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, Bobby Pakzad-Hurson, Stephen Ryan, Frank Schilbach, Jesse Shapiro, 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, Brown University, University of California Berkeley, AHEC 2019, and 2019 Conference on Health IT and Analytics 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 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, Revised June 2020
JEL No. D1,D12,D8,D81,D82,D83,D9,D90,D91,G22,H51,I13

ABSTRACT

Algorithms are increasingly available to help consumers make purchasing decisions. How does algorithmic advice affect human decisions and what types of consumers are likely to use such advice? We conducted a randomized, controlled trial comparing the effects of offering personalized information, either with or without algorithmic expert recommendations, relative to offering no personalized information for consumers choosing prescription drug insurance plans. Treated consumers were more likely to switch plans and to choose a plan that lowered their total spending on drugs. The behavioral response was more pronounced when information was combined with an algorithmic expert recommendation. We develop an empirical model of consumer choice to examine the mechanisms by which expert recommendations affect choices. Our experimental data are consistent with a model in which consumers have noisy beliefs not only about product features, but also about the parameters of their utility function. Expert advice, in turn, changes how consumers value product features by changing their beliefs about their utility function parameters. We further document substantial selection into who demands expert advice. Consumers who we predict would have responded more to algorithmic advice were less likely to demand it.

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 Hea
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 well-being. These types of decisions affect households at all points in the income distribution and include products such as payday loans, mortgages, mobile phone plans, credit cards, life and health insurance, and investment vehicles. Moreover, participation in publicly subsidized benefits, from health insurance to tax-favored retirement arrangements, has evolved in ways that increasingly require relatively sophisticated financial decision making.

The emergence of large-scale data over the past decade and the corresponding development of statistical techniques to analyze these data have the potential to significantly change how consumers make decisions in these environments (Einav and Levin, 2014). Algorithms, which can serve as either substitutes for or complements to human decision-making, provide an opportunity to dramatically scale access to expert recommendations at a fraction of the cost of human assistance (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 surprisingly little empirical evidence on the extent to which and how algorithms influence decision making and what types of consumers use algorithmic assistance.

Our study makes three contributions that begin to close that gap. First, we provide empirical evidence quantifying how much algorithms affect consumer decision making in a real world context. Our study is based on data from a randomized, controlled trial of decision-support software that provided algorithm-based "expert recommendations" to older adults choosing insurance plans for their prescription drugs. In contrast to most studies of the effects of informational interventions, our experimental results demonstrate - in a non-laboratory setting - that consumers are responsive to algorithmic recommendations when making decisions. Not only do people change their choices of insurance plans in response to our treatment, but the response is more pronounced when personalized information is combined with advice from an algorithmic expert.

Second, we provide a simple theoretical framework for understanding how algorithmic advice influences consumer behavior. We propose that recommendations can affect consumers in two conceptually distinct ways: by changing consumer beliefs about product features ("learning"), or by changing consumer beliefs about the mapping of product features into utility ("interpretation"). The design of our trial, motivated by this conceptual distinction, allows us to separately observe how consumers respond to algorithmic expert advice when advice only includes messages about product features versus when advice also implicitly includes information about how to value product features. We find that consumers exhibit differential responses to these two types of interventions. While this evidence allows us to reject the notion of perfectly informed consumers, the results also indicate that consumers have noisy beliefs about *both* product features and how they value those features. In other words, expert recommendations generate both learning and interpretation. The implication is that algorithms create a powerful tool for changing consumers' beliefs about their own preferences.

Finally, we highlight the importance of selection in demand for (algorithmic) expert advice. We document substantial selection into the use of our on-line tool along two margins. First, people who are "active shoppers"

- those who were already considering switching their plan - were more likely to use the decision support tool, conditional on signing up for the experiment. This evidence is consistent with selection on "levels" — consumers with higher levels of potential outcomes even in the absence of treatment were more likely to select into treatment. Second, people for whom we predict the largest treatment effects, using novel machine-learning methods for estimating heterogeneous treatment effects (Wager and Athey, 2018; Athey and Wager, 2019), were the least likely to sign up for the experiment. This is evidence consistent with selection on "slopes" — consumers with the largest predicted treatment effects were the least likely to exhibit demand for expert advice. As we find that treatment effects generally decline with several measures of socio-economic status, the selection on slopes suggests that self-targeting of consumers to expert advice may have undesirable distributional properties.

The empirical context of our study is Medicare Part D, publicly subsidized prescription drug insurance for aged and disabled adults in the US. The program has high participation rates - insuring over 43 million older adults and accounting for over $88 billion in annual public spending (Kaiser Family Foundation, 2018). Under Part D, private insurance plans compete for subsidized enrollees. In this project, we focus on the so-called stand-alone prescription drug plans (PDPs) that offer only prescription drug coverage and no additional medical benefits. Each year during a pre-specified open enrollment period, older adults covered by Medicare may choose from among the approximately 25 stand-alone insurance plans offered in their geographic area (Kaiser Family Foundation, 2018). Plans vary in their premiums and cost sharing, in which drugs are covered, and in which insurer administers the plan.

Our study builds on a large body of literature studying health insurance choices more generally and Medicare Part D specifically that suggests that algorithmic assistance could be valuable to consumers in this context (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). 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; Brown and Jeon, 2019). 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 misconceptions 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), or consider only a subset of plans (Abaluck and Adams, 2017; Coughlin, 2019). Other studies draw stronger, normative conclusions about consumer decision making (Abaluck and Gruber, 2011; Heiss et al., 2010; Heiss et al., 2013, 2016).[1] Ketcham et al. (2015), however, find that consumer decision-making improves over time, suggesting

[1]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.

that choice inconsistencies may be short-lived.

