433***
					(47.17)	(6.86)
Max OOP ($100)					-0.0429***	-0.0163***
					(60.98)	(13.90)
Network Coverage	0.0130***	0.0279***	0.0130***	0.0305***	0.0191***	0.0311***
	(17.52)	(37.39)	(17.46)	(40.20)	(25.60)	(39.70)
Plan Type Dummies	X	X	X	X	X	X
Bran Dummies	X	X	X	X	X	X
Pseudo R-squared	0.152	0.132	0.152	0.154	0.192	0.156
Observations	363,367	317,701	363,367	317,701	363,367	317,701

t statistics in parentheses
* p<0.05 ** p<0.01 *** p<0.001

Note: This table estimates demand models on a censored sample in which we exclude individuals who are in the top 5% in terms of estimated choice error.

C Mean Reversion

To address the potential issue of mean reversion we consider two different types of analyses. First, we look at some of our key results conditioning on the set of agents who have a large number of consumers in our data. Figure 16 shows the ratio of our estimated premium coefficient to our estimated out-of-pocket spending coefficient conditioning on (i) agents who have more than 20 MA enrollees in each year and (ii) agents who have more than 50 MA enrollees in each year. As the number of enrollees per agent gets bigger, our results with our primary sample continue to hold, namely that (i) across the distribution of baseline quality quintiles this ratio moves to near 1 to 1 and (ii) the worst quality quintiles have much higher ratios in 2015. With 50+ MA enrollees per agent, it is highly unlikely that mean reversion from within-agent statistical noise impacts the results.

Table 11: Plan Type and Brand Coefficients (censored sample)

	(1)		(2)		(3)	
	2015	2017	2015	2017	2015	2017
Plan Type						
HMO	-	-	-	-	-	-
PPO	1.072***	0.904***	1.068***	1.158***	1.248***	1.226***
	(50.02)	(48.56)	(48.42)	(57.25)	(53.61)	(55.62)
Other	-1.711***	-0.824***	-1.714***	-0.775***	-3.200***	-0.734***
	(13.46)	(10.13)	(13.49)	(7.83)	(40.95)	(7.25)
Brand						
Regional Carrier	-	-	-	-	-	-
Aetna	0.859***	0.153***	0.859***	0.261***	0.371***	0.211***
	(26.74)	(6.04)	(26.69)	(10.03)	(10.69)	(7.85)
BlueCross BlueShield	1.122***	-0.0454	1.125***	0.133***	0.963***	0.0837**
	(44.65)	(1.61)	(44.33)	(4.68)	(36.96)	(2.90)
Humana	0.714***	-0.621***	0.710***	-0.280***	0.947***	-0.240***
	(25.63)	(20.12)	(25.46)	(8.68)	(31.84)	(7.24)
Kaiser Permanente	3.292***	1.935***	3.293***	2.275***	3.283***	2.203***
	(100.12)	(41.85)	(100.06)	(48.59)	(83.05)	(44.98)
United	0.618***	-0.299***	0.618***	-0.314***	1.123***	-0.323***
	(19.68)	(10.62)	(19.67)	(11.45)	(33.26)	(11.80)
Pseudo R-squared	0.152	0.132	0.152	0.154	0.192	0.156
Observations	363,367	317,701	363,367	317,701	363,367	317,701

t statistics in parentheses
* p<0.05 ** p<0.01 *** p<0.00.1

Note: This table estimates demand models on a censored sample in which we exclude individuals who are in the top 5% in terms of estimated choice error.

