NBER WORKING PAPER SERIES

HOW DO HUMANS INTERACT WITH ALGORITHMS? EXPERIMENTAL EVIDENCE FROM HEALTH INSURANCE

Kate Bundorf
Maria Polyakova
Ming Tai-Seale

Working Paper 25976
http://www.nber.org/papers/w25976

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2019

We thank Palo Alto Medical Foundation (PAMF) patient stakeholders and trial participants, as well as numerous Palo Alto Medical Foundation Research Institute (PAMFRI) team members for making the trial possible. We are grateful to Liran Einav, Aureo de Paula, Jonathan Kolstad, Amanda Kowalski, Jennifer Logg, Matthew Notowidigdo, Stephen Ryan, Justin Sydnor, Kevin Volpp, Stefan Wager and seminar participants at McGill University, University of Pennsylvania, CESifo Digitization, NBER Summer Institute, ASHEcon, Chicago Booth Junior Health Economics Summit, Stanford University, Indiana University, Boston University, and University of California Berkeley for their comments and suggestions. We also thank Sayeh Fattahi, Roman Klimke, and Vinni Bhatia for outstanding research assistance. Research reported in this paper was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDR-1306-03598). The statements in this presentation are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee. The project also received financial support from Stanford Innovation Funds. The experiment reported in this study is listed in the ClinicalTrials.gov Registry (NCT02895295). 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/w25976.ack

How do Humans Interact with Algorithms? Experimental Evidence from Health Insurance
Kate Bundorf, Maria Polyakova, and Ming Tai-Seale
NBER Working Paper No. 25976
June 2019
JEL No. D1,D12,D8,D81,D82,D83,D9,D90,D91,G22,H51,I13

ABSTRACT

Algorithms increasingly assist consumers in making their purchase decisions across a variety of markets; yet little is known about how humans interact with algorithmic advice. We examine how algorithmic, personalized information affects consumer choice among complex financial products using data from a randomized, controlled trial of decision support software for choosing health insurance plans. The intervention significantly increased plan switching, cost savings, time spent choosing a plan, and choice process satisfaction, particularly when individuals were exposed to an algorithmic expert recommendation. We document systematic selection - individuals who would have responded to treatment the most were the least likely to participate. A model of consumer decision-making suggests that our intervention affected consumers' signals about both product features (learning) and utility weights (interpretation).

Kate Bundorf
Health Research and Policy
Stanford University
HRP T108
Stanford, CA 94305-5405
and NBER
bundorf@stanford.edu

Ming Tai-Seale
School of Medicine, Family Medicine
and Public Health
University of California San Diego
9500 Gilman Dr
La Jolla, CA 92093
mtaiseale@ucsd.edu

Maria Polyakova
Department of Health Research & Policy
Stanford University
Redwood Building T111
150 Governor's Lane
Stanford, CA 94305
and NBER
maria.polyakova@stanford.edu

1 Introduction

People increasingly face decisions about complex financial products that have important implications for their health and financial stability. These types of decisions, which affect households across the income distribution, include, but are not limited to, products such as payday loans, mortgages, mobile phone plans, credit cards, life and health insurance, and investment vehicles. Participation in publicly subsidized benefits, such as Medicare, social security and tax-favored retirement arrangements, has evolved in similar ways, increasingly requiring relatively sophisticated financial decision making.

A large literature examining the quality of consumer choices in a variety of areas of household finance, however, suggests that these types of decision are challenging for many people. While many studies have documented that decision-making appears to be costly to consumers and that consumers display many types of behavioral biases that can lead to efficiency losses, there is less evidence on how to help consumers make better decisions. In their review of the household finance literature, Beshears et al. (forthcoming) conclude that many types of interventions designed to influence behavior, such as education and information, have had limited impact.

The emergence of large-scale data over the past decade and the corresponding development of techniques to analyze these data, such as machine learning, have the potential to dramatically change the process of consumer decision making in these environments (Einav and Levin, 2014). By lowering the costs of prediction, algorithms could help consumers make complex decisions by serving as either substitutes for or complements to human decision-making (Agrawal et al., 2019). While the literature on the methods of machine learning and artificial intelligence is expanding rapidly (Liu et al., 2018), there is very little evidence on how consumers incorporate algorithmic assistance into their decision making.

In this paper, we begin to close this gap by reporting on the results of a randomized controlled trial in which we offered older adults access to a decision-support tool incorporating personalized cost estimates and algorithmic expert recommendations for choosing among insurance plans. Our study makes three types of contributions to our understanding of how consumers interact with algorithmic-based recommendations. First, in contrast to most studies on the effects of informational interventions, our experimental results demonstrate - in a non-laboratory setting - that consumers are responsive to personalized information when making decisions. We find that people change their choices of insurance plans in response to our treatment and that the response is more pronounced when personalized information is combined with an "expert recommendation" feature that combines different types of information into a one-dimensional metric, simplifying the choice for consumers.

Second, the experimental set-up, combined with novel, machine-learning methods for estimating heterogeneous treatment effects, allows us to shed light on which types of consumers self-select into the use of electronic decision-support. We find evidence of substantial positive selection into the use of the on-line tool - more "active shoppers" are more likely to use the decision-making support tool, conditional on signing up for the experiment. Using our estimates of the treatment effects function, we also are able to analyze the likely response to the intervention of the people who did not take up the offer to participate in the experiment. We find that the

people who were the least likely to sign up for the experiment were those for whom the effects of our intervention on decision-making would have likely been the greatest.

Finally, we develop a theoretical framework to elucidate the mechanisms by which information affects consumer decision making in our setting and then use our trial data to estimate a structural model of choice to quantify the relative importance of each mechanism. We propose that providing information to consumers can have two conceptually distinct effects: it can change consumers' beliefs about the mapping of product characteristics into utility ("interpretation"), and it can also change consumer beliefs about product characteristics per se ("learning"). We find evidence that both channels are important in this setting and quantify how each one affects consumer welfare.

We examine consumer decision-making in the context of publicly subsidized prescription drug insurance for older adults in the US. Medicare Part D is a social insurance program for aged and disabled Medicare beneficiaries in which private plans compete for subsidized enrollees. The program is heavily subsidized and has high participation rates - insuring over 43 million older adults and accounting for over $88 billion in annual public spending (Kaiser Family Foundation, 2018). Older adults may choose between two types of private plans - a stand-alone prescription drug plan (PDP) or a Medicare Advantage (MA) plan in which coverage for prescription drugs and medical care are bundled in a single plan. In this project, we focus on stand-alone PDPs. Each year during a pre-specified open enrollment period, older adults covered by Medicare may choose a plan from among the approximately 25 stand-alone insurance plans offered in their geographic area (Kaiser Family Foundation, 2018).

Our study builds on a large and active economics literature examining health insurance choice more generally, with many studies focusing specifically on Medicare Part D (Keane and Thorp, 2016). While people with Medicare prescription drug plans are allowed to change their plans during an annual open enrollment period, switching rates are very low, with fewer than 10% of consumers changing their plans each year (Ericson, 2014; Polyakova, 2016; Ho et al., 2017), consistent with the literature documenting inertial behavior in this type of context beginning with Samuelson and Zeckhauser (1988). Estimates of switching costs are generally relatively large - ranging from 20 to 45 percent of annual spending (Handel, 2013; Ericson, 2014; Ho et al., 2017; Polyakova, 2016; Heiss et al., 2016). Several studies have documented that people often do not understand the basic features of their coverage (Cafferata, 1984; Harris and Keane, 1999; Kling et al., 2012; Loewenstein et al., 2013; Handel and Kolstad, 2015) and that their misperceptions influence their plan choices (Harris and Keane, 1999; Handel and Kolstad, 2015). Moreover, many people, when given a choice of plans, often choose a dominated option (Sinaiko and Hirth, 2011; Bhargava et al., 2017). Further, Ericson and Starc (2016) find that consumer choices and inferred utility weights change when health insurance products become standardized.

Other studies draw stronger, normative conclusions about consumer decision making (Abaluck and Gruber, 2011; Heiss et al., 2010; Heiss et al., 2013, 2016). For example, using a structural model of choice, Abaluck and Gruber (2011) find that older adults choosing among prescription drug plans weight premiums more highly than out-of-pocket costs; value plan characteristics, such as deductibles, beyond their effect on OOP spending; and place almost no value on the variance reducing aspects of plans. Ketcham et al. (2016) argue, however, that

these results may be driven at least in part by omitted variable bias - in particular, characteristics of plans such as customer service that are more difficult for econometricians to observe. Other research provides support for these concerns (Harris and Keane, 1999; Handel and Kolstad, 2015). For example, Harris and Keane (1999), adding attitudinal data to a structural model of choice, demonstrate that failing to control for these latent attributes leads to severe bias in estimates of the effects of observed attributes. Ketcham et al. (2015) also find that consumer decision-making improves over time, suggesting that choice inconsistencies may be short-lived. In our theoretical framework and its empirical mapping, we argue that these results can be reconciled if we allow for the possibility of "mistakes" both in consumers' information about product features, as well as in their interpretation of how much (known) product features matter for their utility. Consumer may learn about product features over time and yet not be interpreting this knowledge accurately.

Some recent studies have examined the importance of in-person advice relative to personalized information, but not algorithms, in the context of college funding and the SNAP program (Bettinger et al., 2012; Finkelstein and Notowidigdo, 2019). Few have examined the development and effects of products intended to help consumers choose among health insurance plans. Our paper relates most closely to the randomized field experiment conducted in the second year of Medicare Part D program by Kling et al. (2012). In this experiment, the authors sent Medicare Part D beneficiaries letters with personalized calculations of out-of-pocket costs that they would face in each insurance plan if they continued taking their existing medications. The personalized calculations were based on an out-of-pocket cost calculator made publicly available by the Medicare program. The experimental intervention increased plan switching rates. The authors interpret their findings as demonstrating the existence of "comparison friction" - that people often do not use potentially helpful information that appears readily accessible to them (Kling et al., 2012). Our findings emphasize the importance of these results by providing more direct evidence of the potential benefits of these types of tools for people who are unlikely to use them. We also demonstrate that how personalized information is presented has important implications for its use.

Our randomized field trial ran during the 2017 open enrollment period (November-December 2016). We conducted the project in collaboration with the Palo Alto Medical Foundation (PAMF), a large multi-specialty physician group in California. As part of the project, we designed and developed a software tool with the objective of helping older adults choose among Medicare part D prescription drug plans. Patient and provider stakeholders at PAMF participated in the design and development stages.

In addition to incorporating many aspects of user-centric design specific to this population, the tool incorporated three main features. First, the tool automatically imported a user's prescription drug information from their electronic medical record at PAMF. Second, the tool provided personalized information on expected spending in each available plan, including both the premium and the individual's likely spending on prescription drugs. Finally, the tool incorporated algorithmic expert recommendations. From a third-party vendor, we also obtained a personalized "expert score" for each insurance plan that summarized multi-dimensional plan features into a one-dimensional metric. Our trial population were PAMF patients eligible for Medicare Part D plans.

The experiment had two treatment arms and one control arm. People in the control arm did not receive

access to the decision-support software. Instead, when they logged into the study website, they saw a reminder about the timing of the open enrollment period and information about how to access publicly available resources to help them choose a plan. In the "Information Only" treatment arm patients received access to software that provided a list of all available plans with the individualized cost estimate and information about other plan features. The plans were ordered by the one-dimensional "expert" score, but the score itself was not displayed. The tool provided in the other treatment arm, "Information + Expert," was identical with the exception that the expert score for each plan was displayed and the three plans with the highest personalized score were marked as "Plans recommended for you."

We report three main findings. First, we find that providing consumers with access to a decision-support tool incorporating personalized cost estimates changes their choice behavior. While the effects of the interventions were qualitatively similar in the two arms, the "Information + Expert" arm had more pronounced effects on all outcomes. For our main outcome - switching of plans - exposure to the "Information + Expert" version of the software increased plan switching rates by 10 percentage points, a 36% increase relative to the control arm. We also find that treated individuals were more likely to be highly satisfied with the choice process, spend more time on the choice process, and choose plans anticipated to provide greater cost-savings. Using the generalized random forest analysis (Athey et al., 2019), we find evidence pointing towards heterogeneity (on observables) in treatment effects. The heterogeneity analysis suggests that treatment effects on the probability of plan switching are larger among individuals that are older and have less IT affinity.

Second, we find that selection into software use is quantitatively important. Many people who signed up for the trial and subsequently chose to use the tool if given access, were planning to switch their insurance plan independently of treatment. Those who chose to take up the software were inherently at least 7 percentage points more likely to switch plans, suggesting that the selection effect is nearly as large as the treatment effect and pointing to a strong complementarity in willingness to shop actively for financial products and interest in decision support tools. Using the individual-level prediction of treatment effects from the generalized random forest algorithm and the administrative data on all individuals that were invited to participate in the trial, we can also examine selection into the trial. We find that among individuals that were invited to participate in the trial, people who would have responded the *most* to the intervention were the *least* likely to sign up. These findings have important policy implications - they suggest that merely offering access to decision support (which is current Medicare policy) is unlikely to reach individuals who would be most affected by such decision-making support. Hence, policies with more targeted and intensive interventions may be required to reach consumers who could benefit from algorithmic expert recommendations.

Finally, we offer a conceptual insight for understanding the nature of the complementarity between machine-based algorithms and human decision making. Algorithms can influence decision making by changing either consumers' beliefs about product features, or how they value those features. This distinction has important implications for what types of information consumers need in order to make decisions. If consumer choices are inconsistent with rationality because of "behavioral" utility weights, then a policy of providing information about plan features will not lead to any behavioral responses. In contrast, if consumers know exactly how

to evaluate product features, but have a hard time simply observing the features of different products, then policies that attempt to educate consumers about the meaning of various features of financial products may not be necessary or effective.

Estimating an empirical version of this conceptual model, we find that the behavioral responses that we observe in the data are driven by both the learning and interpretation mechanisms we propose. These results offer one way to reconcile the debate in the literature on whether consumers' choices are inconsistent with the neo-classical preferences, or whether individuals are learning over time. Both of these mechanisms are likely to be taking place, as consumers may be learning about either features of their choice set or the utility weights, which could generate both choice inconsistency and learning at the same time. The model allows us to quantify the normative implications of the information and interpretation effects. We find that on average, the consumers that have "noisy" preferences would choose plans that result in 7% of surplus loss relative to "informed" consumers. This loss is extremely unevenly distributed: while for most consumers the noise in their beliefs about plan features and utility weights does not lead them to select suboptimal plans, for some consumers the losses can be quite significant.

The remainder of the paper is structured as follows. In Sections 2 and 3, we describe the key facts about the economic environment in Medicare Part D and our experimental design, respectively. In Section 4, we report the estimates of the causal effects of our intervention on consumer behavior. In Section 5, we analyze several aspects of selection in our setting. In Section 6, we present our conceptual framework and map our experimental results to an empirical version of the model. We then briefly conclude.

2 Background and Study Setting

Medicare is the public health insurance program in the U.S. for people age 65 and over and those eligible for social security benefits through disability. The program covers over 50 million people with 85% qualifying based on age (Centers for Medicare & Medicaid Services, 2019). Prescription drugs for Medicare beneficiaries are covered by Medicare Part D. In contrast to Medicare-financed medical benefits, prescription coverage is provided exclusively by private plans which compete for highly subsidized enrollees in a tightly regulated market (Duggan et al., 2008). In 2018, approximately 43 million individuals benefited from the program (Kaiser Family Foundation, 2018). Enrolling in Medicare Part D is voluntary for beneficiaries and requires an active enrollment decision in the form of choosing among the private plans offered in the beneficiary's market and paying a premium. Medicare beneficiaries can choose to enroll in either a stand-alone prescription drug plan (PDP) or a plan that bundles their medical and prescription benefits (Medicare Advantage). Fifty-eight percent of people enrolled in Medicare Part D choose a stand-alone plan (Kaiser Family Foundation, 2018).

Medicare beneficiaries who decide to enroll in a PDP typically choose from over 20 plans available in their market and can change their plan each year during the open enrollment period (October 15–December 7). Plans are differentiated along a variety of dimensions. First, premiums vary substantially. In addition, while the program has a statutorily-defined benefit package, insurers are allowed to deviate from that package as long

as the coverage they offer is actuarially equivalent or exceeds the statutory minimum. The statutorily-defined benefit, and as result, most offered benefits are non-linear insurance contracts. During 2017, the time period of our study, the statutorily-defined benefit had an initial deductible of $400 during which enrollees paid 100% of drug spending. After reaching the deductible, the beneficiary paid 25% coinsurance until reaching an initial coverage limit of $3,700. At that point, the enrollee fell into the "donut hole," paying 40% coinsurance for branded drugs and 51% for generics until spending reached a catastrophic threshold of $4,950 after which the beneficiary paid a lower coinsurance rate or a relatively small copayment for each drug (Centers for Medicare & Medicaid Services, 2015; Cubanski et al., 2018). A plan innovating within the statutorily-defined benefit package may, for example, either lower or eliminate the initial deductible while increasing cost sharing on some drugs. Plans are also differentiated along other dimensions such as the composition of pharmacy networks, the availability of mail order, formulary design and customer service. To capture the latter, the Centers for Medicare and Medicaid Services (CMS) has developed a measure of quality based on consumer assessments and annually publishes the "star rating" of plans on a 5-point scale.[1] A consumer enrolled in a plan in a given year who does not actively cancel or change her plan is automatically re-enrolled into the same plan for following year.