Our randomized field trial ran during the 2017 Medicare open enrollment period (November-December 2016). Participants were randomized to either one of two treatment arms or a 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, people received access to software that provided a list of all available plans with a personalized cost estimate and information about other plan features. The plans were ordered by a one-dimensional, algorithmic "expert" score, but the score itself was not displayed. In the "Information + Expert" arm, people had access to an identical tool with the exception that the algorithmic expert score for each plan was displayed and the three plans with the highest personalized score were marked as "Plans recommended for you." The algorithmic expert score was the weighted average of the other features that were shown to consumers and hence did not contain any additional information about product features. We have briefly reported the pre-registered baseline results of this experiment in a companion policy paper (Bundorf et al., 2019).

In this paper, we focus on the economics of how algorithmic advice affects human decision making. We report three main findings. First, we document that providing advice through expert algorithms affects consumer decision making in this context. While consumers' plan choices change when consumers are exposed only to information about product features, their response is somewhat more pronounced when consumers are also exposed to the expert recommendation. For our main measure of behavioral response - 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.

This finding builds on recent studies that have examined the role of personalized information (but not expert algorithms), on its own or relative to in-person advice, in several different contexts, including college funding, the SNAP program, and health insurance (Bettinger et al., 2012; Finkelstein and Notowidigdo, 2019; Kling et al., 2012). Kling et al. (2012), also in the context of Medicare Part D, emphasize the importance of "comparison frictions," by documenting that simply mailing consumers letters with the same (personalized) information that is available from a publicly available website changes consumer behavior. Our findings contribute to this body of work by examining how consumers respond not just to personalized information but to personalized information when it is provided through machine-based expert algorithmic advice.

Our second key finding is that expert algorithms influenced both consumers' information about product features and how they valued those features. This finding is based on a theoretical model of consumers choice behavior in which we propose two mechanisms by which expert advice can influence consumer decision making: by changing consumers' beliefs about product features and the relative importance of features in the utility function. The design of two treatment arms in our experiment allows us to decompose these different channels empirically. In the "Information Only" arm, consumers received truthful messages from the expert about the features of insurance products. In the "Information + Expert" arm, the expert sent messages with information about product features, but also with the expert's judgement on how a consumer should value different product

features. By testing whether consumers differentially respond to these interventions, we can determine whether consumers have noisy beliefs about product features, utility function parameters, or both. We develop and estimate an empirical model of consumer demand for insurance plans. We follow the strand of literature in insurance that models the valuation of contract features directly (e.g. Bundorf et al., 2012; Decarolis et al., forthcoming; Starc and Town, 2019) rather than through a model of the underlying utility function with risk aversion (e.g. Cardon and Hendel, 2001; Cohen and Einav, 2007; Barseghyan et al., 2011; Handel, 2013; Barseghyan et al., 2013).[2] This is a conscious choice in our context. Given that the differences in risk protection across Part D plans are, in practice, quite limited, we pursue a less parametric specification in which the utility weights capture the reduced-form of multiple underlying structural behavioral parameters.[3] The model allows us to reject the notion that consumers have perfect information about product characteristics. Instead, our estimates are consistent with consumers having noisy beliefs about *both* product features and utility function parameters. We further estimate that this noise in beliefs leads about a quarter of consumers in this market to leave substantial financial gains on the table.

Our results raise two fundamental concerns about the idea of scaling up expert advice through algorithms. First, we observe that very simple algorithmic advice manages to move some consumers to more homogeneous behavior by communicating a set of underlying "expert" preferences to consumers. While this type of advice significantly simplifies the dimensionality of the choice problem, it may change not only consumers' information set, but also their preferences. In reality, however, consumer preferences are likely to be heterogeneous. Moreover, any algorithm that incorporates expert preferences into its recommendation, will ultimately have to make a judgement about what constitutes the right preferences. This type of concern extends to policy interventions beyond expert recommendations. For example, Ericson and Starc (2016) find that consumer choices and inferred utility weights change when health insurance products become standardized. In theory, both of these limitations can be overcome by creating individualized weights that are learned from consumer choices using "consumers like you" approaches. This, however, raises a second concern. "Consumers like you" approaches are bound to reinforce any noise, mistakes and behavioral biases that may exist in consumer choices, defeating the purpose of providing algorithm-based expert advice.

Finally, we find strong evidence for both that the degree of selection in demand for expertise is large and that demand for expertise is complementary with existing consumer knowledge and sophistication. Among those participating in the trial, the magnitude of selection into software use on outcomes is comparable to the magnitude of the treatment effect - those who chose to take up the software were inherently at least 7 percentage points more likely to switch plans even if they had not used the tool. This evidence of positive selection on outcomes points 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

[2] See Einav et al. (2010) for a detailed discussion of the dichotomy between the modeling of realized utility with risk aversion versus directly modeling the valuation of insurance contracts.

[3] Risk protection variation across Part D plans is to the first-order driven by the differences in the generosity of the formularies rather than by the differences in plans' financial characteristics, making this a more complex setting than in the contexts where insurance plans only vary in their financial characteristics. As all Part D plans cover most common drugs and since formularies are not transparent, it is not obvious that consumers have meaningful preferences and priors over the coverage of drugs that they are not taking or do not expect to be taking.

generalized random forest algorithm (Athey et al., 2019) and the administrative data on all individuals that were invited to participate in the trial, we are also able to examine selection on treatment effects. 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. In other words, linking this back to our conceptual framework, we find that people for whom information and algorithmic recommendations were most likely to influence their choices, i.e. consumers with the highest degree of noise in their beliefs, were the least likely to seek out expertise. 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 this support. Instead, merely offering access to expertise may create unintended distributional concerns, as consumers with more knowledge are likely to benefit more than those with less knowledge. Hence, policies with more targeted and intensive interventions may be required to reach consumers who could benefit from expert recommendations.