Table 12: 2017 IV: Additional Analysis

	2017 Cost Error	2015 Cost Error
Agent Level Score	-68.73***	6.10
	(6.128)	(6.149)
Age Group		
<=65	-	-
66-70	-43.32	34.84
	(40.52)	(48.66)
71-75	53.27	155.25*
	(43.64)	(53.39)
76+	58.87	212.86***
	(38.87)	(47.12)
Brand		
Regional carrier	-	-
Aetna	-160.86***	436.94***
	(45.72)	(57.95)
Blue	-236.55***	372.23***
	(50.77)	(60.46)
Humana	-238.49***	742.30***
	(52.82)	(56.42)
Kaiser Permanente	-463.83***	-13.34
	(54.32)	(46.42)
United	-207.15***	111.12*
	(50.60)	(53.08)
Constant	6961.19***	323.79
	(524.4)	(534.9)
Observations	10,319	7,977

Standard errors in parentheses
* $p<0.05$ ** $p<0.01$ *** $p<0.001$

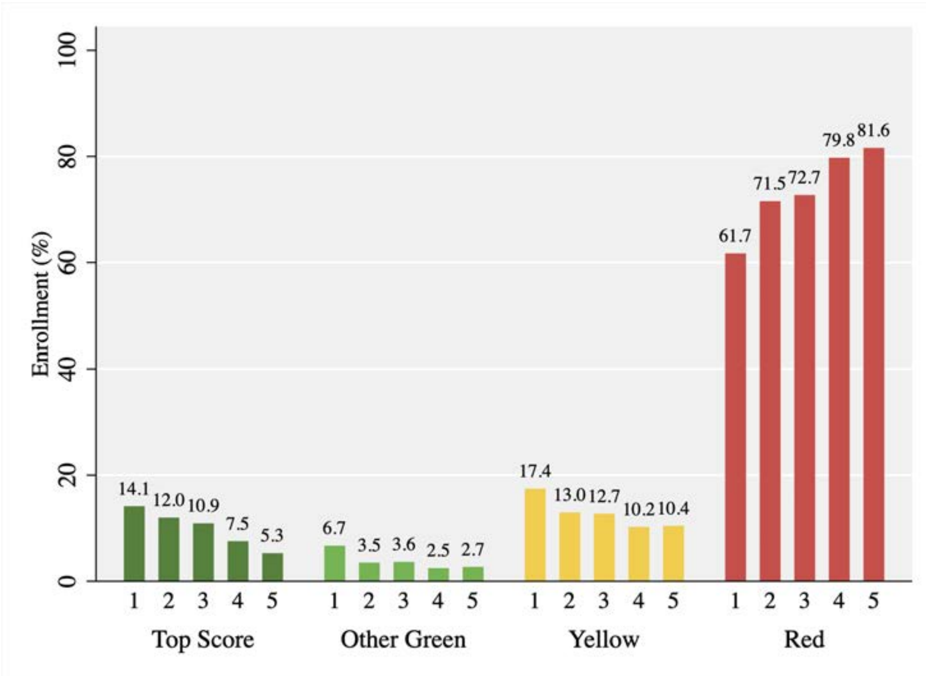

Figure 13: Choice of plan score (categorized by color tier) for 2015, as a function of broker 2015 mean foregone savings quintile (1 is worst, 5 is best).