Our study focuses on aged beneficiaries enrolled in PDPs. During the 2017 open enrollment period (November-December 2016), we conducted a randomized field trial of a software tool designed to help consumers choose among Medicare Part D plans. Study participants lived in California during the 2017 open enrollment period. They were eligible to enroll in one of 22 plans offered by 10 insurers in California at an average monthly premium of $66 (standard deviation of $39). All but one plan offered either a standard deductible of $400, or lowered the deductible to zero, for an average deductible of $216, with a standard deviation of $202. Thirty percent of plans offered some coverage in the "donut hole," and plans covered on average 3,291 drugs, and varied in their formulary breadth (standard deviation of 257 drugs). The average CMS rating of plan quality in California was 3.4 out of 5 stars (s.d. 0.6).

3 Experimental Design and Data

3.1 Intervention

The trial was part of a larger research project funded by the Patient Centered Outcomes Research Institute in which we developed and evaluated a software tool intended to help Medicare beneficiaries choose among Medicare Part D prescription drug plans. The research was conducted in collaboration with patient and provider stakeholders affiliated with the Palo Alto Medical Foundation. Our focus group and qualitative research preceding tool development identified three key features that we incorporated into the software: automatic importation of the user's prescription drug information, user-centric design interface, and the availability of expert recommendations (Stults et al., 2018b,a). In the trial, we examined how two versions of the tool, one with and

[1] More information about the "star rating" measures is available on CMS Part C and D Performance Data page: https://www.cms.gov/medicare/prescription-drug-coverage/prescriptiondrugcovgenin/performancedata.html.

one without explicit algorithmic expert recommendations, performed relative to directing beneficiaries to existing, publicly available resources. Figure 1 provides screen shots of the intervention's user interface in the two treatment and the control arms.

The two versions of the tool were identical with the exception of whether the user interface included information on the expert score. In both versions, when people logged in, they viewed a list of their current prescription drugs based on the drugs recorded in their electronic medical record as of June 30, 2016 and had the opportunity to update the list as needed. They could then proceed to a screen listing all the plans available to them. In both arms, the plan list included the name of the plan, the individual's total estimated spending in each plan based on the entered drugs, and the star rating for each plan. The total estimated spending included the plan premium and out-of-pocket spending for the list of entered drugs based on information about drug-level coverage rules and pricing that Medicare Part D plans annually report to CMS. This computation was based on the user's current drugs and only incorporated drugs that consumers may need in the future if consumers actively entered them into the tool. The star rating is a plan-level measure primarily of service quality developed and disseminated by CMS. The plan in which the user was currently enrolled was highlighted and labeled as "My Current Plan". Users were able to select a subset of plans (up to three) for more detailed comparison. The detailed comparison screens provided information on an extensive list of plan features. Consumers were also able to obtain more information about each plan feature by clicking on a "question mark" icon.

The tool also incorporated algorithmic "expert" recommendations. Using proprietary scoring technology from a third-party provider, each plan available to the beneficiary was assigned an expert score. The expert score was based on the consumer's total spending in the plan for the set of drugs listed in the tool given the plan's benefit design (spending included each plan's premium) and the plan's "star rating". The expert score combined these plan features into a one-dimensional metric. Plans with lower expected spending for a given individual and higher quality scores received higher expert scores. The expert score was not based on any additional information about the individual or the plan other than total cost and the plan's star rating.

The two treatment arms differed only based on how they incorporated the expert recommendation. In both treatment arms, the plans were initially ordered by the expert score with the highest ranking plan at the top of the list. In the "Information Only" arm, although the list was ordered by the expert score, users did not see the score itself. In the "Information + Expert" arm, the three plans at the top of the list with the highest scores were highlighted and labeled as "recommended for you", and the plan information included each plan's expert score. As Panel A and Panel B of Figure 1 illustrate, the user interface for the treatment arms was very similar with the exception of the expert score column and the highlighting of the top three plans in the "Information + Expert" arm.[2]

When participants in the control arm logged into the study website, they received access to information on plan enrollment including a reminder about the open enrollment period in Medicare Part D, some information about the benefits of reviewing their coverage, links to publicly available resources that they could use to

[2]Panel A and Panel B are screenshots for different patients, which explains the different ordering of plans. For the same patient or for two different patients with identical lists of drugs, the ordering of plans would have been the same. Both arms highlighted the incumbent plan, even though it is not visible on B.

evaluate their options, including the Medicare.gov plan finder and Health Insurance Counseling and Advocacy Program counselors, and information about how to access a list of their current prescribed drugs from their electronic medical record. People in the control arm did not receive access to the decision-support software. The control arm is illustrated in Panel C of Figure 1.

3.2 Study Population

We recruited trial participants from patients who receive care at the Palo Alto Medical Foundation (PAMF) to focus on people for whom we had access to electronic information on their use of prescription drugs. Using administrative data from PAMF, we identified a cohort of patients likely to be eligible for the trial based on their age (66 to 85 years), residence (lived in the 4-county primary PAMF service area) and indication of active medication orders (to ensure they were active PAMF patients and thus would have updated medication lists). The administrative data did not allow us to identify people currently enrolled in a Part D plan, our target population. Instead, we excluded people who were unlikely to be enrolled in stand-alone Part D, because they either had a Medicare Advantage or a Medi-Cal (California's Medicaid program) plan. After these and several other minor exclusions primarily for missing or inaccurate data, we identified 29,451 patients potentially eligible to participate in the trial.

During the fall of 2016, we mailed the 29,451 potentially eligible patients invitations to participate in the trial. The invitation provided some basic information about the trial and informed individuals that they would receive a $50 gift certificate for participating in the study following the completion of a questionnaire at the end of the open enrollment period. We sent a follow-up letter approximately two weeks later to those who did not respond to the initial invitation. In the letter, patients received a log-in ID and were directed to an enrollment portal in which they could check their eligibility, provide informed consent and respond to a survey from which we collected baseline data to supplement administrative records (Baseline Survey). Patients also provided information that we used to verify their identity subsequent to their on-line enrollment. We considered those who completed the enrollment portal steps and whose identity was successfully authenticated shortly after their on-line enrollment as enrolled in the trial.

At the point of enrollment into the study, participants were randomized to one of the three arms using a random number generator. After subjects enrolled, we sent them a confirmation e-mail with information on how to access the study website and telling them the website would be available shortly after the open enrollment period began. They then received another email reminder once open enrollment began and the tool was active. In both cases, participants received the same standardized e-mail independent of the arm to which they had been randomized. The subjects thus received no information on their assigned study arm until they accessed the study website during the open enrollment period. When participants logged in to the study website, they accessed content specific to the study arm to which they had been randomized. Just before the open enrollment period ended, we e-mailed another reminder to participate. The day after the open enrollment period ended, we e-mailed those enrolled in the study an invitation to participate in the final survey; we sent a survey reminder in early January. The invitation to complete the final survey was sent to all trial participants, independently of

whether they actually accessed the study website during the open enrollment period. We included people who completed the final survey by January 20th in the final study sample. Figure 2 summarizes this process.

Figure 3 describes the enrollment flow. We invited 29,451 PAMF patients to participate. 1,185 ultimately enrolled in the study and were randomized to one of three arms. Among those randomized to each arm, some entered the study website and some did not. Because we sent the final survey to those enrolled in the trial whether or not they entered the study website, within each arm, the final survey includes both those who entered the study website and those who did not. Table 1 provides descriptive statistics for the sample of people invited to participate in the trial and compares the characteristics of those who did and did not choose to enroll in the trial using administrative data from PAMF. The mean and standard deviation of each dependent variable in the table represents summary statistics for the full sample of 29,451 invited individuals. Invited individuals were on average 74 years of age (s.d. of 5 years), 54 percent were female, 35 percent were non-white,[3] and 54 percent were married. We matched each individual to their census tract based on their address and developed measures of socioeconomic status based on census tract characteristics including median household income and percent of individuals with a college degree. The average (median) household income in our sample was 107 thousand dollars (standard deviation of 46 thousand) and the average percent of the census tract with a college degree was 54 (standard deviation of 0.2), both reflecting the relatively high socioeconomic status of the geographic area from which we recruited patients.

Invited individuals had on average 4.5 active medication orders for prescription drugs (measured from PAMF records prior to the intervention). Drug use varied considerably, with a standard deviation of 3.2 drugs. Column (8) reports the statistics on Charlson score, a common measure of comorbidities based on diagnosis codes (Charlson et al., 1987). The measure counts how many of 22 conditions an individual has, assigning higher weights (weights range from 1 to 6) to more severe conditions. A higher Charlson score reflects an individual in poorer health. In our sample, the score ranges from 0 (no chronic conditions) to 13, with an average of 1.16 and a standard deviation of 1.53. Finally, we measure individuals' IT-affinity at baseline, by recording whether they had logged in to their PAMF electronic medical record over the 3-year period prior to the trial; and if so, how often they communicated with care providers via this system (Tai-Seale et al., 2019). Our measure of communication frequency is based on conversation strand metric which groups individual e-mails into conversations (Tai-Seale et al., 2014). In the full sample of invited participants, 69 percent had accessed their personal medical record within the prior three years. Intensity of use, measured by the number of communication strands, averaged at 3.3 strands but varied considerably, with a standard deviation of 6. The average number of strands was 4.7 among those individuals who ever logged into the electronic medical record and ranged from zero to 174 strands, with significantly more strands (although not a higher probability of using the system) for individuals with a higher Charlson score or more drugs on their record, as would be expected if patients in poorer health are more likely to communicate frequently with their physicians.

Overall, the sample of individuals who were invited to participate in the experiment were higher income, more educated, and likely more IT-savvy than an average Medicare beneficiary. This difference is important to

[3]Includes those who did not have a record of their race or reported "other" in electronic medical records.

keep in mind when interpreting our results and considering the external validity of the experiment. The high average income of our participants makes them unrepresentative of the broader population of older Americans; however, this sample provides us with the opportunity to test whether offering decision support software - in one of the wealthiest and technologically most attuned areas of the country - affects individuals' behavior. Our results likely provide an upper bound for the effects - particularly with respect to take-up - in the general population.

Table 1 demonstrates that there was significant selection into trial participation. The row labelled "randomized" provides estimates of how those who enrolled in the trial differ from those who did not, by reporting the results of a regression of each characteristic specified in columns (1)-(8) on a dummy for whether or not the individual agreed to participate in the trial. Those 1,185 individuals that responded to our invitation and chose to enroll in the trial were on average a year and 8 months younger (column 1), 4 percentage points less likely to be women (column 2), 13 percentage points more likely to be white (column 3), 7 percentage points more likely to be married (column 4), had 5.8 thousand dollar higher (measured at census tract level) household income (column 5), and lived in areas in which residents were 4 percentage points more likely to have a college degree (column 6). All of these differences were highly statistically significant and some were also economically significant - the gender difference of 4 percentage points corresponds to women being 7 percent (relative to the mean) less likely to participate, the difference in race suggests that participants were 37% less likely to be non-white and 13% more likely to be married. Those enrolling in the trial did not have a statistically different number of drugs in their records (column 7), but were significantly healthier with a 16 basis points lower Charlson score (column 8), which is 14% lower than the sample mean. The population taking up the treatment offer was substantially more likely to have used PAMF's patient portal to the electronic medical records - 27 percentage points or almost 40% more likely relative to the mean (column 9) - with 96 percent of individuals in the enrolled population having used the PAMF's electronic health records within the last three years. The enrollees also used these systems more intensively, having sent more than twice as many online messages to their care team relative to the general pool (column 10), despite being in better health on average.

3.3 Randomization

Out of 1,185 individuals, 410 were randomized into the "Information + Expert" arm, 391 into "Information Only" arm, and 384 into the control arm. Randomization was done in real time: just after the participant enrolled in the trial through the enrollment portal, he or she was randomized into one of three arms. We performed a Monte Carlo simulation to confirm that the unequal distribution of individuals within each group is consistent with randomization. Importantly, at the point of randomization, the individual did not learn to which arm they had been randomized - so that when they later received notice that open enrollment had begun and they could access the study website, they did not know whether they were going to have access to the treatment intervention. Tables 2 through 5 examine the quality of randomization, compliance with experimental treatment, and attrition. We discuss each in turn.

Table 2 reports our randomization balance checks. We test whether there are differences in means of

observable characteristic by experimental arm assignment. The table reports the results of regressions for each observable characteristics as the outcome variable on the indicators for being randomized into "Information Only" or "Information + Expert" treatment arms. The constant in this regression captures the mean in the control arm. Two out of ten observable characteristics exhibit differences between the control and treatment arms at conventional levels of statistical significance. We observe that individuals randomized into the control arm were 8 months older (1 percent relative to the sample mean) than individuals randomized into either of the treatment arms. We also observe that individuals randomized into the "Information + Expert" treatment arm were more intensive users of the electronic communication with their physicians. The point estimates for this characteristic are not statistically different from zero for the "Information Only" arm. We do not observe any significant differences between the two treatment arms, as suggested by the F-test, reported in the last row of the table. Differences in two out of ten characteristics are possible by chance and the magnitude of the statistically significant differences, as well as the lack of differences in other outcomes suggests that randomization was not compromised and worked as intended. To account for the realized differences in age and intensity of EMR use, as well as to generally reduce the noise in our estimates, we will control for observable characteristics in our analysis of treatment effects in Section 4.3.

We next examine whether there was systematic attrition in response to the endline survey, which is our key source of outcome measures. After individuals (electronically, through the enrollment portal) agreed to participate in the experiment, they were randomized into one of the study arms and given information about how to access the online tool. At the end of the open enrollment period, we sent a survey to all individuals that were originally randomized (independent of whether they participated in the trial by accessing the study website). 928 individuals responded to at least one question in the survey by a pre-specified cutoff date. Table 3 examines whether, relative to 1,185 randomized individuals, the 928 who responded to the survey differed on their observable characteristics. The table reports the results of a regression of each characteristic on a dummy indicating an individual responded to the endline survey. Eight out of ten characteristics do not differ between those who responded to the survey and those who did not. Race and college education, in contrast, do differ. Individuals who responded to the survey were substantially (9 percentage points relative to 22 percent in the randomized sample) less likely to have their race recorded as white (which includes those who did not agree to their race being recorded in EMR) and were slightly more likely to have a college degree as measured at the census tract level (4 percentage points relative to the sample mean of 59 percent). The lower probability of non-white participants responding to the survey is consistent with the growing literature that documents racial gradients in trust in interactions with government and institutions (e.g. Alsan and Wanamaker, 2018).

Table 4 presents the same analysis of attrition into the endline survey, but separately for each experimental arm. Within each arm, we run a regression of the observable characteristic recorded in each column title on the indicator variable for responding to the endline survey. The results across arms are broadly consistent with the overall attrition results, suggesting no pronounced differential patterns of attrition across arms. We do not observe differential attrition based on race in the control arm, although it is present in both treatment arms. Individuals responding to the survey in the control arm are slightly more likely to have a college degree

(at the census tract level), but are otherwise not different from other individuals in the control arm. In the "Information Only" arm, we observe significant differences in the probability of being non-white. In the "Information + Expert" arm we observe both the race effect as well as the difference in the EMR use intensity - individuals responding to the survey in this arm were slightly more likely to be more intensive EMR users - this difference, however, is not suggesting differential attrition in this arm, since individuals randomized into this arm were higher intensity EMR users at the original randomization stage (as can be seen in column 10 of Table 2).

Finally, in Table 5 we repeat the balance on observable comparison of Table 2 for our main analytic sample of 928 individuals who responded to the endline survey. In column (1), we document that there were no statistically distinguishable differences in survey response rates across three experimental arms. In columns (2) to (11), we report the coefficients of specifications that regress the observable characteristics on the indicator variables for being randomized into two treatment arms. We conclude that randomization was largely preserved at the endline survey stage. We observe that individuals randomized into arm "Information + Expert" are more intensive users of EMR, but this effect was already present at the original randomization. Unlike in the original randomization, we do not estimate statistically significant differences in age across arms, although the point estimates of differences are close to those at the original randomization, suggesting that the differences persist but cannot be detected due to reduced sample size. We detect a slightly more pronounced - relative to the original randomization - coefficient on the probability of being married, suggesting that those who responded to the survey in the "Information + Expert" arm were slightly more likely to be married. In sum, attrition into the endline survey overall appears to be limited; importantly we do not find much evidence for differential attrition across arms above and beyond the differences observed across arms at the original randomization stage. Hence, we proceed to the analysis of outcomes from the endline survey. In all of these analyses, we control for observable characteristics to improve power and to account for any realized differences in observables at randomization and endline survey stages.

3.4 Outcomes

We consider six outcomes across different domains in our baseline specifications, four of which we pre-specified as primary outcomes and two which we pre-specified as secondary outcomes. First, we test whether individuals switched their Medicare Part D plan. We construct our measure of switching using two self-reported measures obtained from the baseline and endline surveys. We are unable to use a measure based on administrative data since PAMF does not have information on the patient's Medicare Part D plan in its administrative records. In both surveys we asked participants to report their Part D plan - the participants were given the list of available plans and could select one of the plans, or choose "None of the above." Our first measure of switch is then an indicator that takes the value of one if the Part D plan reported in the endline survey differs from the plan reported in the baseline survey. Further, in the endline survey we directly ask participants whether they switched their plan, which generates the second measure of switching. To reduce the measurement error in the switching metric, we classify an individual as having switched plans only if both indicators indicate a

plan switch. We use this interacted measure of switching as our outcome variable.