Overall, our study contributes to the broader literature emerging across different scientific disciplines that is trying to understand how humans will interact with artificial intelligence in various contexts. Advising consumers on complex financial decisions appears to be one of the most natural applications of these methods. We emphasize that this application, however, can pose an important conceptual challenge - it requires algorithm makers to take significant stance on what the right choices ought to be. This may be problematic in environments in which consumers exhibit highly heterogeneous preferences. Our findings also raise concerns over using "consumers like you" approaches in settings in which consumer decisions in the absence of informational interventions are inconsistent with rationality, as has been extensively documented in many different household finance domains Beshears et al. (forthcoming).

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

2 Experimental Design and Data

2.1 Study Setting

Medicare is the public health insurance program in the U.S. primarily for people over age 65, covering 50 million people (Centers for Medicare & Medicaid Services, 2019). Prescription drugs for Medicare beneficiaries are covered by Medicare Part D. Part D coverage is administered by private insurance plans (Duggan et al., 2008). Enrolling in Medicare Part D is voluntary for beneficiaries and requires an active enrollment decision in the form of choosing among the private prescription drug plans offered in the beneficiary's market and paying a premium. Medicare Part D benefits can either be sold as stand-alone drug coverage or bundled with medical

benefits. In this project we focus on stand-alone prescription drug plans (PDPs).

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 financial coverage they offer is actuarially equivalent or exceeds the statutory minimum. Plans are also differentiated along dimensions such as the composition of pharmacy networks, the availability of mail order, formulary design and customer service. To capture these non-financial features of plans, 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.[4]

Our study focuses on beneficiaries already 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). The plans varied in their financial features: for example, deductibles ranged from $0 to $400; the plans covered an average of 3,291 drugs (s.d. 257). The average CMS rating of plan quality in California was 3.4 out of 5 stars (s.d. 0.6).

2.2 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, a large multi-specialty physician group in California. 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 individually customized, objective information about the plan's financial features and one with an algorithmic expert recommendation in addition to the financial information, 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 arms and in the control arm.

The two "treatment" versions of the tool were identical with the exception of whether the user interface included an individualized, one-dimensional expert score. In both versions, when consumers logged in, they saw a list of their current prescription drugs that was automatically imported from their electronic medical records 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 treatment arms, the plan list included the name of

[4]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.

the plan (including the insurer's brand), the individual's total estimated spending in each plan based on their drug list, and the CMS star rating for each plan. The total estimated spending included the plan premium and the personalized out-of-pocket spending that was estimated in the background based on information about drug-level coverage rules and pricing that Medicare Part D plans annually report to CMS. This background 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 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 a one-dimensional, individualized expert score. The expert score, based on a 100-point scale, was a combination of the estimated individual-specific total cost of purchasing the plan and the plan's "star rating". 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.

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, they only saw the two underlying plan features - total cost and the star rating. 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 well as the total cost and the star rating). 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.[5]

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 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.

[5]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.

2.3 Study Population

We recruited trial participants from patients who receive care at the Palo Alto Medical Foundation (PAMF). A collaboration with PAMF allowed us to access the electronic medical records of these patients, including 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. In other words, the subjects received no information on their assigned study arm until they chose to access 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 regardless of whether they actually entered the study website, the final survey includes both those who entered the study website (i.e. were exposed to the intervention) 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 old (s.d. of 5 years), 54 percent were female, 35 percent were non-white,[6] and 54 percent were married. We matched each individual to their census tract (roughly equal to a neighborhood, comprised of 2,500 to 8,000 people) to get their socio-economic characteristics. We used the median (in a tract) household income and percent of individuals with a college degree. The resulting average 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 keep in mind when interpreting our results and considering their external validity. The high average income of our participants makes them unrepresentative of the broader population of older Americans; however, this

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

sample provides us with the opportunity to test whether offering algorithmic expert advice - in one of the wealthiest and technologically most attuned areas of the country - affects individuals' behavior. As we discuss in Section 5, we find that high SES and demand for expert advice are complements. Hence, our results on this high SES population provide an upper bound for the likely take-up of these types of tools 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. We report 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.

2.4 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 the 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. 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.

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 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 the arms above and beyond the differences observed across the 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.

2.5 Outcomes

We consider six outcomes across different domains in our baseline specifications, four of which we pre-specified as primary outcomes and two of 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.[7] 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. 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.

3 Experimental Results

3.1 Effect of Offering Expertise

We start by estimating the aggregate behavioral response to being offered the decision support tool, the intent-to-treat (ITT) effects. 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:

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

[7] 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. X_i is a vector of individual observable characteristics that we introduced in Sections 2.3 and 2.4.

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.[8] This is consistent with both a relatively high switching rate in the control group, as well

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

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.[9] 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.

3.2 Effect of Exposure to Expertise

We next proceed to estimate the average causal effect of using the decision support software among those who used it. As there are no always-takers in our setting, using the experimental assignment as an instrument recovers the treatment on the treated effect. 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.

Table 7 reports the first stage coefficients and the 2SLS estimates for the six outcome variables of interest.

percentile of changes.

[9] 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

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 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 ITT 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 treated individuals, 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.