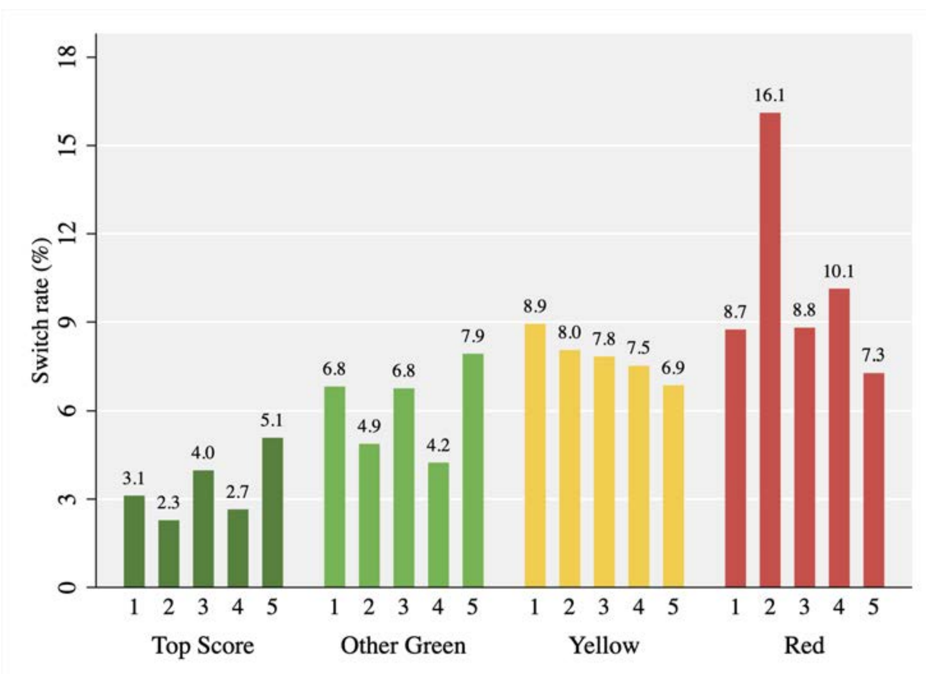

Figure 14: Choice of plan score (categorized by color tier) for 2017, as a function of broker 2015 mean foregone savings quintile (1 is worst, 5 is best).

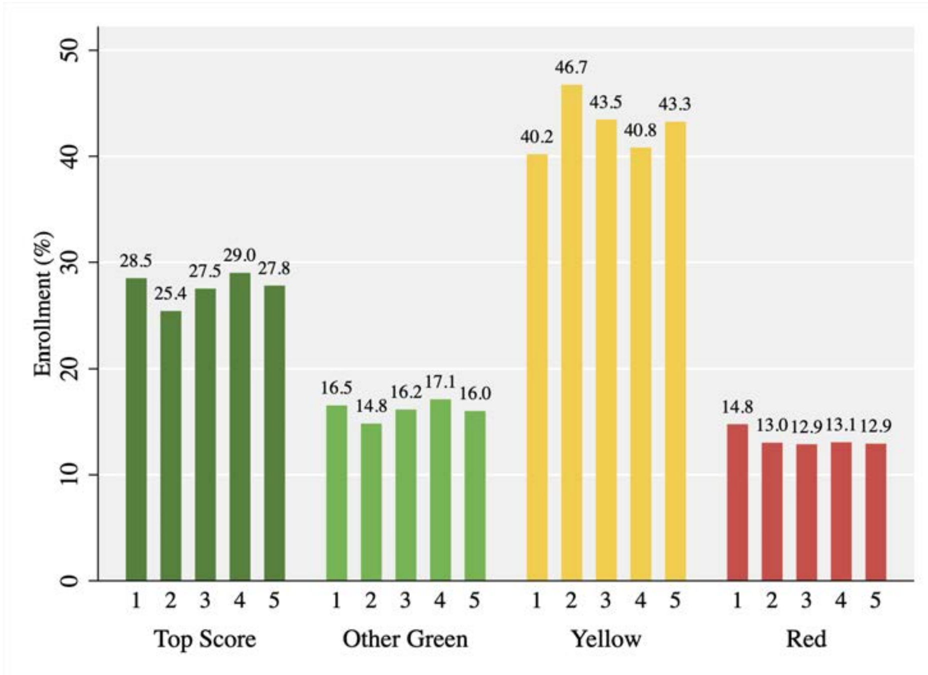

Figure 15: Plan turnover for 2018 as a function of 2015 broker mean foregone savings quintile (1 is worst, 5 is best) and color (score) of plan chosen in 2017.

Second, we note that mean reversion alone would suggest that the better agents in 2015 become worse in 2017, and vice-versa. As shown in Figure 8 (as well as throughout our results) this is not the case. The best agents remain similar in 2017 to 2015, both in terms of average money left on the table and, somewhat, in terms of their premium to expected out-of-pocket choice model coefficients. The worst agents improve markedly on both fronts. Overall, the distributions of money left on the table by baseline quality quintile are close to homogeneous in 2017 and look very similar to the distribution for the top quintile of agents in 2015. Also, importantly, these asymmetric changes by baseline quality quintile look the same when conditioning on agents with more than 50 consumers in each year. Overall, these results show that it is highly unlikely that mean reversion drives our results concerning the heterogeneous impacts of decision support.