The next two outcomes measure different types of consumers' perceived experience. First, we use a self-reported measure of how satisfied individuals were with the choice process. We construct an indicator outcome variable that takes a value of 1 if an individual reported being "Very Satisfied" (other options included: somewhat satisfied, somewhat dissatisfied, and very dissatisfied) with the process of choosing their plan in the endline survey. Second, we measure the degree of decision conflict that an individual experienced around their Medicare Part D plan choices using a validated scale (O'Connor, 1995; Linder et al., 2011). The score is constructed based on individuals' replies to 9 questions about their confidence in their choice, availability of support, and understanding of risks and benefits. A higher score value indicates more decision conflict.

Our fourth outcome is a measure of changes in consumers' expected total (premium + out of pocket) monthly costs. For each consumer, we compute the difference between two levels of expected total costs. One is the level of total cost that consumers would face under the plan they chose in 2017 (as reported in the endline survey). The second is the level of total cost that consumers would have faced in 2017 if they had stayed in their 2016 plan. In both cases, we use the 2016 baseline drug list and the 2017 plan characteristics. Thus, if consumers did not change plans, the difference in total cost would by construction be zero.[4] For consumers who changed plans, this variable measures the difference between expected 2017 costs in the plan chosen in 2017 to what the expected costs would have been if a consumer stayed in her 2016 plan. The comparison of the expected out of pocket costs in the two plans in the same year captures any common trend in costs.

The fifth outcome is the amount of time individuals spent on their choice. The cost of time and effort is frequently considered to be the main barrier to improving individuals' choices, so it is important to understand how much the use of software "cost" individuals who chose to take it up. We create an indicator variable that takes the value of 1 if individuals report spending more than 1 hour on their choice of Medicare Part D plans.

Finally, our sixth outcome is the probability that an individual chooses one of the three plans with the highest algorithmic score ("expert recommended" plans). These plans appeared as the first three plans in each treatment plans, but were highlighted for the participants only in the "Information + Expert" treatment arm.

4 Effects of the Intervention

4.1 Effect of Offering Algorithmic Decision Support

We start by estimating the effect of offering algorithmic decision support to participants using an intent-to-treat analysis (ITT). Let the assignment to experimental arm "Information Only" be denoted with an indicator variable I, while the assignment to experimental arm "Information + Expert" be denoted with an indicator variable E. For outcome variable Y_i, we estimate:

$$Y_i = \alpha_0 + \alpha_1 E_i + \alpha_2 I_i + \delta X_i + \epsilon_i \tag{1}$$

[4]This does not strictly hold true for the interacted switch measure. The difference in costs is measured based on plans that individuals reported at the baseline and endline. While some individuals report different plans and hence we compute a non-zero change in cost, we do not count these individuals as switchers in the more conservative interacted switching measure.

The coefficients of interest, α_1 and α_2, measure whether being randomized into treatment arm "Information + Expert" or treatment arm "Information Only," on average, changed the outcomes of interest. We consider heterogeneity in the treatment effects in detail below. X_i is a vector of individual observable characteristics that were analyzed in Sections 3.2 and 3.3. As these controls are to a large extend balanced through randomization, their primary role is to reduce the standard errors of the point estimates, as our sample size is relatively small. As we would expect, including or excluding control variables has very little effect on the point estimates.

Table 6 reports the ITT results for all six outcome variables of interest. For each regression we report the mean of the outcome variable in the control group, as well as the estimates of α_1 and α_2. The number of observations across different outcome variables varies, since some individuals did not fill out all questions in the endline survey. We report the mean and the standard deviation of each outcome variable for the entire sample at the bottom of the table. The last row of the table reports the p-value of an F-test for whether the estimates of α_1 and α_2 differ from each other.

Column (1) presents the results for the measure of plan switching. We find that a high fraction of people - 28 percent as compared to the national switching rate of approximately 10 percent (Polyakova, 2016) - in our control group switched plans, suggesting that the trial already attracted relatively active shoppers (we explore this point in more detail in Section 5). Being randomized to the "Information Only" treatment increased the switching rate by 1 percentage point, but the estimate is noisy and we cannot reject that the effect of offering decision-making support was zero in this arm. Being randomized into the "Information + Expert" intervention, in contrast, increased the switching probability by 8 percentage points. The estimate is precise and we can reject a zero effect of offering algorithmic decision support at the 95 percent confidence level. The estimate is also economically significant, suggesting a increase in the switching rate of 28 percent relative to the control group. The difference between two intervention arms is economically large and statistically significant at 10% level.

In column (2) we observe that only 39 percent of individuals in the control arm report being very satisfied with the choice process of the Part D plans. Individuals assigned to "Information Only" arm report a 6 percentage point higher satisfaction rate, although we again cannot reject that the effect was zero. Satisfaction with the choice process appears to be improved more by the algorithmic recommendation intervention, with 8 percentage points more people (or 20 percent more) report being very satisfied with the process in the "Information + Expert" arm. As we observe in Column (3), satisfaction with the choice process does not appear to result in a decreased feeling of decision conflict. We cannot reject zero effects of the intervention at any conventional levels on the degree of decision conflict.

In column (4) we note that 75 percent of individuals in the control arm spent more than an hour choosing their Medicare Part D plan. We estimate that individuals assigned to the "Information + Expert" arm were 8 percentage points more likely to spend more than one hour choosing their Part D plan, and yet they also report more satisfaction with the decision process. This suggests that individuals may be willing to invest time in their choices if this time can be spent productively.

In column (5) we effectively get a measure of the return on time investment, estimating how much individuals

save in expected costs by changing their plans. We observe a $112 reduction in expected costs at the baseline in the control group.[5] This is consistent with both a relatively high switching rate in the control group, as well as with either selection or "reminder" effects in the control group, as we discuss below. Relative to the control group, savings are much more pronounced in the group exposed to the "Information + Expert" treatment. Individuals choose plans that have $94 larger decline in expected cost - in other words, individuals choose plans that in expectation would save them 80% more. The point estimate for the "Information Only" arm suggests a magnitude of the effect that is about half the size, but we cannot reject that the effect is zero.

Finally, in column (6) we measure the likelihood that consumers reported choosing one of the "expert recommended" plans - i.e. plans with the highest algorithmic scores. These plans were relatively popular among consumers prior to the intervention.[6] 39 percent of individuals in the control group enrolled in (what would have been) an expert recommended plan for them in 2017. The probability of enrolling in an expert-recommended plan increased 5 to 6 percentage points (15 percent) from the exposure to either treatment. Both coefficients, however, are noisy and we cannot reject a zero effect at 95% confidence level. The effect appears to be slightly more pronounced in the "Information + Expert" arm, both in absolute levels and in statistical precision, relative to the "Information Only" arm.

4.2 Effect of Using Algorithmic Decision Support

We next proceed to estimate the average causal effect of using the decision support software among treatment compliers. We estimate a 2SLS model, in which being randomized into either the "Information Only" or "Information + Expert" arms serve as instruments for using the corresponding version of software. Let the use of "Information Only" version of software be denoted with an indicator variable UI, while using the software in "Information + Expert" arm be denoted with an indicator variable UE. For outcome variable Y_i (same outcomes as above), we estimate:

$$Y_i = \gamma_0 + \gamma_1 UE_i + \gamma_2 UI_i + \phi_0 X_i + \epsilon_{i0} \tag{2}$$

$$UE_i = \pi_{10} + \pi_{11} E_i + \pi_{12} I_i + \phi_1 X_i + \epsilon_{i1} \tag{3}$$

$$UI_i = \pi_{20} + \pi_{21} E_i + \pi_{22} I_i + \phi_2 X_i + \epsilon_{i2} \tag{4}$$

Here, variables UE_i and UI_i take the value of 1 if the individual logged-in into the software, which we can track through individualized login information linked to encoded patient id. π_{11}, π_{12}, π_{21}, and π_{21} measure the take-up of the software across experimental arms. The coefficients of interest are the 2SLS estimates of γ_1 and γ_2. These coefficients measure the impact of using the algorithmic decision support (or at least logging into the software) on individuals' behavior.

[5] As the cost estimates are extremely skewed, we trim the regression to only include cost changes between the 1st and 99th percentile of changes.

[6] This decreases our power to detect changes in the probability of enrolling in an expert recommended plan. To increase power, in this regression specification we control for the whether individuals were enrolled in a plan that would have been one of three top plans for the at the baseline

Table 7 reports the first stage coefficients and the 2SLS estimates for the six outcome variables of interest. As we observe in Column (1), the take up of the software tool conditional on being randomized into a treatment arm was very high. Being randomized into "Information + Expert" arm increased the take up of the "expert recommendation" version of software from zero (by construction, individuals in the control arm did not have access to the software) to 81 percent. Similarly, being randomized into "Information Only" arm increased the take up of the individualized information version of the software from zero to 80 percent.

The estimates reported in columns (2) to (8) of Table 7 are the same as coefficients in Table 6, but re-scaled by the first stage (with the exception of column 6). Hence, the direction of the effects is the same and we observe only a change in the magnitude that reflects the imperfect treatment take up. The LATE (or in this case, treatment on the treated) estimates suggest that using the algorithmic expert software increases plan switching rates by 10 percentage points relative to the baseline rate of 28 percent in the control group (36% increase). We do not observe a significant increase in average switching rates relative to the control group from the use of the "individualized information" version of the software (column 2). As in the intent-to-treat results, we see a notable increase in the probability that individuals using software report being more likely to be highly satisfied with the choice process. The effect of the "expert recommendation" version of the software has a slightly more pronounced effect, increasing the subjective choice process satisfaction by 23 percent (column 3). We also observe that individuals that use software are 10 percentage points more likely to spend more than an hour on choosing their Part D plans (column 5).

In column 6, we introduce a new outcome - an index that measures the intensity of software use. The index outcome measure comprises five underlying outcomes: whether the consumer viewed explanation buttons within the software, how often these buttons were clicked, the total number of actions within the software, the number of actions per login, and the total time that the individual spent within the software tool as measured by clicks and login behavior. The index is defined to be an unweighted average of z-scores of each component outcome, where all of the outcomes are oriented such that a positive sign implies more intensive website use. The z-scores are in turn computed by subtracting the mean in "Information Only" group and dividing by the standard deviation in "Information Only" group. All underlying outcomes can only be defined for individuals that were assigned to either of the treatment arms; they are further only defined for individuals that used the software. Hence, for this measure we can only compare individuals that used the "Information Only" version of software to those who used the "Information + Expert" version, excluding all individuals in the control arm. We estimate that individuals assigned to the "Information + Expert" version of the software were using the decision-support tool much more intensely than those in the "Information Only" group. This is an interesting finding, as it suggests that algorithmic advice serves as a complement to human decision making, inducing more consumer engagement (Agrawal et al., 2019).

The reduction in expected costs as reported in column 7 becomes more pronounced relative to the ITT results, as we now focus on compliers, who we know were more likely to switch their plans. Individuals using "Information + Expert" version of the software choose a plan with $116 lower expected cost. As the reduction in the cost is driven by individuals that actually switch plans, we analyzed the reduction of costs among switchers

further. Among those who switch in the "Information + Expert" arm, expected spending in the plan chosen post-intervention was $595 lower than if the consumer stayed in the incumbent plan. For the "Information Only" arm, the decline was $485. In both treatment arms, consumers were 7 percentage points (imprecisely measured) more likely to have one (of three) "expert-recommended" plans relative to the control arm.

Overall, we conclude that being exposed to the algorithmic recommendation increased the propensity of consumers to shop for plans and decreased their costs. Being exposed to individualized information had effects in the same qualitative direction but quantitatively, the effects on switching and costs were less pronounced, although, except for plan switching, we cannot formally reject the equivalence of the effects. Using both versions of the decision support software increased consumers' search time, but also their subjective satisfaction with the process. The intensity (including time) of software use was significantly more pronounced among consumers exposed to the treatment arm with the "algorithmic expert advice" feature.

Two issues are important to keep in mind when interpreting our LATE estimates. First, we feel reasonably confident in interpreting these results as treatment on the treated, since we do not believe that individuals outside of treatment groups had access to the treatment software. The trial enrollment process insured that no two individuals in the same household were participating in the experiment. In addition, PAMF patients who participated in the experiment are not concentrated in a small geographic area and are unlikely to be acquainted. Hence, it is not very likely that the control group including always-takers - people who used the software even though they were not randomized to a treatment arm. Second, in theory, being randomized into a treatment arm could affect individuals in ways other than through software use or through information about Part D within the software. One plausible alternative hypothesis is that being randomized to a treatment arm reminded people about the prescription drugs they were taking (after those were imported from the electronic medical records). This reminder could have changed individual behaviors relative to the control group, who were informed about the possibility of seeing their drug lists in the electronic medical records, but were not shown their list of drugs explicitly. While this channel may affect our estimates of behavioral responses when comparing the treated individuals to the control group, this difference does not exist in the comparison of "Information Only" and "Information + Expert" treatments - individuals were shown their drugs in both treatment arms. Hence, the differences in behavior between treatment arms provide compelling estimates for the effects of exposure to different types of information rather than other channels.

4.3 Heterogeneity of Treatment Effects

We next examine heterogeneity in the estimated treatment effects. We focus on the intent-to-treat analysis, as being offered decision support algorithms is most relevant for policy. Given the small sample size of the intervention, estimates of treatment effects among subgroups in our population are unlikely to be precise; however, the estimates may still be informative about the degree and direction of heterogeneity.

We use generalized random forests to systematically analyze heterogeneity in treatment effects in the sample of people enrolled in the trial along the same ten observable demographic and health-related characteristics that we examined in Sections 3.2 and 3.3. These include: age, gender, race, marital status, income at the census

tract level, share of college-educated individuals at the census tract level, the number of prescription drugs, the Charlson score, the use of online patient records, and the intensity of its use as measured by message strands. The generalized random forest methods are discussed in detail in the emerging literature on the use of machine learning methods for causal inference (Wager and Athey, 2018; Athey et al., 2019; Davis and Heller, 2017; Hitsch and Misra, 2018; Asher et al., 2018). The basic idea is to create - under the assumption of unconfoundedness - a decision tree that identifies splits in observable demographics in a way that maximizes differences in the treatment effect along the split line. As there are many possible permutations of such trees, the random forest algorithm bootstraps the tree, generating a more robust prediction (aggregated through an adaptive weighting function across individual draws of trees) of treatment effects as a function of observables.

For each of our six outcomes we use the estimates of the generalized random forest algorithm to compute the predicted treatment effect (separately for the "Information Only" and "Information + Expert") for each individual that participated in the trial, based on observable characteristics. We observe pronounced heterogeneity in point estimates of the predicted treatment effects across individuals. While we cannot formally reject a uniform treatment effect due to the limited number of individuals in-sample, two suggestive patterns emerge when comparing the two treatment arms in the context of plan switching outcome.[7] For the "Information Only" arm, the treatment appears to have induced some consumers to be more likely to stay in their incumbent plans. This evidence of asymmetry in treatment effects may explain the small average intent to treat effect that we estimated in Table 6, as this average combines a positive treatment effect for some individuals and a negative treatment effect for others. "Information + Expert" recommendation treatment effects have little mass at zero, with the majority of individuals having a positive treatment effect on plan switching from algorithmic expert recommendation.

In addition to providing a sense of the degree of heterogeneity in treatment effects in the estimation sample, the same method allows us to predict treatment effects out of sample. Table 8 summarizes the results of this prediction exercise. We compute a treatment effect for each individual that was invited to participate in the trial (i.e. for 29,451 individuals). We split these individuals into five equal-size groups, by quintiles of the treatment effect distribution. Within each quintile, we then report the average value of the observed demographic. This allows us to qualitatively characterize the outcome of the generalized random forest procedure. We observe several clear patterns. Treatment effects are greater among older individuals; they are also more pronounced among women and non-white beneficiaries. The starkest differences emerge on the IT affinity dimension. Individuals who are less likely to have ever used the electronic medical records and use it much less intensively have much larger estimated behavioral responses to the intervention. While this analysis provides initial insights into what types of people were likely to enroll in a trial providing access to a web-based tool, we return to this

[7] To test the quality of our causal forest estimates and our ability to formally reject the null of no heterogeneity in the treatment effects, we implement a calibration test motivated by Chernozhukov et al. (2018) as described in detail in Athey and Wager (forthcoming). The calibration test produces two coefficients. The first coefficient (α) tests the accuracy of the average predictions produced by the generalized random forest, while the second (β) is a measure of the quality of the estimates of treatment heterogeneity. If $\alpha = 1$, then we can generally say our forest is well-calibrated, while if β is statistically significant and positive, we are able to reject the null of no heterogeneity. Our estimates of α are close to 1 for both treatment arms, although the estimate is very noisy for the "Information Only" arm - α=0.98 (s.e. 0.45) for "Information + Expert" arm and α=1.04 (s.e. 2.6) for "Information Only" arm. These results suggest that our forest is well-calibrated. For both arms our estimates of βs, however, are too noisy to interpret, suggesting that we cannot formally reject the null of no heterogeneity in treatment effects.

idea in more detail in the context of selection discussion in the next Section.

5 Selection

We use three empirical strategies to quantify the importance of selection in the take up of the decision support software. Understanding who chose to take up the intervention is crucial for interpreting the external validity of the experiment and for understanding how to target policies offering consumers algorithmic decision-making support tools.