The first stage and the ITT estimates allow us to report our results in terms of the "persuasion rate" (DellaVigna and Kaplan, 2007), which measures how effective messages are at changing consumer behavior. Following DellaVigna and Gentzkow (2010), we compute the persuasion rate as:

$$f = 100 * \frac{\alpha_1}{\pi_{11}} \frac{1}{1 - \gamma_0} \tag{5}$$

The persuasion rate allows us to compare the effects of intervention to those of other types of interventions that are intended to influence consumer behavior. The persuasion rate is around 13% for the main outcomes related to plan choice (13.9% for plan switching and 12.3% for choosing an expert-recommended plan). These are quite large relative to the rates of 1 to 7% that are observed in two studies focused on consumer persuasion as reported in the overview work by DellaVigna and Gentzkow (2010). These rates are also quite high in the context of all different "persuasion" studies. The upper bound of the persuasion rate was 29.7% in DellaVigna and Gentzkow (2010). We observe an even higher persuasion rate on the outcome that measures whether consumers spent more than 1 hour engaging in their choice of plan, which has the persuasion rate of 40%.

Overall, we conclude that being exposed to the algorithmic recommendation increased the propensity of consumers to shop for plans and decreased their costs, and that this exposure had a relatively strong effect relative to other types of interventions intended to influence consumer behavior through persuasion, or more generally non-price mechanisms. Being exposed to individualized information had effects in the same qualitative direction but quantitatively, the effects on switching and costs were less pronounced. Using 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.

4 Mechanisms

4.1 Model

In this section we develop a stylized model of how advice impacts decision makers to analyze the mechanisms underlying the behavioral changes that we observe in Section 3. We follow the rich literature on the economics of communication between a less informed principal and a better informed expert who gives advice and has an interest in persuading the principal to follow that advice (Ottaviani, 2000 and DellaVigna and Gentzkow, 2010 provide detailed overviews of this literature). In our setting, the expert has a paternalistic rather than a strategic objective, aiming to convince the consumer to make the "right" choice according to a set of preferences

deemed desirable by the expert.[10] This conceptual model motivates our experimental design. In this section, we formalize the model to highlight the assumptions that we need to identify empirically how expert advice—in this case delivered through an algorithm—affects consumer actions, and why it may affect consumers differently than pure provision of information about plan features.

Set-up Consider consumer i who faces a choice set J of products. Each product j is characterized by a (possibly individual-specific) vector of features and product prices, ϕ_{ij}. Let $U_{ij}(\phi_{ij};\beta_i)$ be the utility that consumer i experiences from product j with features ϕ_{ij}. Here, β_i are parameters of consumer's i utility function and determine consumer "type." Consumers of different types—for example, of different risk aversion, health, or brand loyalty—may experience different levels of utility from the same vector of product features ϕ_{ij}. A consumer who knows both her own type (β_i) and the actions chosen by insurance companies, in the form of product features ϕ_{ij}, chooses product j^* such that U_{ij^*} is greater than U_{ij} for all other $j \in J$.

In practice, consumers may not have full information about either product features or their own types. Such partitioning of choice uncertainty into uncertainty about ϕ_{ij} and β_i maps into the distinction between games of imperfect and incomplete information (Harsanyi, 1967). Being uncertain about ϕ_{ij} is equivalent to having imperfect information about the actions chosen by other players in the game—in our application, insurers defining plan features. Consumers may, however, also have incomplete information and be uncertain about the parameters of their own payoff functions.[11]

Denote consumer beliefs about vectors ϕ_{ij} and β_i with $\tilde{\phi_{ij}}$ and $\tilde{\beta}_i$. Consumer i maximizes her perceived utility given these beliefs and chooses a product $\tilde{j}$ such that $\tilde{j} = \underset{j}{\text{argmax}}\, U_{ij}(\tilde{\phi_{ij}};\tilde{\beta}_i)$. The welfare loss L is given by the difference in the experience utility from plan j^* relative to plan $\tilde{j}$, i.e. $L = U_{i\tilde{j}} - U_{ij^*}$. The welfare loss is zero when consumer beliefs are such that consumer i chooses the optimal plan j^* even when she is uninformed.

Exposure to Expertise An uninformed consumer may want to consult an expert to help her choose the right product. The expert sends messages to the consumer and the consumer then evaluates her alternatives in J considering both these messages and her own priors. The expert has accurate information about plan

[10] The idea of a paternalistic objective in a communication game is related to the model in Lightle (2014), except in the latter paternalistic objectives lead the sender to send over-stated messages, while we assume that a paternalistic algorithm sends sincere messages, which is the case in our empirical setting. This idea is also related to the broader literature on paternalistic policy-making with non-optimizing households, see, for example, Fadlon and Laibson (2017) and Camerer et al. (2003). Tsai (2014) provides an in-depth discussion of the normative and philosophical considerations for thinking about rational persuasion and paternalism.

[11] Uncertainty about parameters of the utility function, or more specifically, utility weights on product features is one way to capture the idea that consumers, for example, may not understand the full implications of complex provisions in financial contracts, or involved cost-sharing constructs in insurance plans, even if they have perfect information about the existence of these features (Bhargava et al., 2017). Distinguishing between the two sources of uncertainty implies that there are two types of information a consumer may want to 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 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, an extensive literature in 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 when thinking about how advice may affect consumers.

features ϕ. The expert also has information about β; however, this information is more subjective as it reflects the expert's judgement on the level of β that the expert believes is right for the consumer. The expert has no private information about a specific consumer's β_i, for example, her risk aversion or brand preferences.