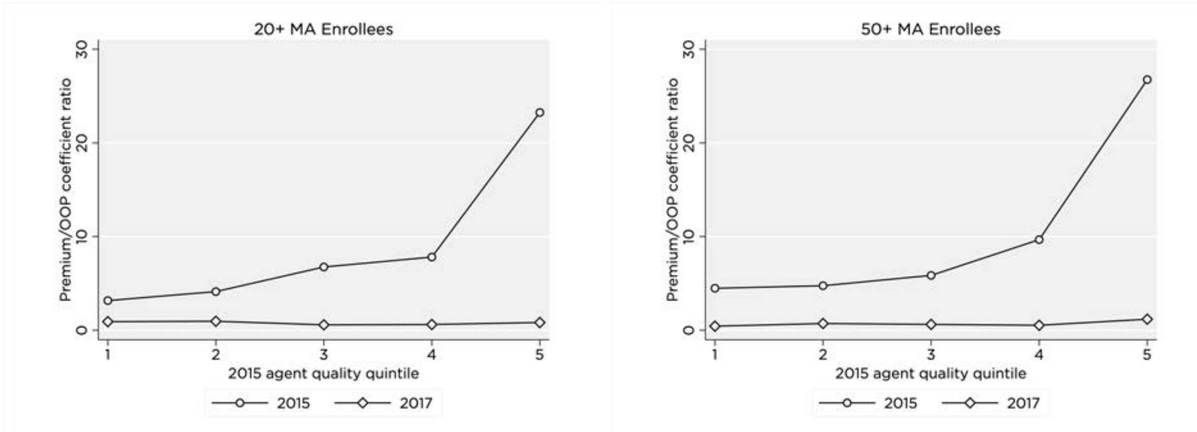

Figure 16: Ratio of coefficient estimates for premium to expected out-of-pocket spending for agents with larger number of MA enrollees.

D Additional Brand Preference Analysis

In the main text we discuss an analysis that compares brand choices over time, pre and post decision support. We compare what share of plans agents actually choose by brand, by year, to what they would have chosen if they randomly chose brands from the choice sets they engage with. To do this exercise, we:

1. Keep top 200 agents by volume in each year to get higher within-agent sample sizes.
2. For each agent in each year, for each brand, only keep consumers that actually had that brand in their choice set when calculating the share for that brand.
3. Compute the share that each agent would have in each brand if choices were made randomly. This controls for choice set size.
4. Subtract the randomly chosen share from the actually chosen share.
5. Plot the histogram of this statistic for all agents in the sample.

Positive values of the statistic for a given brand indicate a positive brand preference, relative to random choice, with negative values indicating a negative brand preference.

Figure 17 plots the results of this exercise for 6 brands, including all regional carriers together as one 'brand.' The results show that the changes to the average brand preferences estimated in our structural choice model come from shifts across the entire distribution of agents, rather than from shifts to specific agents who have very strong preferences for certain carriers. Each brand's histogram reflects a level shift in the distribution to the left or the right, rather than a shift in the shape of the distribution. For example, the Kaiser distribution of shares chosen relative to random choice (i) has a similar shape in 2015 and 2017 (ii) is strongly positive in both years and (iii) is lower in magnitude in 2017 relative to 2015, reflecting the substitution on the margin away from