5.1 Lower Bound of Selection

Our first strategy exploits the simple idea that the IV estimates in our setting correct selection bias. Hence, the difference between the IV and OLS estimates are informative about the degree of selection into the use of software among those who signed up for the trial. OLS estimates of the effects of using software on outcomes among those enrolled in the trial capture both treatment and selection effects in the treatment group relative to the control group. For example, trial participants who are more active shoppers and are considering changing their plan even in the absence of our intervention are likely to disproportionately select into using the software. To quantify this selection bias, we first estimate the following OLS relationship:

$$Y_i = \tau_0 + \tau_1 UE_i + \tau_2 UI_i + \kappa_0 X_i + \epsilon_i \tag{5}$$

In this equation, τ_1 and τ_2 are biased estimates of the treatment effects, as the exposure to software conditional on being randomized into a treatment arm is determined by the individual's decision to take up the intervention, which, for example, could be correlated with the latent propensity of switching plans. We use this omitted variable bias to learn about the magnitude of selection. Panel A of Table 9 reports OLS results for our six outcome variables of interest. These estimates of the effects of the intervention are much larger than the IV estimates for both treatment arms. We estimate that in the "Information + Expert" arm, using the software was associated with a 17 percentage point increase (9 percent in "Information Only" arm) in the probability of switching plans (column 1). For both arms, this is 7 percentage points larger than the IV estimates (reported again in the second section of Panel A in the same table for convenience). We conclude that out of 17 percentage point increase (9 for the "Information Only" treatment arm) in switching rates as suggested by OLS, 10 percentage points (2 for "Information Only") was the treatment effect and 7 percentage points was selection. In other words, individuals that took up the experimental software were inherently 7 percentage points more likely to switch their plans than those individuals who were assigned to treatment arms, but chose not to use the software (or those assigned to the control arm).

The comparison of OLS and IV estimates in column (2) suggests little selection on the satisfaction with the Part D shopping process, although the emerging direction of selection appears to be negative. In other words, individuals that were inherently less likely to be satisfied with the selection process were possibly more likely to

take up the decision support tool. We observe only very noisy estimates of differences in decision conflict score (column 3) and no selection on the time search dimension (column 4).

Individuals choosing to use the software appear to be those who would have experienced greater savings absent the intervention (column 5) and would have been more likely to choose one of the three expert recommended plans (column 6).

Overall, the evidence is consistent with the idea that, even among those who chose to participate in the trial, individuals who actively accessed algorithmic advice were inherently more likely to revise their plan choices towards lower cost plans absent the intervention. The magnitude of selective take up is substantial relative to the treatment effect, especially with respect to the inherent propensity to switch plans. Notably, these results are estimated relative to the average outcome among those assigned to the control group. In this exercise, outcomes in the control group serve as a control for selection into the intervention. The average outcome of the control group could itself, however, are potentially comprised of both selection and treatment effects. In particular, simply entering the study website could have generated a "reminder effect." On the other hand, the reminder effect may be either very small or non-existent suggesting that selection into software is even larger than the difference between the OLS and the IV estimates. In this sense, this difference between the OLS and the IV represents a lower bound for the degree of selection captured in the OLS estimate. We next estimate the upper bound.

5.2 Upper Bound of Selection

We take advantage of our two-step experimental design that allows us to directly observe the selection mechanism in the control group to estimate the upper bound of the selection effect. Consumers who were randomized to the control group did not know that they were in the control group until they logged into the experimental website. Since we can observe who in the control group logged into the website, we can measure the difference in outcomes between those who chose to access the software and those who did not. As discussed above, this difference represents a combination of selection and treatment effects in the control group. Under the assumption that the reminder screen did not generate a treatment effect among those individuals in the control group who chose to log in, the difference between those who did and those who did not log in to the website in the control group would represent the pure selection effect. Since in practice some of this difference may be due to the treatment effect of the reminder screen, this comparison gives us the upper bound of selection. Given the low impact of generic reminders that has been found in the broader literature, we believe the selection interpretation plays an important role (Ericson et al., 2017), but the difference likely includes some of both. To measure this upper bound, We estimate the following OLS regression among the control group individuals only:

$$Y_i = \xi_1 LOGIN_i + \xi_2 X_i + \epsilon_i \tag{6}$$

Panel B of table 9 reports the estimates. Individuals that logged into the software website - before knowing whether they were assigned to the treatment or the control arm - were 21 percentage points more likely to

switch plans than those that did not log in (column 1). They also had a 15 percentage point higher probability of choosing an expert recommended plan (column 6), and were saving $169 in expected total cost of their Part D plan (column 5). We did not observe differences in the choice process satisfaction, decision conflict score, or search time (columns 2, 3, 4).

Our results on the selective take-up of intervention software indicate that caution is warranted when interpreting the positive effects of algorithmic decision support software for the development of policy. While offering people algorithmic decision support affects their choices, it is also much more likely to attract "active shoppers" and thus could be a poorly targeted policy instrument for rolling out in the general population. Without additional targeted interventions encouraging those who are not active shoppers to use such a tool, algorithms may not reach those who would benefit most from them.

5.3 Selection on Treatment Effects

We next examine the importance of self selection into decision-support tools by comparing the likely benefits of algorithmic recommendations among those who enrolled in the trial relative to those who did not.

We use the results of the generalized random forest algorithm - as discussed in 4.3 - to predict (intent-to-treat) treatment effects on the full sample of individuals who were originally invited to participate in the experiment. Recall that we originally invited 29,451 individuals to participate in the study and that 4% took up the invitation and were randomized into three arms. While we do not have survey data for the original 29,451 individuals, we observe their administrative records which we used to analyze the selection into the experiment on observables in Table 1. We now use the same observables to predict treatment effects (for each treatment type) among all invited individuals. In Table 8 and Section 4.3 we have already characterized the heterogeneity in treatment effects. Here we examine whether there were systematic differences in predicted treatment effects between those who decided to participate in the experiment and those who did not.

Table 10 reports the results of a regression of the predicted treatment effect for each outcome of interest on an indicator that takes a value of one if the individual was *not* among those who participated in the experiment. We estimate these regressions separately for "Information + Expert" (Panel A) and "Information Only" (Panel B) treatment arms. We observe pronounced selection on treatment effects. Individuals who did not participate in the trial would have overall responded *more* to either type of intervention than those individuals who did. Individuals that chose not to participate would have been 3-4 percentage points more likely to switch plans than those who did participate (column 1). They would have also been slightly more satisfied with the choice process as the result of using the tool (column 2), would have saved approximately 10% more under the algorithmic recommendation treatment (column 5), and would have been up to 50% more likely to enroll in one of the expert recommended plans (column 6). At the same time, they would have been less likely to increase their search time beyond one hour as compared to those who did choose to participate in the experiment (column 4).

Figure 4 documents the non-linearity of the experimental take up as a function of predicted treatment effects. This figure plots the take-up rate of the experiment for each ventile of the predicted treatment effect. For the probability of switching plans, we observe that the take-up rate declines sharply with the estimated treatment

effect, suggesting that individuals that would have responded most to the software intervention (in terms of switching their plans), were least likely to participate in the experiment. The same holds true for cost savings (those who would have saved more are less likely to participate), although the pattern is slightly noisier.

5.4 Implications of Selection

Overall, our analyses provide strong evidence of selective take-up. As in many other settings, we document that more sophisticated consumers are more likely to shop for coverage and demand more information, in this case, in the form of accessing information tools. A main contribution of our study is to demonstrate that the expected benefits of algorithmic recommendations, in particular, appear to have the greatest benefits for those who are least likely to use them.

Our analyses provide some insight into the potential barriers to greater use of algorithms in the setting we study. We demonstrate empirically that the expected benefits of personalized information are negatively correlated with participation in the trial. Because consumers access information when the expected benefits of information exceed the costs of obtaining it (Stigler, 1961), our finding implies that, for those with relatively high estimated treatment effects, either the expected benefits of accessing information were low or the costs of search were high. In our empirical work, we find some evidence supporting the potential importance of the costs of search. In particular, those with relatively large estimated treatment effects had the lowest rates of EMR use, suggesting relatively low familiarity with information technology. In other words, consumers may have rationally chosen not to enroll in the trial because they correctly expected that for them the costs to them of using the online tool exceeded the benefit. Alternatively, consumers for whom the estimated treatment effects were largest may have systematically underestimated the benefits of information. For example, those with high estimated treatment effects may have underestimated the likelihood that an alternative plan would have covered their drugs more generously. A different version of this mechanism is that consumers observe the expected benefits with noise. If the variance in perceived benefits increases with the mean, then it is more likely that consumers with high benefits on average will underestimate their expected benefit relative to the cost. This interpretation is consistent with evidence on noise in consumer beliefs that we present in the next section. In sum, our results suggest that offering decision-support software without additional targeting efforts or even a requirement to go through algorithmic decision-support when enrolling into a plan, is unlikely to reach individuals who would have benefited most from having access to such software. We speculate that reducing the noise in perceived benefits of algorithmic support (for example, through mailings that first highlight individualized potential savings, as in Kling et al., 2012, and encourage consumers to seek out algorithmic support), may provide a way to improve targeting.

6 Theory and Welfare

In this section we develop a simple theoretical framework that allows us to conceptually differentiate between two related ideas: *information* versus (non-strategic) *advice*. While the former allows consumers to learn about

product features, the latter also helps consumers interpret these features. We map this framework into our trial data. We use the estimates to quantify the welfare effects of offering consumers an algorithm that provides them with information and/or advice.[8]

We argue that consumer choices may deviate from a full-information benchmark due to two conceptually distinct reasons. First, consumers may have imperfect information about the features of the products among which they are choosing. Second, consumers may have only noisy signals about the mapping of each product feature into utility. Uncertainty about utility weights is one way to capture the idea that consumers may not understand contract features even if they have perfect information about these features (Bhargava et al., 2017). Allowing for two sources of uncertainty implies that there are two types of information a consumer may acquire: (i) information about features that allows the consumer to *learn* about the good, and (ii) advice about the valuation of features that allows the consumer to *interpret* the value of the good. This conceptual distinction between information and non-strategic advice is related to several ideas in the prior literature. For example, Çelen et al. (2010) asked, in a laboratory experiment, whether the subjects would like to get advice or the underlying information. Further, a literature on advertising has made a related distinction between informative versus persuasive advertising (Braithwaite, 1928; Ackerberg, 2001). The general idea that external advice and information may alter preferences relates closely to the rich literature on persuasion (DellaVigna and Gentzkow, 2010), except in our setting advice transmission is non-strategic. The idea that consumers are unsure about their payoffs or may overvalue more salient characteristics of goods is common in the models with rational inattention (e.g., Steiner et al., 2017; Sallee, 2014; Matejka and McKay, 2015), salience and context-dependent choice (Bordalo et al., 2013), as well as experience goods (Riordan, 1986). In these frameworks, however, one usually does not distinguish between the uncertainty about product features and the uncertainty about the relative importance of these features for utility, which we argue is an important distinction in our setting.

6.1 Model

Consider consumer i who faces a choice set J of insurance contracts. Each contract j is characterized by a vector of characteristics ϕ_{ij} that can be individual-specific. Let $U_{ij}(\phi_{ij};\beta_i)$ be the utility that consumer i gets from choosing plan j with characteristics ϕ_{ij}. This utility depends on plan characteristics ϕ_{ij} and the parameters of the utility function for consumer i, β_i. Under perfect information about both ϕ_{ij} and β_i, consumer i chooses contract j^* such that U_{ij^*} is greater than U_{ij} for all other $j \in J$.

In practice, the consumer may only have a noisy prior about the characteristics of each plan. In other words, the elements of ϕ_{ij} may be observed imprecisely. Further, the consumer may be uncertain about how to aggregate the elements of ϕ_{ij} into utility-relevant objects. In other words, the elements of β_i may be observed imprecisely. For example, figuring out which drugs are covered by any given insurance plan is costly, as that

[8] Our goal here is to provide *one* potential framework that allows us to think about the systematic differences in behavior we observe across experimental arms. Alternative explanations for the differences in behavior exist and are equally plausible. For example, the differences in consumer behavior when they face the "expert" recommendation could stem from the framing effects, anchoring, or other ways of "coherent arbitrariness" in which the presentation of expert scores and highlighting of plans as "recommended" could change individual choice behavior and hence the preferences that we estimate (John G. Lynch, 1985; Ariely et al., 2003; DellaVigna, 2009, 2018).

information is frequently complicated and difficult to find. At the same time, obtaining information about which drugs are covered by a plan - i.e. obtaining a document from an insurer that lists all covered drugs - may not resolve consumer's uncertainty about how to interpret this information and consequently how much utility weight to assign to this feature of the product.

Denote consumer beliefs about vectors ϕ_{ij} and β_i with $\tilde{\phi_{ij}}$ and $\tilde{\beta}_i$. The consumer maximizes her utility given these beliefs and chooses a plan $\tilde{j}$ such that:

$$\tilde{j} = \underset{j}{\text{argmax}} \quad \tilde{\beta}_i \tilde{\phi_{ij}} \tag{7}$$

The welfare loss L from noisy beliefs is given by the differences in the underlying utility from plan j^* relative to plan $\tilde{j}$:

$$L = U_{i\tilde{j}} - U_{ij^*} \tag{8}$$

Let the wedge between beliefs about plan features and true features be ξ^{ϕ}, and the wedge between true utility weights and beliefs about weights be ξ^{β}. We can then re-write the decision utility as being (omitting individual-specific subscripts):

$$\tilde{U_j} = (\beta + \xi^{\beta}\beta)(\phi_j + \xi^{\phi}\phi_j) \tag{9}$$

Exposure to pure information about produce features can reduce the wedge in consumer beliefs about plan features, ξ^{ϕ}, but should not affect utility weights. Advice, on the other hand, is different from information, as it provides a way to interpret information in addition to information itself. We model non-strategic advice as a reduction in ξ^{β}, which improves consumer's signal about the mapping of features into utility. Let $1-\kappa$ denote the "strength" of a decision-support intervention that exposes consumers to information or information and advice. κ measures the share of ξ^{ϕ} and ξ^{β} that remain despite the intervention. Consumer's decision utility with a decision support intervention then becomes:

$$\tilde{U_j} = \begin{cases} (\beta + \xi^{\beta}\beta)(\phi_j + \kappa\xi^{\phi}\phi_j) & \text{if exposed to information} \\ (\beta + \kappa\xi^{\beta}\beta)(\phi_j + \kappa\xi^{\phi}\phi_j) & \text{if exposed to information and advice} \end{cases} \tag{10}$$

where $\kappa \in [0,1]$. If $\kappa = 0$, the decision support intervention completely eliminates the noise in beliefs, meaning that $\tilde{j} = j^*$ and $L = 0$. If $\kappa = 1$, the intervention has no effect on consumer beliefs and consumer choices.

To summarize, this simple framework provides us with a key basic insight. Any intervention aimed at helping consumers make choices can change their choices through two mechanisms: by either changing their beliefs about the features of the products, or by changing their utility weights for these features. The two mechanisms generate very different policy implications. If consumer choices are affected by noisy priors about how product features map into utility, then a policy of providing information about plan features will not generate any behavioral responses. In contrast, if consumers know exactly how to evaluate product features,

but have a hard time accessing that information, policies that improve information access may be effective. For example, because cost-sharing can be complex and vary by drug and plan, consumers may not have perfect knowledge of the cost-sharing features of their plans are. In contrast, they may be aware that their plan has a very high deductible, but not able to evaluate the implications of a high deductible for their utility. The distinction between the two mechanisms is of central practical relevance for complex financial products, where the knowledge of product features may not be enough for consumers to make informed decisions.

Our framework accommodates multiple types of consumer behaviors that have been documented in the literature. In particular, it offers a way to reconcile the divergent conclusions of two strains of work that have explored consumer choices in Medicare Part D specifically. The first set of papers (Abaluck and Gruber, 2011; Abaluck and Gruber, 2016) argues that consumers make choices that are inconsistent with rational decision-making. The second argues that consumers are behaving rationally and learn over time (Ketcham et al., 2012). Our framework demonstrates that both behaviors could in fact be taking place at the same time. The idea of choice inconsistencies in (Abaluck and Gruber, 2011; Abaluck and Gruber, 2016) can be thought of as a non-zero ξ^{β} - consumers observe deductibles, coverage in the gap, and other features of the plans, but have biased utility weights for these features. Ketcham et al. (2012); Ketcham et al. (2015), on the other hand, argue that consumer choices are improving over time. This could be true if original choices are affected by the noise in the knowledge about plan characteristics - ξ^{ϕ} - that decrease over time as consumers learn about product features. Learning about characteristics, however, doesn't preclude that the wedge in utility weights - ξ^{β} - and hence "inconsistent" choices continue to exist.

6.2 Estimation

Set-up The conceptual model outlined above can be directly mapped to an empirical discrete choice problem with random utility. We start with a standard discrete choice framework, in which consumer i is choosing a product j from the set of available products J. The consumer picks j that maximizes her decision utility that we empirically specify as follows:

$$u_{ij} = \beta_i \phi_{ij} + \epsilon_{ij} \tag{11}$$

Here, ϕ_{ij} is a vector of characteristics of product j that are allowed to be individual-specific. Vector β_i maps product characteristics into utility. An entry in vector β_i that multiplies a dollar-denominated feature, such as the expected out of pocket spending gives us the marginal utility of income that "translates" monetary objects into utils. This marginal utility of income can vary across individuals i. When re-normalized to the marginal utility of income, other entries in vector β_i, provide the measure of individual's willingness to pay for the corresponding product feature. ϵ_{ij} captures any consumer-product specific parts of utility that are not observable to the researcher, but are observable to the consumer and affect consumer choices.