The expert can send two types of messages. One type only includes objective information about the state of the world $m_\phi = \phi$. The other includes the same information about ϕ and as well as the expert's judgement about the parameters of the utility function, m_β — m_β may or may not equal β_i. In our setting we assume that the expert's only objective is to persuade the consumer to choose the right product for herself (as viewed by the expert). There are no other strategic objectives in the expert's problem. As a result, the expert sends truthful messages that fully reflect the information about the state of the world and the expert's truthful opinion about the correct product choice. When faced with the message about both ϕ and β, the consumer decides whether to follow the expert's advice (choose j^{expert} that maximizes $U_{ij}(m_\phi, m_\beta)$ or to update her beliefs, but not fully adopt the expert's choice.

The distinction between uncertainty that consumers may have about ϕ_{ij} versus β_i creates sharp predictions about consumer responses to different types of messages that an expert may send. For example, if consumers have perfect information about product features, but are uncertain about how to map those feature into utility (i.e. uncertain about their own types), then messages of the form $m_\phi = \phi$ should have no impact on consumer behavior. Reversely, if consumers have uncertainty only about ϕ_{ij}, and not about β_i, then advising consumers on how to interpret information about ϕ_{ij} should have no additional effects on consumer decision-making after information about ϕ, m_ϕ, has been provided.[12] It follows that if we were able to observe consumer behavior when the consumer is exposed to an "information" message m_ϕ only versus an "advice" message (m_ϕ, m_β), we could test whether consumers have uncertainty about only one or both parts of the partition that we proposed here. Our experimental design allows us to do just that. We experimentally vary whether consumers are exposed to no expert messages, to messages that contain objective information about ϕ only, or information about ϕ combined with a recommendation for the best plan, which is implicitly sending a joint message (m_ϕ, m_β). This allows us to shed light on the mechanism by which advice can affect consumers, as we can separately test for the existence of uncertainty in consumers' beliefs about product features ϕ versus parameters of the utility function, β.

The idea that expert advice differs from pure information in that it can alter consumer beliefs about the parameters of the utility function (β) as well as product features (ϕ) provides a useful way to think about the potential impact of increasingly popular algorithmic, or more generally AI-based, decision support systems that are aiming to democratize and dramatically scale access to expertise. Our framework implies that interventions aimed at helping consumers make choices can change their choices through two mechanisms: by changing their beliefs about the features of the products, and by changing their utility weights for these features. These two

[12] Our framework provides insight into how to interpret studies of consumer choice in the setting we study - Medicare Part D prescription drug plans. Abaluck and Gruber (2011), for instance, argue that consumers exhibit choices that are inconsistent with rationality, putting more weight on premiums than on out of pocket expenditures in their utility function. This is equivalent to arguing that $\tilde{\beta}_i \neq \beta_i$, but $\tilde{\phi_{ij}}$ is equal to ϕ_{ij}. Related evidence in Kling et al. (2012), however, rejects the notion that $\tilde{\phi_{ij}}$ is equal to ϕ_{ij}, since providing information about ϕ_{ij} to consumers changes their choice behavior. Taken together, these results indirectly suggest that both sources of uncertainty likely exist in this setting, but we have no direct evidence on their joint existence and relative importance.

mechanisms generate very different implications for policies intended to change behavior through information provision. If consumer choices are driven by noisy priors about how product features map into utility, then a policy of providing information about 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 make information more accessible may be effective. This distinction is of central practical relevance in the markets for complex financial products, where the knowledge of product features may not be enough for consumers to make informed decisions, and where algorithms that purport to merely simplify consumers' choice by aggregating complex information into a uni-dimensional decision metric may end up altering the parameters of consumers' utility functions.

4.2 Estimation

Estimation approach We now map our experimental data into an empirical version of the conceptual model outlined above. We start with a random utility 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 in general we empirically specify as follows:

$$u_{ijt} = \beta_i \phi_{ijt} + \epsilon_{ij}, \;\; \epsilon_{ijt} \sim \text{iid EV Type I} \tag{6}$$

Here, ϕ_{ij} is a vector of characteristics of product j that are allowed to be individual-specific. The vector β_i are parameters of the utility function that map 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" dollars into utils. This marginal utility of income can vary across individuals. When normalized to the marginal utility of income, other entries in vector β_i, provide a measure of individual's willingness to pay for each product feature. ϵ_{ij} captures any consumer-product specific parts of utility that affect consumer choice, but are not observed by the econometrician. Allowing for consumer uncertainty, the decision utility becomes: $u_{ijt} = \tilde{\beta}_i \tilde{\phi}_i + \epsilon_{ij}, \;\; \epsilon_{ijt} \sim$ iid EV Type I.

Is it possible to separately identify uncertainty in β_i and ϕ_{ij} in the data? Conceptually, we would need interventions that separately affect only $\tilde{\beta}_i$ or $\tilde{\phi_{ij}}$. We argue that our two treatment arms provide us with exactly that type of variation. Arm "Information Only" provides individuals with personalized information about expected costs and CMS plan quality rating. Individuals receive information about plan features, but they do not receive any further guidance about how much different plan features should matter for their utility function. In other words, for individuals enrolled in the "Information Only" arm, the treatment affects only $\tilde{\phi_{ij}}$. Individuals in the "Information + Expert" arm receive the same information as those in "Information Only" arm, but they also get 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 implicitly sends consumers a message with the expert's opinion on the appropriate relative weighting of various plan features.