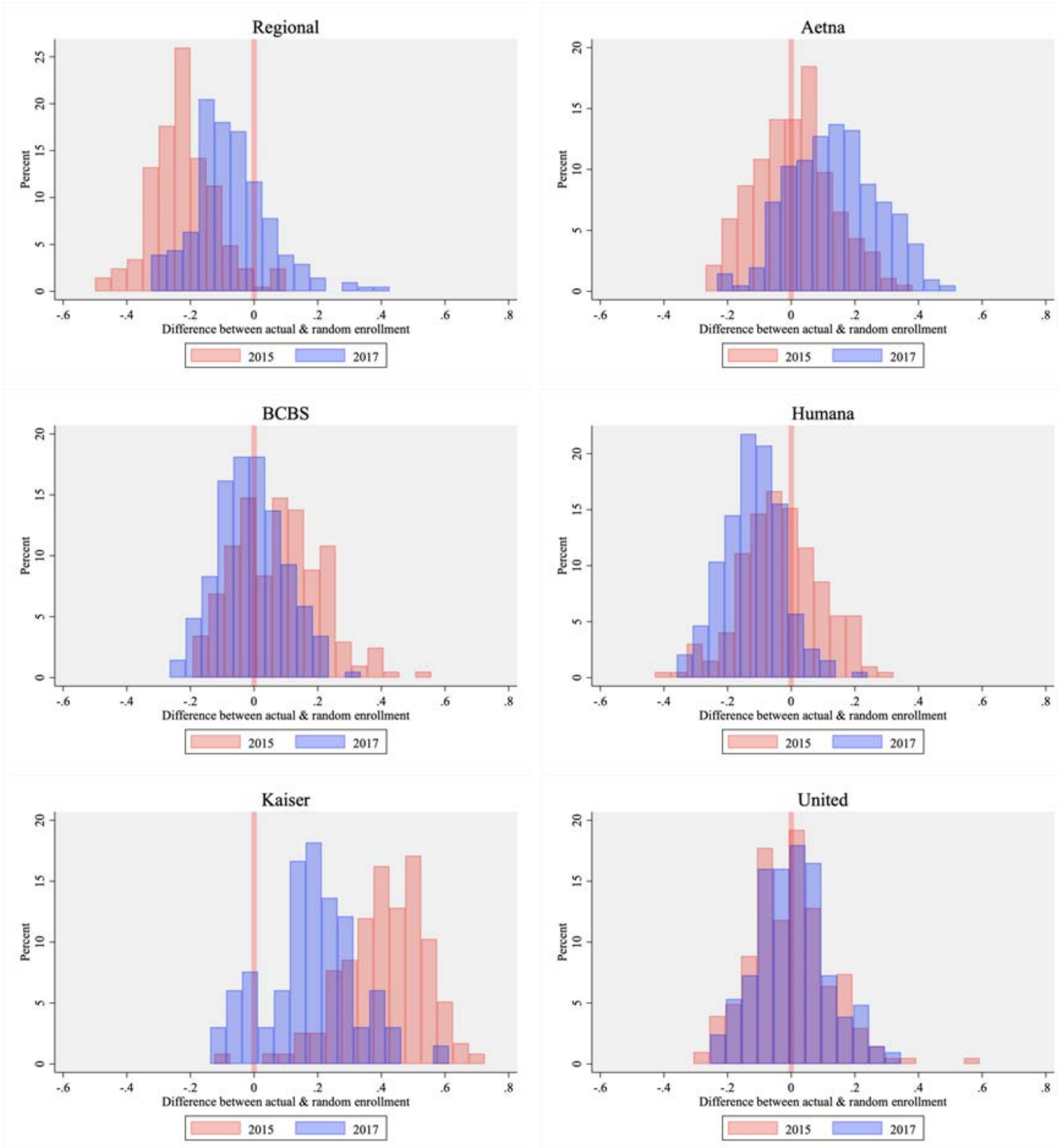

Figure 17: Agent Level Heterogeneity in Brand Choices

Kaiser by the consumers losing the most money from joining Kaiser. The distributions for BCBS and Humana clearly shift to the left, moving from positive on average to negative on average, while regional plans capture a lot of this lost brand equity, moving from quite negative in 2015 to near 0 in 2017. Thus, once decision support is used, the additional financial benefits from regional carriers as opposed to national carriers overcomes some of the brand effects for similar options broad network PPO options that had been present in the market. Importantly, the fact that that the shifts occur across the entire distribution of brand preferences for each brand suggests that the mean brand preference effects captured in Table 3 do a good job of

E Measurement Error

Table 13 reports the coefficients from the measurement error analysis. See Section 4.1.1 for a discussion of our approach to assessing measurement error and the implications of our results.