In most applications, when estimating a discrete choice model of demand, researchers include product features ϕ_{ij} as they are observed to the researcher, which is usually an "objective" measure of these product features. This, however, may not be the ϕ_{ij} that enters consumer decision-making if consumers observe ϕ_{ij}

with some noise. Further, when estimating β_i from revealed preferences for product features, we would typically capture the utility weights that entered the decision utility function. The weights in the utility function, however, reflect only consumer's current information set and may be a noisy signal of the underlying welfare-relevant weights.

Following the argument in Section 6.1, consider the following reformulation of the standard utility specification that includes noise in features and utility weights. Adding multiplicative friction terms to Equation 11, we get:

$$u_{ij} = (\beta_i + \xi_i^{\beta}\beta_i)(\phi_{ij} + \xi_i^{\phi}\phi_{ij}) + \epsilon_{ij} \tag{12}$$

or re-arranging,

$$u_{ij} = (1 + \xi_i^{\beta})(1 + \xi_i^{\phi})\beta_i\phi_{ij} + \epsilon_{ij} \tag{13}$$

Is it possible to separate ξ_i^{β} and ξ_i^{ϕ} empirically? Conceptually, to do that we need an intervention that plausibly affects only ξ_i^{β} or ξ_i^{ϕ}. We argue that our two treatment arms provide us exactly with that type of variation. Arm "Information Only" provides individuals with personalized information about expected costs, CMS plan quality rating, and plan brands. Hence, in this arm, individuals receive information about expected out of pocket costs, but they do not receive any further guidance about how to combine different plan features into a utility function. In other words, for individuals enrolled in the "Information Only" arm, the treatment affects only ξ_i^{ϕ}.

Individuals in the "Information + Expert" arm receive the same information as those in "Information Only" arm, but they also receive the personalized expert scores and a recommendation to choose one of three plans with the highest expert scores. The expert score does not provide *additional* information about plan features, as it is a combination of out of pocket cost prediction and the star rating. However, it provides a suggestion to the consumer of how to weight plan features by combining the personalized cost estimate with the plan-level star rating into a one-dimensional metric. Hence, we can interpret arm "Information + Expert" as changing both the information about features and the utility weights that consumers ought to place on these features, i.e. changing both $\xi_i^{\phi}$and ξ_i^{β}. This implies that by comparing the choice behavior across control arm and treatment arm "Information Only," and then treatment arm "Information + Expert" we can quantify the presence of ξ_i^{ϕ} and ξ_i^{β} in consumer's decision utility.

To illustrate our approach, consider an example with $\kappa = 0$, so that an informational intervention completely removes noise terms. The decision utility of individual i from choosing plan j in the control arm is then given by:

$$u_{ij} = (1 + \xi_i^{\beta})(1 + \xi_i^{\phi})\beta_i\phi_{ij} + \epsilon_{ij} \tag{14}$$

While for individual i choosing plan j in the "Information Only" arm, utility is:

$$u_{ij} = (1 + \xi_i^{\beta})\beta_i\phi_{ij} + \epsilon_{ij} \tag{15}$$

And similarly, utility for an individual in the "Information + Expert" arm becomes:

$$u_{ij} = \beta_i \phi_{ij} + \epsilon_{ij} \tag{16}$$

The latter corresponds to the standard discrete choice utility that we started with in Equation 11, as it "restores" the case of complete information.

We can now proceed to estimate Equations 14 to 16. Our goal is to estimate β_i, ξ_i^β and ξ_i^ϕ. We achieve this by estimating how revealed preferences for ϕ_{ij} vary across experimental arms. Assuming that, by the virtue of randomization, there should be no latent differences in utility weights across the experimental arms (i.e. no differences in underlying β_i), we will attribute any variation in estimated preferences across arms to differences in beliefs.[9] Comparing utility weight estimates between the "Information Only" and "Information + Expert" arms allows us to measure how much of the behavioral change in response to the intervention is coming from changes in ξ_i^β versus changes in ξ_i^ϕ.

We estimate the following specification for consumer i in year t (recall that we observe consumer plan choices at the baseline and endline of the experiment, which spans two years of choices):

$$u_{ijt} = \widetilde{\beta}_i \phi_{ijt} + \epsilon_{ij}, \;\; \epsilon_{ijt} \sim \text{iid EV Type I} \tag{17}$$

We allow for unobserved heterogeneity in consumer preferences that is assumed to have a normal distribution. We also assume that the part of utility not observed by the researcher is distributed iid with Type 1 extreme value distribution. We let ϕ_{ij} include the expected total cost of the plan, CMS star rating and indicators for one of three most popular insurer brands. This is the full set of plan features that study participants observe on the main page of the experimental software in the two treatment arms (see Figure 1). This information is also in principle readily available to participants in the control arm from the government-run online Medicare Part D calculator. To increase the precision of our estimates given the small sample size, we pool observations from all three experimental arms and years 2016 and 2017 choices of plans. The specification then becomes:

$$u_{ijt} = \mu_1 Cost_{ijt} + \mu_2 CMSStar_{jt} + \mu_3 AARP_{jt} + \mu_4 Humana_{jt} + \mu_5 Silverscript_{jt} + \epsilon_{ijt} \tag{18}$$

$$\mu_n = \psi_n + \lambda_n I + \eta_n E \quad \forall\, n \in [1, 5] \tag{19}$$

Estimating this model allows us to quantify the wedges in beliefs for each plan feature. First, we aggregate our estimates to derive one revealed preference parameter for each plan feature in each experimental arm. Consider the expected costs. For this feature, the estimate of revealed preferences in the control arm $\hat{\beta_1}^C$ is equal to $\hat{\psi}_1$. For treatment arm "Information Only", $\hat{\beta_1}^I = \hat{\psi}_1 + \hat{\lambda}_1$. For treatment arm "Information + Expert," $\hat{\beta_1}^E = \hat{\psi}_1 + \hat{\eta}_1$.

Now we map the three estimates of revealed preferences in each arm into the underlying model parameters.

[9] We verify this assumption empirically by estimating the differences in revealed preference parameters at the baseline, prior to the intervention. We find no differences in estimated β_i:s across experimental arms.

For control arm:

$$\hat{\beta_1}^C = (1+\xi_i^\beta)(1+\xi_i^\phi)\beta_i \tag{20}$$

For treatment arm "Information Only":

$$\hat{\beta_1}^I = (1+\xi_i^\beta)\beta_i \tag{21}$$

And finally, for treatment arm "Information + Expert":

$$\hat{\beta_1}^E = \beta_i \tag{22}$$

These are three equation in three unknowns that give us β, ξ_i^β, and ξ_i^ϕ once we have $\hat{\beta_1}^C$, $\hat{\beta_1}^I$, and $\hat{\beta_1}^E$.

Estimation results Panel A of Table 11 reports model estimates. Column (1) reports ψ^1, τ^1, and η^1 - coefficients on the "cost" feature of the plans. We estimate τ^1 to be negative and large (relative to the control group) in absolute value, suggesting that "Information Only" intervention makes consumers appear more sensitive to costs. The change in the sensitivity to cost is substantially less pronounced under the "Information + Expert" treatment. Column (2) in turn suggests that consumers become more sensitive to CMS star rating under "Information Only" intervention, while columns (3) to (5) suggest that the intervention changes consumers' ranking of brands. We observe similar patterns for the "Information + Expert" arm, except that it makes AARP-branded plans appear less desirable to consumers.

To interpret these estimates in the context of our conceptual framework, we substitute the point estimates into Equations 20 to 22 to get (for the cost feature as an example):

$$-0.13 = (1+\xi_i^\beta)(1+\xi_i^\phi)\beta_i \tag{23}$$

$$-0.21 = (1+\xi_i^\beta)\beta_i \tag{24}$$

$$-0.17 = \beta_i \tag{25}$$

It follows that $1+\xi_i^\beta$=1.27 and $1+\xi_i^\phi$=0.62, as we report in Panel B.1. This in turn suggests that consumers tend to underestimate the expected costs of plans, but have a higher willingness to pay for each \$100 reduction in the out of pocket costs than they would under full information. Panel B of Table 11 also reports similar computations for other plan features. Except for the Silverscript brand indicator, we find a similar qualitative pattern across all features - that consumers have a negative ξ^ϕ, underestimating the features of available plans (for the brand indicators, this can be interpreted as noisy signal about the probability that any given plan has a particular brand), and yet have a positive ξ^β, suggesting a higher - than under full information - willingness to pay for each feature.

In Panel B.2 we examine how our results change when we assume that the exposure to either treatment arm only "corrects" 80% of noise in beliefs. The magnitude of noise estimates change accordingly, but provide very

similar qualitative take away. For example, we still find that individuals underestimate the total costs they are likely to face in a plan, and overestimate how the costs map into the utility function.

Taken together these estimates illustrate that changes in utility weights that could be impacted by advice may have a substantial effect on consumers' behavioral response. We also conclude that the experimental data is consistent with a hypothesis that consumers may have noisy priors about both the product features and their interpretation, or utility weights, on these features.

6.3 Welfare

We next use our estimates to shed light on how the provision of information may affect consumer welfare. To accomplish this, we simulate consumer choices and the corresponding welfare loss function from equation 8 under four scenarios Recall that we defined the welfare loss as the difference between the "true" utility the consumer experiences from the plan chosen under noisy beliefs ($\tilde{j}$) and the plan that would have been chosen under perfect information (j^*).

The first scenario simulates consumer choices using the preferences as estimated under the "Information + Expert" treatment arm. Put differently, this scenario switches off both $1+\xi_i^{\beta}$ and $1+\xi_i^{\phi}$. We take consumer choices and their utility in this scenario as our normative benchmark, U_{ij^*}. In the other three simulation scenarios we switch on $1+\xi_i^{\beta}$, or $1+\xi_i^{\phi}$, or both, respectively. Each of these simulations with wedges in beliefs switched on gives us a $\tilde{j}$, allowing us to compute $U_{i\tilde{j}}$ and $L = U_{i\tilde{j}} - U_{ij^*}$. In essence, this exercise measures how much $1+\xi_i^{\beta}$ and $1+\xi_i^{\phi}$ alter the ordinal ranking of plans in utility terms. If consumers have noisy beliefs, but these beliefs lead them to choose the same product as they would have under perfectly informed beliefs, then there is no welfare loss from the noise in beliefs and informational interventions would be an unnecessary cost.

Table 12 reports our simulation results. We simulate our model for all 29,451 individuals who were invited to participate in the trial. In Panel A we report several moments of the distribution of surplus loss (L) from relying on "noisy" beliefs. On average, the welfare loss is relatively modest. We estimate the average loss to vary between \$48 to \$68 depending on which wedges in beliefs we allow for. This represents a 4.1% to 6.8% loss in utility. The relative loss is the highest when we allow for wedges in both types of beliefs, which is intuitive, as that increases the likelihood that the wedges change the ordinal raking of plans.

The modest average loss masks a substantial amount of heterogeneity in how much the noise in beliefs about product characteristics or the mapping of characteristics into the utility function affects consumer utility. For half of the consumers, the noise in the utility function does not in fact lead to any surplus losses. These consumers end up choosing the same plan across all specifications of the utility function. For some consumers, however, the noise in beliefs lead to significant welfare losses, both in absolute and relative terms. For these consumers, noise in beliefs lead them to choose a plan that is far from the optimum. At the 95th percentile of the distribution, individuals that choose plans according to preferences as estimated from the control group (i.e. those that allow for both sources of noise in beliefs), would lose nearly \$300 or 15% of their benchmark normative utility. This is a significant loss, equal to nearly six monthly premiums.

This analysis suggests that while for many consumers misconceptions in their beliefs about plan features

or the mapping of features into the utility function is inconsequential, some consumers experience significant losses and choose sub-optimal plans when they don't have perfect information. An cost-effective informational intervention would want to target consumers that experience the highest welfare losses. Panel B of Table 12, however, reveals that offering consumers a decision-support software - i.e. self-targeting - would not lead to optimal targeting. Among consumers who were offered to participate in the trial, those who we predict would have benefited the most, were not more likely (and if anything were slightly less likely) to participate. This finding is consistent with our earlier results on selection outlined in Section 5 and once again underscores the challenge of targeting an informational intervention in this domain.

7 Conclusion

Personalized decision support software providing consumers with varying levels of decision autonomy is increasingly prevalent in many markets. In theory, delegating consumer decisions to individualized predictive algorithms could significantly alter consumption patterns, especially in complex decision environments. The rise of algorithms could thus substantially alter market allocations across a range of settings. In practice, we know little about how consumers interact with algorithms or which type of consumers choose to engage in such interactions in the first place. Much of the research on algorithms to date has focused on examining the potential for strategic or unintended biases of algorithmic decision support, while little evidence exists on consumer responses to this new technology.

In this paper, we provide novel evidence from a randomized-controlled study in which older adults were offered individualized decision support software for the choice of prescription drug insurance plans. The treated groups received two versions of the software. One version offered a more intensive intervention by providing consumers with "expert" machine-generated one-dimensional scores for each choice option. The other treated group received personalized information about the expected total cost in each plan and a plan-level quality assessment, but was not given the expert score summarizing this information. The control group was offered a reminder.

We draw three main conclusions from our experimental results. First, exposure to the decision support tool changed consumer behavior. More specifically, providing (individualized) information coupled with a one-dimensional algorithmic recommendation significantly increased the probability of plan switching, the time spent on the choice process, the expected cost-savings and self-reported satisfaction with the choice process. While providing individuals with individualized information without the one-dimensional algorithmic recommendation moved the outcomes in a similar direction, the magnitudes of the effects were less pronounced economically and statistically.

Second, there is strong selection into the use of decision support software. We document two types of selection. We find that individuals who actually used the softward conditional on having access to it were inherently more active shoppers who likely would have changed their plan and chosen a lower cost plan without an intervention. Quantitatively, this selection effect is close in magnitude to the treatment effect, allowing us

to conclude that there is strong complementarity in the willingness to shop actively for financial products and the interest in decision support algorithms. Further, we find that individuals whom we predict would have responded most strongly to the treatment intervention, were the least likely to enroll in the trial. While the findings of strong selection do not invalidate the idea that intuitive tools with clear, simplified, algorithmic recommendations could improve choices if rolled out in a general population, they do suggest that a policy of merely offering algorithmic recommendations within a software tool is unlikely to reach those who would respond the most to them. Hence, more targeted and intensive interventions may be required for populations who are unlikely to take-up algorithmic advice but are likely to benefit from it.

Finally, using a simple model of consumer decision-making that offers a lens through which to interpret our findings, we find that the behavioral responses that we observe in the data are driven by both the (i) updating of consumers' signals about the features of the products, and (ii) adjustments in consumers' utility weights - or mapping - of these features into utility. The noise in consumer beliefs leads to relatively small welfare losses, on average; however, a small set of consumers experience significant losses in utility of up to 15%. The distinction between consumer's misconception about the characteristics of a financial object versus the mapping of object features into utility is important for interpreting the findings on consumer "mistakes" in a variety of financial settings. This distinction is also crucial for policy-making in the realm of algorithmic advice. Existing algorithmic recommendations not only allow consumers to learn about product features, but usually also aim to change how consumers interpret the value of these features. Our results indicate that the interpretation channel is quantitatively important in the setting we examine. While the ability of algorithms to change individual preferences creates opportunities to improve consumer choices, it also raises concerns over the possibility that algorithms may influence decision-making in ways that have poorly understood or unintended consequence for consumers. Algorithms may generate biases in decision making, either strategic or inadvertent, that have important downstream consequences. Because consumers are responsive to algorithmic recommendations, it will be increasingly important not only to understand how consumers respond to algorithms but also the implications of those responses for societal welfare.

References

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

Abaluck, Jason and Jonathan Gruber, "Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance," *American Economic Review*, 2016, *106* (8), 2145–2184.

Ackerberg, Daniel A., "Empirically Distinguishing Informative and Prestige Effects of Advertising," *The RAND Journal of Economics*, 2001, *32* (2), 316–333.

Agrawal, Ajay K, Joshua S Gans, and Avi Goldfarb, "Prediction, Judgment and Complexity: A Theory of Decision-Making and Artificial Intelligence," in Ajay K Agrawal, Joshua S Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, 2019, pp. 89–110.

Alsan, Marcella and Marianne Wanamaker, "Tuskegee and the health of black men," *Quarterly Journal of Economics*, 2018, *133* (1), 407–455.

Ariely, Dan, George Loewenstein, and Drazen Prelec, ""Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences," *The Quarterly Journal of Economics*, 2003, *118* (1), 73–106.

Asher, Sam, Denis Nekipelov, Paul Novosad, and Stephen P. Ryan, "Moment Forests," 2018.

Athey, Susan and Stefan Wager, "Estimating Treatment Effects with Causal Forests: An Application," *Observational Studies*, forthcoming.

_ , Julie Tibshirani, and Stefan Wager, "Generalized random forests," *Annals of Statistics*, 2019, *47* (2), 1179–1203.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian, "Behavioral Household Finance," in Douglas B. Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics*, Elsevier, forthcoming.