It follows that by comparing the choice behavior across three experimental arms, we can separately test for

the existence of two channels of uncertainty in consumers' beliefs and quantify the extent to which algorithms affect these beliefs. The key identifying assumption is that there are no latent differences in utility weights across the three experimental arms, so that we can attribute differences in estimated willingness to pay parameters across the arms to differences in beliefs about ϕ_{ij} and βi. We verify this assumption empirically by testing for differences in revealed preference parameters at the baseline, prior to the intervention. We find no differences in estimated willingness to pay for product features across experimental arms.

Under this assumption, the model delivers two predictions. First, we can reject the hypothesis that consumers have no uncertainty about product features ϕ if the revealed preference estimate of the willingness to pay for ϕ differs between the control arm and the "Information Only" arm. Second, we can reject the hypothesis that consumers have no uncertainty about the parameters of their own payoff function if willingness to pay for ϕ differs between the "Information Only" and "Information+Expert" arms. The latter implies that consumers are uncertain about how to evaluate product features in this context. More substantively, it also implies that algorithmic recommendations can influence consumer preferences.

In practice, we pool the data from all experimental arms, both pre- and post- intervention, and estimate the following specification. For consumer i in year t, consumer's utility is:

$$u_{ijt} = \mu_1 Cost_{ijt} + \mu_2 CMSStar_{jt} + \mu_3 AARP_{jt} + \mu_4 Humana_{jt} + \mu_5 Silverscript_{jt} + \epsilon_{ijt} \tag{7}$$

$$\mu_n = \psi_n + \lambda_n I + \eta_n E \quad \forall n \in [1,5] \tag{8}$$

In this specification, ϕ_{ij} includes the individual-specific expected cost of enrolling into a plan ($Cost_{ijt}$), CMS star ratings, and indicators for 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). Participants in the control arm can gather this information from publicly available sources, so their utility function over these objects is still well-defined. μ_n is a vector of coefficients that map each product feature into utility. μ_1, which multiplies the dollar-denominated cost of enrolling into a plan, captures the value of a dollar in utils. The ratio of $\mu_n, n > 1$ to μ_1 gives us an estimate of the consumer's willingness to pay for each product feature. As can be seen in Equation 8, we allow the revealed preference estimate of the willingness to pay parameters to vary between the control and treatment arms, where I denotes "Information Only" arm and E denotes "Information + Expert" arm.[13] ψ_n captures the willingness to pay as revealed in the control group. λ_n and η_n then capture how the revealed willingness to pay for the same product feature differs between the control group and "Information Only" as well as "Information + Expert" group, respectively. We further allow for unobserved heterogeneity in consumer preferences for expected costs, which we implement by assuming an individual specific ψ_{1i} that is distributed normal; the mean and variance of this normal distribution will be in

[13]Strictly speaking, the differences in the estimates of β and ϕ between the control and treatment groups will capture not only the differences in information signals about product features and preferences, but also any differences in consideration sets. While treated consumers were exposed to a very salient (and observable) list of plans, making it reasonable to assume that all plans were part of the consideration set, we do not observe which plans were considered by consumers in the control group. If, in practice, the consideration set of the control group was substantially smaller than that of the treated consumers, our estimates of λ_n and η_n (relative to the control group, not relative to each other) will also reflect these differences(Abaluck and Adams, 2017; Caplin et al., 2018; Barseghyan et al., 2019; Coughlin, 2019).

the set of parameters that we estimate.

λ_n and η_n are the main coefficients of interest that directly map to the predictions of the model:

1. We can reject that consumers have no uncertainty about product features ϕ if $\lambda_n \neq 0$.

2. We can reject that consumers have no uncertainty about the parameters of their payoff function if $\lambda_n \neq \eta_n$. Further, $\lambda_n \neq \eta_n$ provides evidence that consumers are uncertain about how to interpret product features and that algorithmic recommendations influence their preferences.

Estimation results Panel A of Table 8 reports model estimates. Column (1) reports ψ_1, λ_1, and η_1 - coefficients on the individual-specific expected cost of purchasing a plan and its interaction with experimental arm; Column (2) reports the analogous coefficients on CMS star rating; and Columns (3)-(5) on three most popular insurer brands. In Panel B, we convert the interacted coefficients into the implied utility weights for each experimental arm by adding the coefficients for product features in the control arm to the coefficients on the interaction term in each treatment arm. In Panel C, we convert the coefficient estimates into willingness to pay for product features in each treatment arm.

We first consider the estimates of ψ_1 to ψ_5 which represent decision utility in the control arm. We estimate a negative coefficient (-0.13) on cost as expected since we anticipate that consumers prefer lower cost plans. This coefficient also provides an estimate of the marginal utility of income. Our estimate of the standard deviation of the normal distribution of this parameter is 0.16, indicating that the marginal utility of income varies substantially across consumers. The coefficients on the CMS star rating and insurance company brands are economically large and statistically precise. Consumers place significant value on brands in this market despite the fact that insurance products are relatively standardized and tightly regulated.

Rows two and three in Panel A report the main coefficients of interest. The estimates of λ_1 to λ_5 allow us to clearly reject the notion that consumers have perfect information about product features. All brand coefficients except for one are meaningfully different from zero, both in economic and statistical terms. Consumers receive information that reduces uncertainty in their beliefs about ϕ and, since the parameters of their utility function should be unchanged, we attribute changes in their choices to their updated beliefs about ϕ. With updated beliefs about ϕ, we are able to estimate $\tilde{\beta}$ by adding λ_n and ψ_n. This follows from the argument that $u_{ijt} = \tilde{\beta}_i \phi_{ij} + \epsilon_{ij}$ describes the decision utility in the "Information Only" arm. Row 1 in Panel B provides estimates of the implied utility weights when consumers have updated information about product features. In particular, we find that consumers appeared to have inaccurate information about the CMS star rating for particular plans. Providing them with more accurate information affects their decision-making and changes our estimate of how they value the CMS star rating. We estimate, for example, that consumers perceive the value of an extra CMS star rating to be \$740 ($\frac{\psi_2+\lambda_2}{-\psi_1-\lambda_1} * 100$) (See Panel C of Table 8).