Table 13: Measurement Error Analyses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Rounding		**Noise (by Standard Deviation)**				
	Picwell OOP	$500	$1,000	$200	$500	$1,000	$2,000	$3,000
Annual Premium ($100)	-0.0980	-0.0975	-0.0929	-0.0971	-0.0949	-0.0883	-0.0784	-0.0724
	(0.00177)	(0.00177)	(0.00173)	(0.00176)	(0.00173)	(0.00169)	(0.00158)	(0.00156)
Predicted OOP ($100)	-0.102	-0.0984	-0.0994	-0.0977	-0.0943	-0.0808	-0.0569	-0.0413
	(0.00175)	(0.00178)	(0.00177)	(0.00176)	(0.00172)	(0.00161)	(0.00136)	(0.00120)
Deductible ($100)	0.00465	0.00808	0.00169	0.00653	0.00130	-0.0115	-0.0139	-0.00747
	(0.00605)	(0.00601)	(0.00602)	(0.00603)	(0.00600)	(0.00600)	(0.00595)	(0.00593)
Max OOP ($100)	-0.000719	-0.00143	-0.000680	-0.00130	-0.000686	-0.00129	-0.000743	-0.00138
	(0.000750)	(0.000747)	(0.000757)	(0.000750)	(0.000748)	(0.000749)	(0.000736)	(0.000745)
Risk Penalty	0.205	0.202	0.206	0.208	0.202	0.196	0.178	0.165
	(0.00481)	(0.00485)	(0.00489)	(0.00487)	(0.00481)	(0.00480)	(0.00474)	(0.00473)
Actuarial Value	-0.0165	-0.0149	-0.0161	-0.0147	-0.0117	-0.0144	-0.0153	-0.0102
	(0.00249)	(0.00248)	(0.00250)	(0.00248)	(0.00246)	(0.00247)	(0.00241)	(0.00239)
Network Coverage	0.0187	0.0187	0.0185	0.0181	0.0183	0.0178	0.0173	0.0157
	(0.000736)	(0.000737)	(0.000740)	(0.000735)	(0.000733)	(0.000723)	(0.000707)	(0.000696)
Plan Type Dummies								
HMO	-	-	-	-	-	-	-	-
PPO	-3.115	-3.039	-3.101	-3.043	-2.947	-2.794	-2.584	-2.433
	(0.0357)	(0.0348)	(0.0353)	(0.0347)	(0.0336)	(0.0317)	(0.0291)	(0.0278)
Other	-0.0899	-0.119	-0.195	-0.162	-0.0641	-0.141	-0.107	-0.128
	(0.0627)	(0.0623)	(0.0632)	(0.0632)	(0.0627)	(0.0619)	(0.0599)	(0.0595)
Brand Dummies								
Regional carrier	-	-	-	-	-	-	-	-
Aetna	0.338	0.293	0.327	0.329	0.299	0.281	0.307	0.282
	(0.0381)	(0.0385)	(0.0385)	(0.0382)	(0.0378)	(0.0371)	(0.0354)	(0.0351)
Blue	0.953	0.946	0.981	0.954	0.929	0.911	0.843	0.769
	(0.0239)	(0.0240)	(0.0240)	(0.0240)	(0.0238)	(0.0234)	(0.0230)	(0.0228)
Humana	0.995	0.999	1.016	1.010	0.960	0.915	0.853	0.810
	(0.0250)	(0.0250)	(0.0251)	(0.0250)	(0.0249)	(0.0247)	(0.0245)	(0.0241)
Kaiser Permanente	3.095	3.093	3.125	3.082	2.989	2.845	2.496	2.407
	(0.0342)	(0.0340)	(0.0340)	(0.0340)	(0.0338)	(0.0339)	(0.0337)	(0.0334)
United	1.018	1.025	1.061	1.038	1.001	0.970	0.893	0.860
	(0.0301)	(0.0300)	(0.0301)	(0.0300)	(0.0298)	(0.0294)	(0.0289)	(0.0286)
Pseudo R-squared	0.277	0.273	0.275	0.273	0.263	0.243	0.206	0.186
Observations	385,804	385,804	385,804	385,804	385,804	385,804	385,804	385,804

Standard errors in parentheses