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu, "The Role of Application Assistance and Information in College Decisions: Results from the HR Block FAFSA Experiment," *The Quarterly Journal of Economics*, 07 2012, *127* (3), 1205–1242.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor, "Choose to Lose: Health Plan Choices from a Menu with Dominated Options," *The Quarterly Journal of Economics*, 2017, *132* (3), 1319–1372.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience and Consumer Choice," *Journal of Political Economy*, 2013, *121* (5), 803–843.

Braithwaite, Dorothea, "The Economic Effects of Advertisement," *The Economic Journal*, 1928, *38* (149), 16–37.

Cafferata, Gail Lee, "Knowledge of Their Health Insurance Coverage by the Elderly," *Medical Care*, 1984, *22* (9), 835–847.

Çelen, Boğaçhan, Shachar Kariv, and Andrew Schotter, "An Experimental Test of Advice and Social Learning," *Management Science*, 2010, *56* (10), 1687–1701.

Centers for Medicare & Medicaid Services, "CMS Fast Facts," Technical Report February 2019.

Centers for Medicare & Medicaid Services, "Closing the Coverage Gap - Medicare Prescription Drugs are Becoming More Affordable," Technical Report 2015.

Charlson, Mary E., Peter Pompei, Kathy L. Ales, and C.Ronald MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, 1987, *40* (5), 373 – 383.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and IvÃ¡n FernÃ¡ndez-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," Working Paper 24678, National Bureau of Economic Research June 2018.

Cubanski, Juliette, Tricia Neuman, and Anthony Damico, "Closing the Medicare Part D Coverage Gap: Trends, Recent Changes, and What's Ahead," 2018.

Davis, Jonathan M. V. and Sara B. Heller, "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs," *American Economic Review*, 2017, *107* (5), 546–550.

DellaVigna, Stefano, "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, June 2009, *47* (2), 315–72.

_ , "Chapter 7 - Structural Behavioral Economics," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics - Foundations and Applications 1*, Vol. 1, North-Holland, 2018, pp. 613 – 723.

DellaVigna, Stefano and Matthew Gentzkow, "Persuasion: Empirical Evidence," *Annual Review of Economics*, 2010, *2* (1), 643–669.

Duggan, Mark, Patrick Healy, and Fiona Scott Morton, "Providing Prescription Drug Coverage to the Elderly: America's Experiment with Medicare Part D," *Journal of Economic Perspectives*, 2008, *22(4)*, 69–92.

Einav, Liran and Jonathan Levin, "Economics in the age of big data," *Science*, 2014, *346* (6210).

Ericson, Keith M. Marzilli and Amanda Starc, "How product standardization affects choice: Evidence from the Massachusetts Health Insurance Exchange," *Journal of Health Economics*, 2016, *50*, 71 – 85.

Ericson, Keith Marzilli, Jon Kingsdale, Tim Layton, and Adam Sacarny, "Nudging leads consumers in Colorado to shop but not switch ACA Marketplace Plans," *Health Affairs*, 2017, *36* (2), 311–319.

Ericson, Keith Marzilli, "Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange," *American Economic Journal: Economic Policy*, 2014, *6* (1), 38–64.

Finkelstein, Amy and Matthew J Notowidigdo, "Take-up and Targeting: Experimental Evidence from SNAP," *The Quarterly Journal of Economics*, May 2019.

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

Handel, Benjamin and Jonathan Kolstad, "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 2015, *105(8)*, 2449–2500.

Harris, Katherine M and Michael P Keane, "A model of health plan choice: Inferring preferences and perceptions from a combination of revealed preference and attitudinal data," *Journal of Econometrics*, 1999, *89* (1-2), 131–157.

Heiss, Florian, Adam Leive, Daniel McFadden, and Joachim Winter, "Plan Selection in Medicare Part D: Evidence from Administrative Data," *Journal of Health Economics*, 2013, *32* (6), 1325–1344.

_ **, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou**, "Inattention and Switching Costs as Sources of Inertia in Medicare Part D," Working Paper 22765, National Bureau of Economic Research October 2016.

Heiss, Florian, Daniel Mcfadden, and Joachim Winter, "Mind the Gap! Consumer Perceptions and Choices of Medicare Part D Prescription," in "Research Findings in the Economics of Aging," The University of Chicago Press, 2010, pp. 413–481.

Hitsch, Günter J and Sanjog Misra, "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation," 2018.

Ho, Kate, Joseph Hogan, and Fiona Scott Morton, "The impact of consumer inattention on insurer pricing in the Medicare Part D program," *RAND Journal of Economics*, 2017, *48* (4), 877–905.

Kaiser Family Foundation, "An Overview of the Medicare Part D Prescription Drug Benefit," Technical Report October 2018.

Keane, Michael P. and Susan Thorp, "Complex Decision Making," in "Handbook of the Economics of Population Aging," 1 ed., Elsevier B.V., 2016, pp. 661–709.

Ketcham, Jonathan D., Nicolai V. Kuminoff, and Christopher A. Powers, "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Comment," *American Economic Review*, 2016, *106* (12), 3932–3961.

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

Ketcham, Jonathan D., Claudio Lucarelli, and Christopher A. Powers, "Paying Attention or Paying Too Much in Medicare Part D," *American Economic Review*, 2015, *105* (1), 204–233.

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

Linder, Suzanne K., Paul R. Swank, Sally W. Vernon, Patricia D. Mullen, Robert O. Morgan, and Robert J. Volk, "Validity of a Low Literacy Version of the Decisional Conflict Scale," *Patient Education and Counseling*, 2011, *85* (3), 521–524.

Liu, Jiaying, Xiangjie Kong, Feng Xia, Xiaomei Bai, Lei Wang, Qing Qing, and Ivan Lee, "Artificial intelligence in the 21st century," *IEEE Access*, 2018, *6*, 34403–34421.

Loewenstein, George, Joelle Y Friedman, Barbara McGill, Sarah Ahmad, Suzanne Linck, Stacey Sinkula, John Beshears, James J Choi, Jonathan Kolstad, David Laibson, Brigitte C Madrian, John A List, and Kevin G Volpp, "Consumers' misunderstanding of health insurance," *Journal of Health Economics*, 2013, *32* (5), 850–862.

Lynch, JR. John G., "Uniqueness Issues in the Decompositional Modeling of Multiattribute Overall Evaluations: An Information Integration Perspective," *Journal of Marketing Research*, 1985, *22* (1), 1–19.

Matejka, Filip and Alisdair McKay, "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, January 2015, *105* (1), 272–98.

O'Connor, Annette M., "Validation of a decisional conflict scale," *Medical decision making*, 1995, *15* (1), 25–30.

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.

Riordan, Michael H, "Monopolistic Competition with Experience Goods," *The Quarterly Journal of Economics*, 1986, *101* (2), 265–279.

Sallee, James M., "Rational Inattention and Energy Efficiency," *The Journal of Law and Economics*, 2014, *57* (3), 781–820.

Samuelson, William and Richard Zeckhauser, "Status Quo Bias in Decision Making," *Journal of Risk and Uncertainty*, 1988, *1* (1), 7–59.

Sinaiko, Anna D and Richard A Hirth, "Consumers, health insurance and dominated choices," *Journal of Health Economics*, 2011, *30* (2), 450–457.

Steiner, Jakub, Colin Stewart, and Filip Matějka, "Rational Inattention Dynamics: Inertia and Delay in Decision-Making," *Econometrica*, 2017, *85* (2), 521–553.

Stigler, George J., "The Economics of Information," *Journal of Political Economy*, 1961, *69* (3), 213–225.

Stults, Cheryl D., Alison Baskin, Ming Tai-Seale, and M. Kate Bundorf, "Patient Experiences in Selecting a Medicare Part D Prescription Drug Plan," *Journal of Patient Experience*, 2018, *5* (2), 147–152.

Stults, Cheryl D., Sayeh Fattahi, Amy Meehan, M. Kate Bundorf, Albert S. Chan, Ting Pun, and Ming Tai-Seale, "Comparative Usability Study of a Newly Created Patient-Centered Tool and Medicare.gov Plan Finder to Help Medicare Beneficiaries Choose Prescription Drug Plans," *Journal of Patient Experience*, 2018, *6* (1), 81–86.

Tai-Seale, Ming, Caroline J. Wilson, Laura Panattoni, Nidhi Kohli, Ashley Stone, Dorothy Y. Hung, and Sukyung Chung, "Leveraging electronic health records to develop measurements for processes of care," *Health Services Research*, 2014, *49* (2), 628–644.

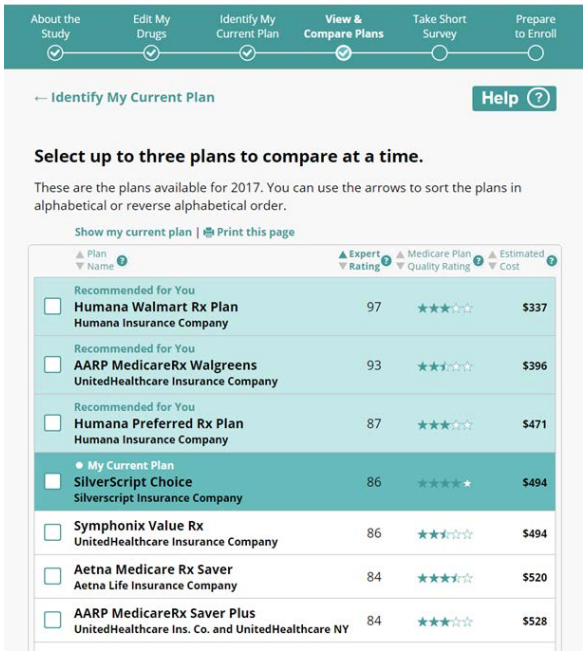
_ , N. Lance Downing, Veena Goel Jones, Richard V. Milani, Beiqun Zhao, Brian Clay, Christopher Demuth Sharp, Albert Solomon Chan, and Christopher A. Longhurst, "Technology-Enabled Consumer Engagement: Promising Practices At Four Health Care Delivery Organizations," *Health Affairs*, 2019, *38* (3), 383–390. PMID: 30830826.

Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

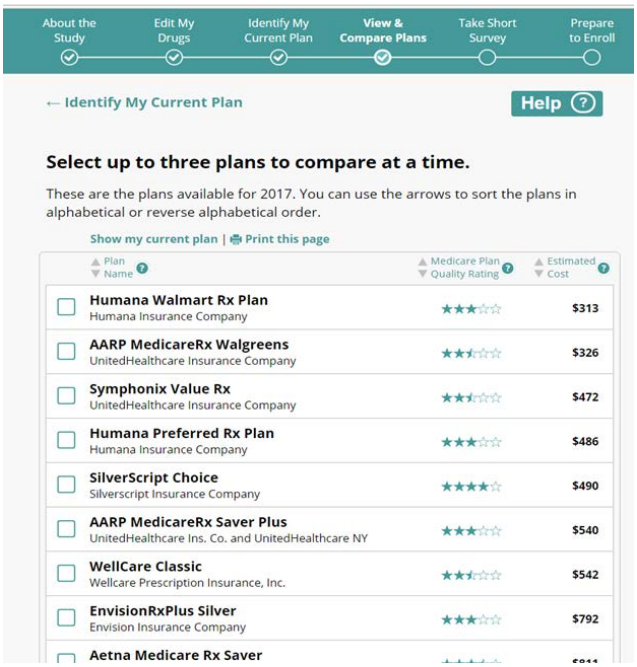
Figures and Tables

Figure 1: User Interface by Experimental Arm

A. Information + Expert Arm



B. Information Only Arm



C. Control Arm

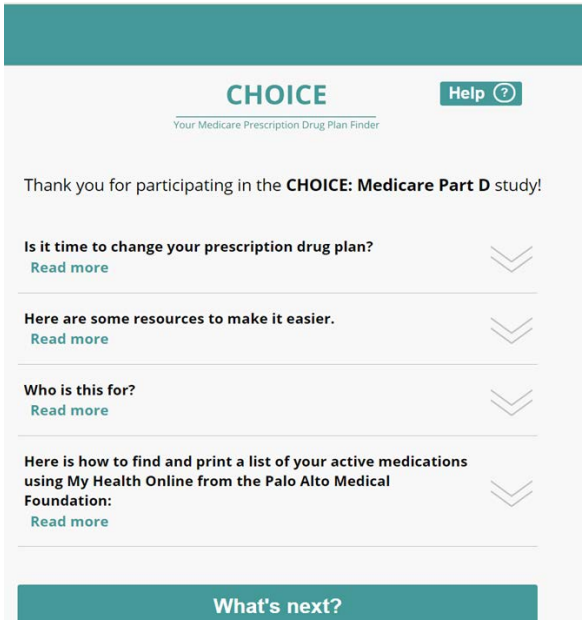


Figure 2: Experimental Design

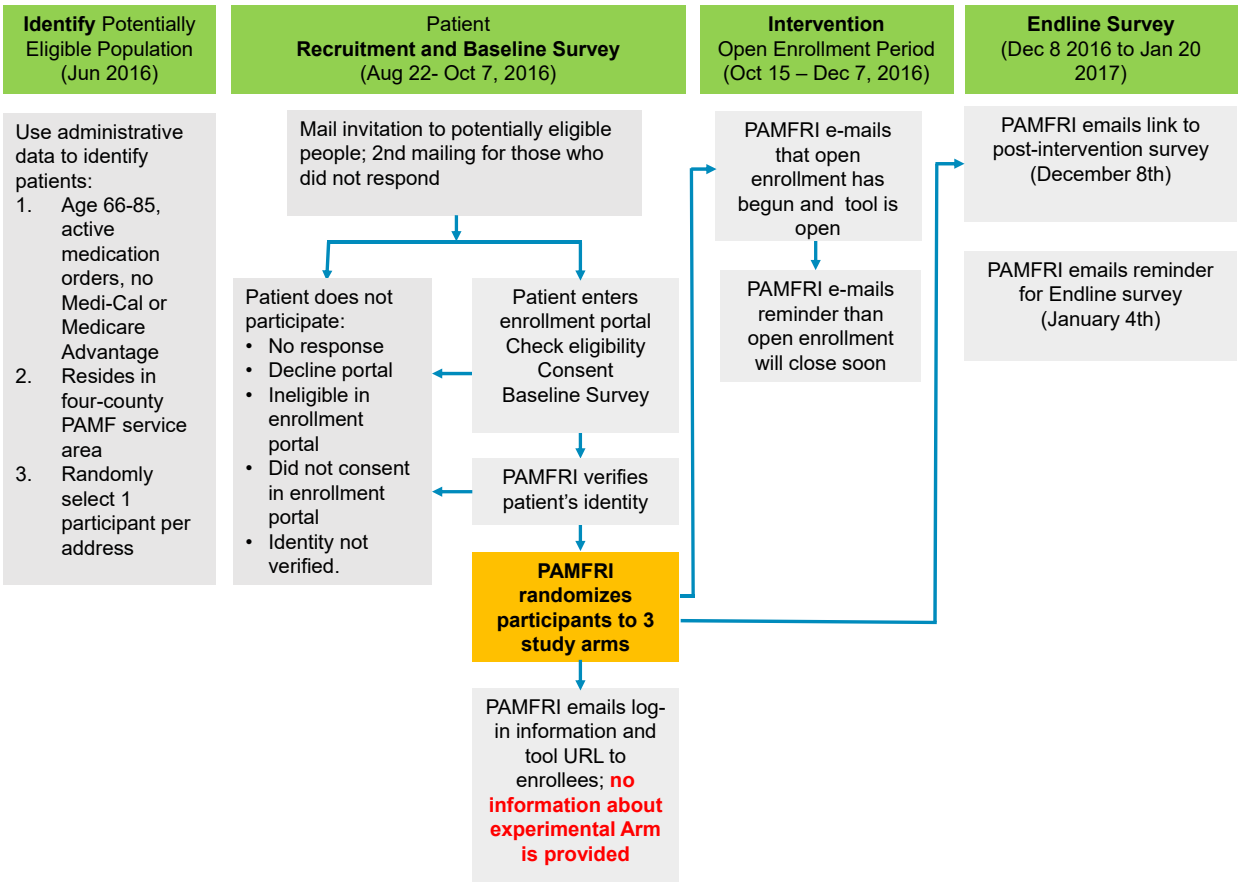
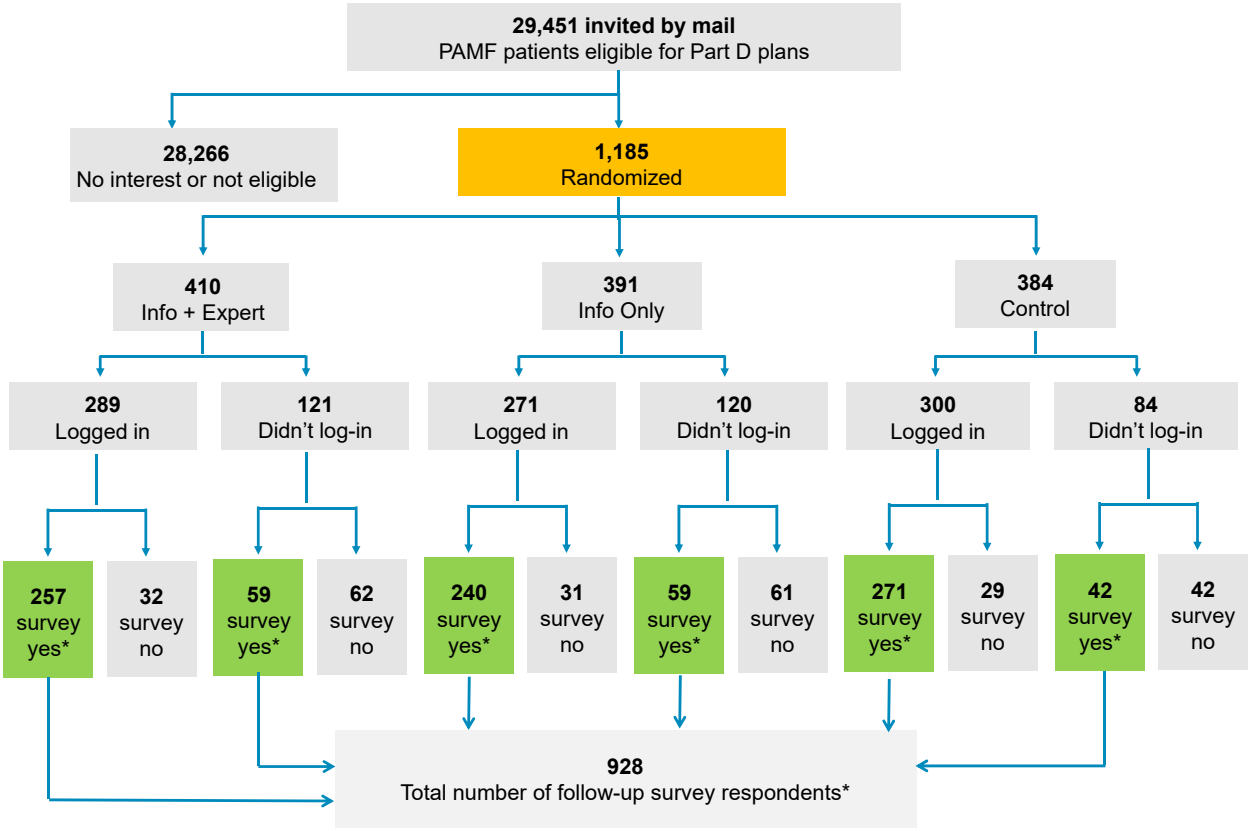