Similarly, the estimates of η_1 to η_5 allow us to reject that consumers have no uncertainty in the parameters of their own payoff functions. As Panel A reports, η_1 to η_5 are statistically distinct from λ_1 to λ_5 at conventional levels for cost parameters, CMS star rating features, and AARP brand. They are also distinct in economic

terms. For example, the estimates suggest that under the "Information and Expert" arm, consumers value an additional CMS star rating at $257 less than consumers in the "Information Only" Arm ($\frac{\psi_2+\eta_2}{-\psi_1-\eta_1} * 100$). More generally, under "expert" preferences, consumers appear to put less weight on non-pecuniary plan features such as CMS quality ratings and insurer brands.

Taken together these estimates suggest: first, consumers are likely both to have imperfect information about plan features and to be uncertain how to evaluate the features that they observe. Second, expert recommendations—delivered through an online decision-support tool—not only update consumer information about product features, but also change how consumers value product characteristics.

4.3 Normative Implications

We next use our estimates to quantify how much the provision of expertise improves consumer welfare. To accomplish this, we start by simulating—for all 29,451 individuals who were invited to participate in the trial—consumer choices under three different scenarios using utility functions as estimated under the control arm, the "Info Only" arm, and the "Info + Expert" arm. In addition, we record the top plan that was recommended by our algorithmic expert.

In general, without knowing true normative preferences, one cannot quantify the extent to which interventions that provide information improve welfare. In other words, while we observe changes in choices in response to informational interventions, we lack an empirical benchmark for the consumers true normative preferences. However, we can still shed some light on the normative aspects of the interventions we analyzed through two complementary exercises. First, we calculate the share of consumers for whom the top ranked plan varies depending on which utility function we use to simulate consumer choices. We estimate that only for 24% of consumers do plan rankings from all three utility models and the expert ranking arrive at the same highest-rank plan. In that sense, many consumers could plausibly be making sub-optimal choices and leaving surplus on the table.

The second exercise allows us to quantify how much surplus is at stake. Taking the expert recommended plan as a benchmark (without loss of generality), we compute for each consumer the difference in expected out of pocket spending between the expert recommended plan and the plan that each consumer picks in each of three other utility-based simulations. In Panel A of Table 9 we report several moments of the distribution of differences in out of pocket spending from three (control and two treatment arms) utility rankings relative to the expert's choice. For those consumers for whom the choice of the top plan differs across rankings, the stakes are significant. The average cost savings that a consumer "foregoes" by choosing based on her own information rather than expert advice varies between $403 to $537 depending on which set of preferences we use.[14] These foregone savings are very large relative to the average spending in the top expert recommended plan of $843.

The average masks substantial heterogeneity in foregone savings. A quarter of all consumers would not be

[14] We use the term foregone cost savings here for simplicity. Conceptually, consumers may be making optimal choices conditional on their private information about their risk aversion and their expectations about future pharmaceutical needs. Under this interpretation, consumers are not foregoing savings, but rather exhibit a large willingness to pay for features that are not valued by the expert (such as insurance brands) or are choosing based on information about future risk that the expert does not utilize.

foregoing substantial savings: the 25th percentile of foregone savings relative to the expert plan is $0 under preferences from 'Information Only" and "Information + Expert" arms, and $29 if using preferences of the control arm. For consumers at the top of this loss distribution, however, substantial amounts are at stake. At the 95th percentile of the savings distribution, as Panel A of Table 9 reports, plans that have the highest utility under the "control group" preferences, are nearly $1,550 more expensive on average than the expert-recommended plan.

A cost-effective informational intervention would want to target consumers who are likely to benefit the most from the intervention. Suppose we take foregone savings as the measure of potential benefit. Are consumers with higher foregone savings more likely to take up algorithmic expertise that could steer them to lower-cost plans? Panel B of Table 9 reveals that this is not the case. The probabililty of taking up the intervention is very similar across the percentiles of likely savings, suggesting that self-targeting does not result in an efficient exposure to information in our setting. Among consumers who were offered to participate in the trial, those who had the potential to save the most (relative to the expert recommended plan) by changing their plan, were not more likely to participate. This finding points to the potential challenges of targeting informational interventions in the financial products domain, which we turn to next.

5 Who Demands Expertise?

We next examine what types of consumers demand algorithmic expert advice. To accomplish this, we take advantage of the imperfect take-up of our experiment as well as the imperfect compliance with experimental assignment. We use two empirical strategies to quantify and characterize selection. We consider selection on both levels and slopes. First, how do (potential) outcomes in the absence of the intervention differ between those who take up software and those who do not (levels); and second, what is the difference in the size of the likely treatment effects between those who express interest in receiving expert advice and those who do not (slopes). Understanding who ultimately chooses to take up decision support software is crucial for predicting which types of consumers policy-makers could reach by using algorithms to scale up expert advice.