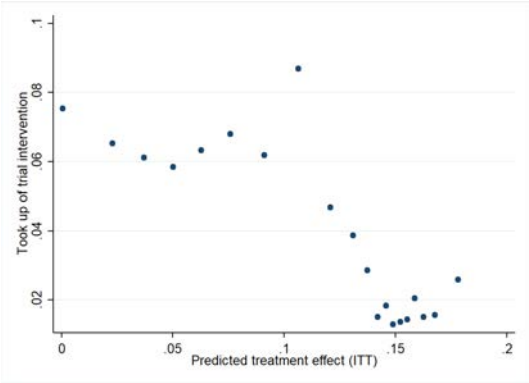
Figure 3: Enrollment Flow



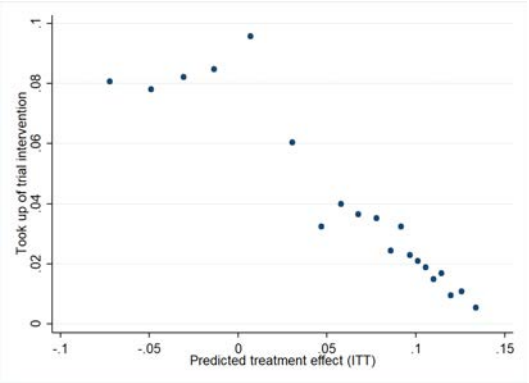
* Number of participants that responded to at least one survey question by pre-specified cutoff date

Figure 4: Take-up of Experiment by Predicted Treatment Effect

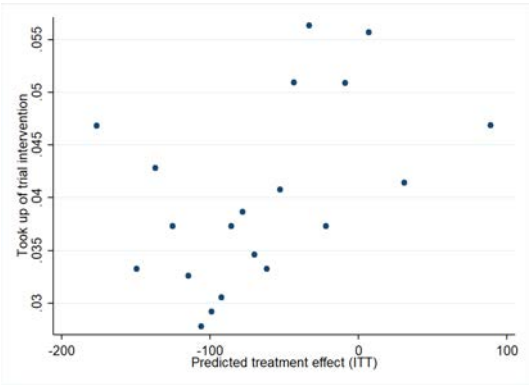
A. Switched Plans: Info + Expert



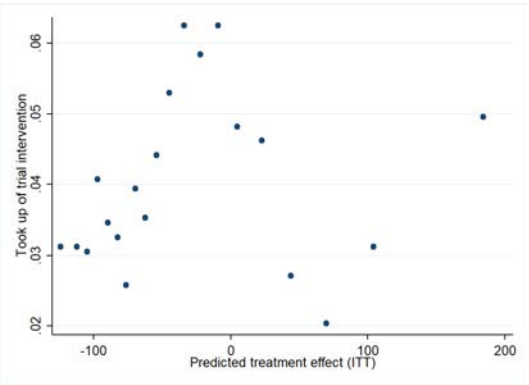
B. Switched Plans: Info Only



C. Change in expected cost: Info + Expert



D. Change in expected cost: Info Only



The figures plot the relationship between the probability of participating in the experiment and predicted treatment effects in the full sample of 29,451 individuals that were invited to participate. For these individuals we observe the demographics that are recorded in administrative data, allowing us to estimate treatment effects for this sample. Individual-level treatment effects of offering decision-support software are estimated using the generalized random forest (GRF) algorithm (Wager and Athey, 2018) as described in the text. Panels A and C report the results for "Information + Expert" arm; Panels B and D for "Information Only" arm. Panels A and B plot the probability of signing up for the experiment as a function of treatment effects for the outcome that is an indicator for whether an individual changed plans (outcome in column 1 of Table 6). Panels C and D plot the probability of signing up for the experiment as a function of predicted treatment effects for the change in expected total cost of the plan (outcome in column 5 of Table 6). Each figure is a binned scatterplot, where the outcome on the y-axis is computed within each ventile-sized bin of the treatment effect recorded on the x-axis.

Table 1: Selection into Experiment

	Age	Female	Non-White‡	Married	Income, $'000†	Share College†	Number Drugs	Charlson Score	Any EMR Use§	Intensity of EMR Use§~
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Randomized	-1.68***	-0.04**	-0.13***	0.07***	5.83***	0.04***	0.08	-0.16***	0.27***	3.74***
	(0.14)	(0.01)	(0.01)	(0.01)	(1.34)	(0.01)	(0.09)	(0.04)	(0.01)	(0.23)
No. of Obs.	29451	29451	29451	29451	29451	29451	29451	29451	29451	29451
Mean of Dep. Var.	73.96	0.54	0.35	0.54	106.81	0.54	4.45	1.16	0.69	3.30
Std. Dev. Of Dep. Var.	5.21	0.50	0.48	0.50	45.85	0.20	3.17	1.53	0.46	6.01

Table shows the relationship between baseline demographic characteristics of individuals and their take-up of the offer to participate in the experiment. 29,451 individuals were invited to participate. 1,185 entered the on-line enrollment portal, verified that they were eligible to participate, participated in a pre-enrollment survey and authenticated their identity. These individuals were randomized across three experimental arms. In columns (1) through (10) we report the results of separate regressions of each baseline demographic characteristic as the dependent variable on the indicator variable for whether an individual was a part of the 1,185 people that were randomized across arms. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

‡ Non-white includes "other" and missing responses

† Computed at census tract level

§ Measured within 3 years prior to the intervention

~ Number of strands of electronic conversations

Table 2: Randomization - Balance on Observables

	Age	Female	Non-White‡	Married	Income, $'000†	Share College†	Number Drugs	Charlson Score	Any EMR Use§	Intensity of EMR Use§~
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Information + Expert	-0.68*	-0.04	-0.03	0.06	-1.29	0.01	0.18	0.12	0.00	1.28*
	(0.33)	(0.04)	(0.03)	(0.03)	(3.23)	(0.01)	(0.23)	(0.10)	(0.02)	(0.55)
Information Only	-0.70*	-0.04	0.00	0.04	-3.57	-0.01	-0.00	0.01	0.01	0.91
	(0.33)	(0.04)	(0.03)	(0.04)	(3.30)	(0.01)	(0.21)	(0.10)	(0.01)	(0.51)
Mean of Dep. Var. in Control	72.81	0.53	0.23	0.57	114.02	0.59	4.46	0.96	0.95	6.15
No. of Obs.	1185	1185	1185	1185	1185	1185	1185	1185	1185	1185
Mean of Dep. Var.	72.35	0.50	0.22	0.60	112.40	0.59	4.52	1.01	0.96	6.89
Std. Dev. Of Dep. Var.	4.56	0.50	0.41	0.49	45.18	0.19	3.07	1.36	0.21	7.91
F-test across Arms, p-value	0.95	0.98	0.34	0.65	0.47	0.14	0.40	0.28	0.58	0.54

Table shows the relationship between baseline demographic characteristics of individuals who participated in the experiment (1,185 individuals) and their experimental arm assignment. Individuals were randomized across three experimental arms. In columns (1) through (10) we report the results of separate regressions of each baseline demographic characteristic as the dependent variable on two indicator variables representing the treatment arms, and a constant that captures the average value of the dependent variable in the control arm. We report the coefficients on the indicators for being randomized into treatment arms. The last row reports the F-test for the difference in the coefficients on the two treatment arm indicators. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

‡ Non-white includes "other" and missing responses

† Computed at census tract level

§ Measured within 3 years prior to the intervention

~ Number of strands of electronic conversations

Table 3: Attrition at Endline Survey

	Age	Female	Non-White‡	Married	Income, $'000†	Share College†	Number Drugs	Charlson Score	Any EMR Use§	Intensity of EMR Use§~
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Responded to endline survey	-0.32	0.00	-0.09**	0.03	3.32	0.04*	-0.16	0.04	0.03	0.57
	(0.32)	(0.04)	(0.03)	(0.03)	(3.26)	(0.01)	(0.22)	(0.09)	(0.02)	(0.55)
No. of Obs.	1185	1185	1185	1185	1185	1185	1185	1185	1185	1185
Mean of Dep. Var.	72.35	0.50	0.22	0.60	112.40	0.59	4.52	1.01	0.96	6.89
Std. Dev. Of Dep. Var.	4.56	0.50	0.41	0.49	45.18	0.19	3.07	1.36	0.21	7.91

Table shows the relationship between baseline demographic characteristics of randomized individuals and their participation in the endline survey, defined as responding to at least one endline survey question by the pre-specified cutoff date. 1,185 individuals were invited to complete the endline survey; 928 individuals responded to at least one question by the cutoff date. In columns (1) through (10) we report the results of separate regressions of each baseline demographic characteristic as the dependent variable on the indicator variable for whether an individual responded to at least one endline survey question. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

‡ Non-white includes "other" and missing responses

† Computed at census tract level

§ Measured within 3 years prior to the intervention

~ Number of strands of electronic conversations

Table 4: Attrition at Endline Survey by Experimental Arm

	Age	Female	Non-White‡	Married	Income, $'000†	Share College†	Number Drugs	Charlson Score	Any EMR Use§	Intensity of EMR Use§~
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Information + Expert Recommendation Arm										
Responded to endline survey	-0.45	-0.06	-0.13*	0.09	-1.95	0.00	-0.22	0.09	0.04	2.09*
	(0.54)	(0.06)	(0.05)	(0.06)	(5.23)	(0.02)	(0.36)	(0.13)	(0.03)	(0.90)
No. of Obs.	410	410	410	410	410	410	410	410	410	410
Mean of Dep. Var.	72.13	0.49	0.20	0.62	112.73	0.60	4.64	1.08	0.95	7.43
Std. Dev. Of Dep. Var.	4.58	0.50	0.40	0.48	43.79	0.19	3.22	1.39	0.21	9.25
Panel B: Information Only Arm										
Responded to endline survey	0.01	0.06	-0.13*	0.06	7.08	0.04	0.10	-0.14	0.02	0.16
	(0.50)	(0.06)	(0.05)	(0.06)	(5.31)	(0.02)	(0.34)	(0.17)	(0.03)	(1.00)
No. of Obs.	391	391	391	391	391	391	391	391	391	391
Mean of Dep. Var.	72.11	0.49	0.23	0.61	110.45	0.58	4.46	0.98	0.96	7.06
Std. Dev. Of Dep. Var.	4.41	0.50	0.42	0.49	44.76	0.19	2.77	1.34	0.19	8.07
Panel C: Control Arm										
Responded to endline survey	-0.70	-0.00	-0.02	-0.07	4.82	0.06*	-0.38	0.20	0.04	-0.61
	(0.62)	(0.07)	(0.06)	(0.06)	(6.65)	(0.03)	(0.44)	(0.15)	(0.03)	(0.95)
No. of Obs.	384	384	384	384	384	384	384	384	384	384
Mean of Dep. Var.	72.81	0.53	0.23	0.57	114.02	0.59	4.46	0.96	0.95	6.15
Std. Dev. Of Dep. Var.	4.67	0.50	0.42	0.50	47.08	0.19	3.19	1.34	0.22	5.93

Table shows the relationship between baseline demographic characteristics of randomized individuals and their participation in the endline survey, defined as responding to at least one endline survey question by the pre-specified cutoff date. The relationship is estimated separately by experimental arm in Panels A, B, and C. Out of 928 individuals that responded to at least one question in the endline survey by the cutoff date, 316 were in arm "Information + Expert"; 299 were in arm "Information Only"; and 313 were in the control arm. In columns (1) through (10) we report the results of separate regressions of each baseline demographic characteristic as the dependent variable on the indicator variable for whether an individual responded to at least one endline survey question. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

‡ Non-white includes "other" and missing responses

† Computed at census tract level

§ Measured within 3 years prior to the intervention

~ Number of strands of electronic conversations

Table 5: Balance on Observables at Endline Survey

	Responded to endline survey	Age	Female	Non-White‡	Married	Income, $'000†	Share College†	Number Drugs	Charlson Score	Any EMR Use§	Intensity of EMR Use§~
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Information + Expert	-0.04	-0.65	-0.06	-0.05	0.09*	-2.62	-0.00	0.20	0.10	0.00	1.87**
	(0.03)	(0.37)	(0.04)	(0.03)	(0.04)	(3.57)	(0.01)	(0.26)	(0.11)	(0.02)	(0.63)
Information Only	-0.05	-0.57	-0.03	-0.03	0.07	-2.79	-0.01	0.10	-0.06	0.01	1.05
	(0.03)	(0.37)	(0.04)	(0.03)	(0.04)	(3.67)	(0.01)	(0.24)	(0.11)	(0.02)	(0.55)
Mean of Dep. Var. in Control	0.82	72.68	0.53	0.22	0.55	114.91	0.60	4.39	1.00	0.96	6.04
No. of Obs.	1185	928	928	928	928	928	928	928	928	928	928
Mean of Dep. Var.	0.78	72.28	0.50	0.20	0.61	113.12	0.59	4.49	1.02	0.96	7.01
Std. Dev. Of Dep. Var.	0.41	4.57	0.50	0.40	0.49	44.73	0.18	3.07	1.40	0.19	7.97
F-test, p-value	0.84	0.82	0.50	0.40	0.55	0.96	0.51	0.67	0.16	0.76	0.26

Table shows the relationship between the probability of responding to the endline survey (column 1) and baseline demographic characteristics (columns 2-11) of individuals who responded to at least one question on the endline survey and their experimental arm assignment. Individuals were randomized across three experimental arms. In colum (1) we report the results of a regression of an indicator variable for whether an individual responded to the endline survey on the indicator variables for experimental arms. In columns (2) through (11) we report the results of separate regressions of each baseline demographic characteristic as the dependent variable on the indicators for experimental arms, and a constant that captures the average value of the dependent variable in the control arm. We report the coefficients on the indicators for being randomized into treatment arms. The last row reports the F-test for the difference in the coefficients on the two treatment arm indicators. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

‡ Non-white includes "other" and missing responses

† Computed at census tract level

§ Measured within 3 years prior to the intervention

~ Number of strands of electronic conversations

Table 6: Intent-to-Treat Effect of Offering Algorithmic Decision Support

	Switched plans	Very satisfied w/ process	Decision conflict score	Search time > 1 hour	Change in expected OOP cost	Chose an "expert" plan
	(1)	(2)	(3)	(4)	(5)	(6)
Information + Expert	0.08*	0.08*	-0.14	0.08*	-94.27*	0.06
	(0.04)	(0.04)	(1.86)	(0.03)	(38.84)	(0.03)
Information Only	0.01	0.06	-1.46	0.06	-58.67	0.05
	(0.04)	(0.04)	(1.87)	(0.03)	(36.22)	(0.03)
Mean of Dep. Var. in Control	0.28	0.39	21.06	0.75	-111.55	0.39
No. of Obs.	896	928	883	918	880	898
Mean of Dep. Var.	0.31	0.44	20.51	0.80	-160.23	0.41
Std. Dev. Of Dep. Var.	0.46	0.50	22.22	0.40	462.67	0.49
F-test between arms (p-value)	0.10	0.60	0.48	0.58	0.34	0.83

Table shows the intent to treat estimates. Columns (1) through (6) report the results of separate regressions for six outcome variables as reported by participants in the endline survey. We report coefficients of a regression of the dependent variable as specified in the column headers on the indicator variables for whether an individual was assigned to one of the two treatment arms, as well as control variables. The dependent variables are defined as follows. Column (1) uses a variable that interacts the response to the question (in endline survey) of whether the consumer switched her plan with a variable that was constructed by comparing which plans individuals reported having in the baseline and endline surveys. Column (2) outcome is an indicator for whether the individual chose "very satisfied" on a 5-point scale satisfaction with the choice process question. Column (3) dependent variable is a decision conflict score constructed from underlying responses as described in the manuscript. Column (4) is a self-reported assessment of how much time the individual spent choosing a Medicare Part D Plan. Column (5) measures the savings in expected out of pocket costs between the plan that the individual had before the trial and the plan chosen after the intervention. This column restricts the regression to observations with cost changes within the 1st and 99th percentile of the distribution of cost change as this variable is highly skewed. Column (6) dependent variable is an indicator that take a value of one if the individual choose one of the plans with top 3 algorithmic expert scores in the endline survey. All regressions include the following controls: age, indicator for being female, non-white, married; median household income in census tract, percent of college graduates in census tract, count of prescription drugs in electronic medical records, Charlson score, indicator for using electronic medical records, number of message strands in electronic medical record system. In column 6 we in addition control for the baseline value of the outcome variable to reduce the noise. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

Table 7: Treatment-on-the-Treated Effect of Algorithmic Decision Support

	Used software	Switched plans	Very satisfied w/ process	Decision conflict score	Search time > 1 hour	Index: software use intensity†	Change in expected OOP cost	Chose an "expert" plan
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Information + Expert	0.81***	0.10*	0.10*	-0.18	0.10*	0.14*	-115.98*	0.07
	(0.02)	(0.05)	(0.05)	(2.27)	(0.04)	(0.07)	(47.06)	(0.04)
Information Only	0.80***	0.02	0.08	-1.82	0.08	-	-73.11	0.07
	(0.02)	(0.05)	(0.05)	(2.32)	(0.04)	0.00	(44.66)	(0.04)
Mean of Dep. Var. in Control	0.00	0.28	0.39	21.06	0.75	-	-111.55	0.39
No. of Obs.	928	896	928	883	918	497	880	898
Mean of Dep. Var.	0.54	0.31	0.44	20.51	0.80	0.08	-160.23	0.41
Std. Dev. Of Dep. Var.	0.50	0.46	0.50	22.22	0.40	0.79	462.67	0.49
F-test between arms (p-value)	0.74	0.10	0.62	0.47	0.59	-	0.35	0.84

Table shows the 2SLS estimates. Column (1) reports the first stage: difference in the probability of using the online tool by treatment arm assignment. By construction, individuals randomized into the control group had zero use of the software tool. The coefficients on the indicator variables for treatment arms thus measure compliance with assigned treatment. Columns (2) through (6) report the results of separate regressions for six outcome variables as reported by participants in the endline survey. We report coefficients of a regression of the dependent variable as specified in the column headers on the indicator variables for whether an individual was assigned to one of the two treatment arms, as well as control variables. The dependent variables are defined as follows. Column (2) uses a variable that interacts the response to the question (in endline survey) of whether the consumer switched her plan with a variable that was constructed by comparing which plans individuals reported having in the baseline and endline surveys. Column (3) outcome is an indicator for whether the individual chose "very satisfied" in a 5-point scale satisfaction with the choice process question. Column (4) dependent variable is a decision conflict score constructed from underlying responses as described in the manuscript. Column (5) is a self-reported assessment of how much time the individual spent choosing a Medicare Part D Plan. Column (6) is an index measure that combines the five outcomes: whether the consumer viewed explanaiton buttons within the software, how often these buttons were clicked, the total number of actions within the software, the number of actions per login, and the total time that the individual spent within the software tool. Column (7) measures the savings in expected out of pocket costs between the plan that the individual had before the trial and the plan chosen after the intervention. This column restricts the regression to observations with cost changes in between the 1st and 99th percentile of the cost change variables that is highly skewed. Column (8) dependent variable is an indicator that take a value of one if the individual choose one of the plans with top 3 algorithmic expert scores in the endline survey. All regressions include the following controls: age, indicator for being female, non-white, married; median household income in census tract, percent of college graduates in census tract, count of prescription drugs in electronic medical records, Charlson score, indicator for using electronic medical records, number of message strands in electronic medical record system. In column 6 we in addition control for the baseline value of the outcome variable to reduce the noise. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

† Comparison between "Information Only" and "Information + Expert," since the outcome is not defined for the control group that did not have access to the software

Table 8: Out-of-Sample Treatment Effect Heterogeneity - Plan Switching

Plan switch treatment effect quintile	Age (1)	Female (2)	Non-White‡ (3)	Married (4)	Income, $'000† (5)	Share College† (6)	Number Drugs (7)	Charlson Score (8)	Any EMR Use§ (9)	Intensity of EMR Use§~ (10)
Panel A: Information + Expert Recommendation Arm										
1	72.77	0.51	0.27	0.61	89.64	0.47	3.86	1.06	0.99	4.07
2	73.32	0.53	0.28	0.60	121.48	0.62	5.20	1.20	0.99	6.39
3	73.30	0.50	0.36	0.61	110.09	0.53	3.99	1.27	0.67	3.06
4	75.02	0.55	0.42	0.48	104.66	0.52	4.25	1.23	0.42	1.10
5	75.38	0.60	0.40	0.37	108.17	0.58	4.93	1.02	0.40	1.88
Panel B: Information Only Arm										
1	73.88	0.51	0.24	0.60	111.78	0.58	5.41	1.41	0.99	8.70
2	74.14	0.56	0.31	0.66	145.65	0.68	5.09	1.16	0.83	5.97
3	73.15	0.56	0.39	0.55	113.78	0.59	2.91	0.59	0.67	1.05
4	73.96	0.56	0.38	0.46	87.89	0.50	3.40	0.78	0.55	0.53
5	74.66	0.50	0.41	0.41	74.93	0.38	5.41	1.85	0.43	0.25

Table shows the mean of baseline demographic characteristics of the full sample of individuals that were invited to participate in the trial (29,451 individuals), by the quintile of their predicted indiviudal-level treatment effect (ITT; Arm Information + Expert in Panel A and Arm Information Only in Panel B) on the probability of switching plans. In columns (1) through (10) we report the within quintile average of each baseline demographic characteristic as recorded in column headers. The unit of observation is individuals.

‡ Non-white includes "other" and missing responses

† Computed at census tract level

§ Measured within 3 years prior to the intervention

~ Number of strands of electronic conversations

Table 9: Selection into Software Use Conditional on Trial Participation

	Switched plans	Very satisfied w/ process	Decision conflict score	Search time > 1 hour	Change in expected OOP cost	Chose an "expert" plan
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Lower bound of selection; OLS versus 2SLS						
OLS						
Information + Expert	0.17***	0.07	-1.68	0.10**	-158.12***	0.11***
	(0.04)	(0.04)	(1.81)	(0.03)	(39.16)	(0.03)
Information Only	0.09*	0.06	-3.34	0.08*	-91.72**	0.08*
	(0.04)	(0.04)	(1.84)	(0.03)	(35.08)	(0.03)
2SLS (Treatment on the Treated)						
Information + Expert	0.10*	0.10*	-0.18	0.10*	-115.98*	0.07
	(0.05)	(0.05)	(2.27)	(0.04)	(47.06)	(0.04)
Information Only	0.02	0.08	-1.82	0.08	-73.11	0.07
	(0.05)	(0.05)	(2.32)	(0.04)	(44.66)	(0.04)
Implied Magnitude of Selection						
Magnitude of Selection - Arm A	0.07	-0.03	-1.50	0.00	-42.14	0.04
Magnitude of Selection - Arm B	0.07	-0.02	-1.52	0.00	-18.61	0.01
No. of Obs.	896	928	883	918	880	898
Mean of Dep. Var.	0.31	0.44	20.51	0.80	-160.23	0.41
Std. Dev. Of Dep. Var.	0.46	0.50	22.22	0.40	462.67	0.49
Panel B: Upper bound of selection: Outcomes among those who take up treatment in control						
Logged-in into trial web page	0.21***	-0.014	-4.53	0.12	-168.7**	0.15**
	(0.05)	(0.09)	(4.80)	(0.08)	(64.20)	(0.06)
No. of Obs.	301	313	302	310	295	302
Mean of Dep. Var.	0.28	0.39	21.06	0.75	-111.55	0.39
Std. Dev. Of Dep. Var.	0.45	0.49	22.56	0.44	458.34	0.49

Table quantifies how much selection is present in the take-up of treatment. Panel A reports OLS estimates of the association between software use and outcomes. Software use is set to zero for the control group that is not given access to software. Columns (1) through (5) report the results of separate regressions for six outcome variables as reported by participants in the endline survey. We report coefficients of a regression of the dependent variable as specified in the column headers on the indicator variables for whether an individual used software as provided in each treatment arm, as well as control variables. The dependent variables are defined in the same way as in the main ITT and LATE result tables. We also repeat the results of 2SLS regressions to make the comparison convenient. The implied magnitude of selection in each arm is the difference between OLS and 2SLS coefficients. Panel B restricts the sample for individuals assigned to the control group. For these individuals, we report coefficients of a regression of the dependent variable as specified in the column headers and an indicator for whether an individual logged in the software page to receive the "control group" message that reminded individuals to choose their Part D plans, as well as control variables. All regressions include the following controls: age, indicator for being female, non-white, married; median household income in census tract, percent of college graduates in census tract, count of prescription drugs in electronic medical records, Charlson score, indicator for using electronic medical records, number of message strands in electronic medical record system. In column 6 we in addition control for the baseline value of the outcome variable to reduce the noise. The unit of observation is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

Table 10: Selection into Trial Participation and Predicted Treatment Effects

	Switched plans	Very satisfied w/ process	Decision conflict score	Search time > 1 hour	Change in expected OOP cost	Chose an "expert" plan
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Information + Expert Treatment Effects						
Not randomized	0.03***	0.00*	0.87***	-0.02***	-6.68**	0.01***
	(0.00)	(0.00)	(0.05)	(0.00)	(2.03)	(0.00)
Mean among randomized	0.09	0.05	0.82	0.08	-59.96	0.02
Std. dev. among randomized	0.05	0.04	1.61	0.07	68.85	0.06
No. of Obs.	29451	29451	29451	29451	29451	29451
Mean of Dep. Var.	0.11	0.05	1.66	0.07	-66.37	0.03
Std. Dev. Of Dep. Var.	0.05	0.03	1.59	0.07	63.86	0.05
Panel B: Information Only Treatment Effects						
Not randomized	0.04***	0.02***	0.08	-0.03***	-3.92	0.01***
	(0.00)	(0.00)	(0.08)	(0.00)	(2.33)	(0.00)
Mean among randomized	0.02	0.05	-1.36	0.07	-24.00	0.03
Std. dev. among randomized	0.06	0.07	2.83	0.07	78.64	0.04
No. of Obs.	29451	29451	29451	29451	29451	29451
Mean of Dep. Var.	0.06	0.06	-1.28	0.05	-27.76	0.05
Std. Dev. Of Dep. Var.	0.06	0.06	2.56	0.07	78.74	0.03

Table shows the difference in predicted treatment effects between individuals who responded to the invitation to participate in the experiment and those who did not. Columns (1) through (6) report the results of separate regressions where the left hand side variable is the individual-level prediction of the treatment effect form "Information + Expert" intervention (Panel A) or "Information Only" intervention (Panel B). We report coefficients on the indicator variable for whether an individual was in the randomized sample. 29,451 individuals were invited to participate. 1,185 entered the on-line enrollment portal, verified that they were eligible to participate, participated in a pre-enrollment survey and authenticated their identity. These individuals were randomized across three experimental arms. Individual-level treatment effects for each treatment arm are computed based on the generalized random forest algorithm (Wager and Athey 2018) as described in the text. The GRF algorithm was estimated using ten observables about individuals that are available in PAMF's administrative data and can hence be observed for the full starting sample of 29,451 individuals. The unit of observation in the regressions is individuals. Standard errors in parentheses are robust to heteroskedasticity. * p<0.05; ** p<0.01; *** p<0.001.

Table 11: Utility Model and Estimates of Noise in Beliefs

	OOP Cost	CMS Star Rating	AARP Brand	Humana Brand	Silverscript Brand
	(1)	(2)	(3)	(4)	(5)
Panel A - model estimates					
ψ (Control Arm)	-0.13	0.66	2.46	1.45	1.19
	(0.01)	(0.10)	(0.08)	(0.08)	(0.12)
Interaction: λ (Info Only Arm)	-0.08	0.90	0.53	0.70	-0.10
	(0.02)	(0.25)	(0.23)	(0.24)	(0.25)
Interaction: η (Info+Expert Arm)	-0.03	0.14	-0.38	0.36	-0.35
	(0.01)	(0.21)	(0.20)	(0.20)	(0.25)
Panel B - estimates of noise					
Panel B.1 - assume treatment corrects 100% of noise					
Utility weight under algorithmic treatment	-0.17	0.80	2.08	1.81	0.84
Noise in beliefs about utility weight, $1+\xi^{\beta}$	1.27	1.95	1.44	1.19	1.29
Noise‡ in beliefs about characteristic, $1+\xi^{\Phi}$	0.62	0.42	0.82	0.67	1.09
Panel B.2 - assume treatment corrects 80% of noise					
Utility weight under algorithmic treatment	-0.17	0.70	1.93	1.86	0.76
Noise‡ in beliefs about utility weight, $1+\xi^{\beta}$	1.37	2.56	1.62	1.25	1.39
Noise‡ in beliefs about characteristic, $1+\xi^{\Phi}$	0.57	0.37	0.79	0.62	1.12

Tables reports the estimates of empirical utility model and implied size of wedges in consumer's assessment of utility weights and product features. Panel A reports model. Each column corresponds to a plan feature included in the utility function. The model is restricted to plan features that consumers can observe on the first screen of experimental software. The model includes but we do not report a random coefficient on the OOP Cost parameter. Standard errors are reported in parentheses. Panel B translates coefficient estimates in Panel A into the estimates of the magnitude of noise wedges that can explain the differences in consumer choices across consumers that are exposed to treatment and consumers that are not exposed to treatment. Panel B.1 reports the estimates of wedges under the assumption that informational treatment completely eliminates the wedge in the perception of product features, and information + expert treatment completely eliminates the wedge in both the perception of product features and utility weights. In Panel B.2 we report the wedge estimates under the assumption that each treatment intervention eliminates only half of each wedge.

‡ Noise terms are assumed to be multiplicative relative to the underlying utility parameters, as in the following: $u_{ij} = \left(1 + \xi^{\beta}\right)\beta_i(1 + \xi^{\varphi})\varphi_{ij}$.

A noise in beliefs about utility weights > 1, suggests that consumers put too much weight on the characteristic. A noise in beliefs about characteristics <1, suggests that consumers have a downward biased beliefs about the level of the characteristic or the probability that a particular product has a certain characteristic.

Table 12: Normative Implications of Noise in Beliefs

	Mean	5th percentile‡	25th percentile	50th percentile	75th percentile	95th percentile
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A - welfare loss (L), in $/year‡‡					
Allow for $1+\xi^{\beta}$	48.0	0	0	0	57.6	237.6
as % of U_{ij*}	*4.1*	*0*	*0*	*0*	*2.4*	*10.7*
Allow for $1+\xi^{\phi}$	68.1	0	0	0	92.4	259.0
as % of U_{ij*}	*4.8*	*0*	*0*	*0*	*3.9*	*11.6*
Allow for both $(1+\xi^{\beta})$ and $(1+\xi^{\phi})$	65.4	0	0	0	100.6	296.6
as % of U_{ij*}	*6.8*	*0*	*0*	*0*	*4.3*	*15.0*
	Panel B - probability of trial take-up					
Probability of trial participation	0.040	0.041	0.041	0.041	0.038	0.039

The table reports the outcomes of utility model simulations on the sample of all 29,451 individuals that were originally invited to participate in the trial. For each individual we compute the level of the utility function for each plan under four scenarios: (1) using "true" utility as implied by the estimates of utility parameters under algorithmic treatment; (2) allowing for the noise in beliefs about utility weights as estimated in the model; (3) allowing for the noise in the beliefs about product characteristics as estimated in the model; (4) allowing for both sources of noise. To compute utility, for each individual we draw one random draw of a random coefficient and add the term that captures unobserved part of utility (ε_{ij}) computed as an average of 100 random draws from Type II extreme value distribution for each individual. Each utility simulation generates a ranking of insurance plans. In Panel A, we report, for simulations 2, 3, and 4, how much consumers loose in "true" utility (as measured in simulation 1) when they choose a plan guided by plan ranking generated in simulations 2-4. Utility loss is reported in dollars. The dollar value is obtained by diving the utility value by the absolute value of the coefficient on the out of pocket cost as estimated for the "true" utility model. For each dollar-value of welfare loss, we also report the relative loss, as a percent of utility in simulation 1. For each simulation, we report the average loss or percent loss across the whole population (column 1), as well as the quintiles of the loss distribution (columns 2-6). Pancel B reports the rate of trial participation for the whole sample (column 1), as well as within each moment of the loss distribution as specified in columns 2-6 from simulation #4 that allows for both wedges in beliefs.

‡ Percentiles computed across 29, 451 individuals that were invited to participate in the trial
‡‡ See equation (8) in the manuscript for the definition of the welfare loss function