5.1 Selection on Potential Outcomes: Lower Bound

Our first strategy exploits two ideas. First, the IV estimates correct selection bias from imperfect compliance to the experimental assignment. Second, the 2SLS estimates in our setting deliver the average treatment effect among those consumers treated by our intervention rather than only among compliers. This happens because only one-sided non-compliance is possible in our experimental setup—individuals not assigned to treatment do not get login passwords for the online tool. Hence, there are no always-takers.[15] The difference between the OLS estimates and the treatment on the treated estimates—given by 2SLS in our case— quantifies selection into the use of software among those who signed up for the trial. For example, trial participants who are more

[15]It is conceivable that patients who know each other could share their passwords. Indeed, we assume that this is likely the case for spouses, which is why the randomization was done at the household level. We do not see any evidence of take-up among those who were assigned to the control group, however.

active shoppers and are considering changing their plan even in the absence of our intervention, may be more likely to use our software. Hence, to quantify the selection bias, we first estimate the following OLS relationship:

$$Y_i = \tau_0 + \tau_1 UE_i + \tau_2 UI_i + \kappa_0 X_i + \epsilon_i \tag{9}$$

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. Using the potential outcomes notation, where Y_{1i} denotes whether i switches plans when using software and Y_{0i} denotes whether i switches plans when not using software, we know that τ_1 estimates (omitting conditioning on X_i to simplify notation):

$$\begin{aligned} &E[Y_i|UE_i = 1] - E[Y_i|UE_i = 0] = \\ &\qquad E[Y_{1i}|UE_i = 1] - E[Y_{0i}|UE_i = 1] + E[Y_{0i}|UE_i = 1] - E[Y_{0i}|UE_i = 0] \end{aligned} \tag{10}$$

τ_1 from the OLS regression estimates the left hand side of this expression. On the right hand side, we have the sum of the treatment effect on the treated (the first difference) as well as the selection effect (the second difference). We can use our 2SLS results to estimate the first difference. In general, 2SLS with heterogeneous treatment effects estimates the average treatment effect among compliers and thus does not give us the first difference on the right hand side of Equation 10. However, in the special case when there are no always takers - which is the situation we have - 2SLS captures the average treatment effect among the treated and thus gives us exactly the first difference on the right hand side of Equation 10. The difference between OLS and 2SLS, in turn, allows us to learn about the magnitude of the selection bias: $E[Y_{0i}|UE_i = 1] - E[Y_{0i}|UE_i = 0]$. Panel A of Table 11 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 treatment-on-the-treated 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—in the absence of intervention—than those individuals who were assigned to treatment, but chose not to use the software.

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 would have been more likely to revise their plan choices towards lower cost plans even in the absence of the intervention. The magnitude of selection is substantial relative to the treatment effect, especially in plan switching.

There is one important caveat to consider, however. These results are estimated relative to the average outcome among those assigned to the control group. The average outcome in the control group could itself be subject to treatment effects if simply entering the study website generated a "reminder effect" that affected consumers' behavior. In this sense, we could be underestimating the treatment effect and also underestimating the degree of selection. In this sense, the difference between the OLS and IV gives us a lower bound of selection. We next estimate an upper bound.

5.2 Selection on Potential Outcomes: Upper Bound

Our two-step experimental design allows us to directly observe selection within the control group. Consumers who were randomized into the control group did not know that they were in the control group until they logged in into the experimental website. Since we can observe whether someone in the control group logged in or not, we can measure the difference in outcomes between those who chose to (try) access the software and those who did not. The difference in behavior between these two groups of consumers captures selection plus any "reminder" effects that the screen in Figure 1C could have generated. Hence, the comparison of these two groups gives us an upper bound estimate of selection. Given the low impact of generic reminders that has been found in the literature (Ericson et al., 2017), it is likely that the true magnitude of selection is close to this upper bound.

To measure the difference in behavior between those who logged in into the experimental website and those who did not, we estimate the following OLS regression among individuals in the control group only:

$$Y_i = \xi_1 LOGIN_i + \xi_2 X_i + \epsilon_i \tag{11}$$

Panel B of table 11 reports the estimates. Individuals who logged in - before knowing whether they were assigned to the treatment or the control arm - were 21 percentage points more likely to switch plans than those who 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 the intervention indicate that caution is warranted when interpreting the positive effects of algorithmic decision support software for policy-making. 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 the most from them.

5.3 Selection on Treatment Effects

We next examine the importance of self selection "on slopes": whether individuals that express their interest in decision-support tools by signing up for our trial are likely to have higher or lower treatment effects from exposure to algorithmic support relative to those who do not sign up.

We start by estimating heterogeneous (intent-to-treat) treatment effect functions. 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 2.3 and 2.4. 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 we compare the two treatment arms.[16] 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

[16]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 (2019). 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.

+ 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 10 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 predicted behavioral responses to the intervention.

Using the out of sample predictions, we next examine whether there were systematic differences in predicted treatment effects between those who decided to participate in the experiment and those who did not. 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. Table 12 reports the results of a regression of the predicted treatment effect for each outcome on an indicator that takes the 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 the intervention than those individuals who did participate. 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 experimental take up as a function of predicted treatment effects graphically. This figure plots the take-up rate of the experiment for each ventile of the predicted treatment effect. We observe that the take-up rate declines sharply with the estimated treatment effect, suggesting that individuals that would have responded to the software intervention (in terms of switching their plans) the most, were the 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 Drivers of Selection

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.

We find some evidence supporting the importance of both channels. On the cost side, we observe that consumers with relatively large estimated treatment effects had the lowest rates of EMR use, suggesting relatively low familiarity with information technology. Consumers may have rationally chosen not to enroll in the trial because they correctly expected that for them the costs of using the on-line tool exceeded the benefit. The alternative explanation is that consumers for whom the estimated treatment effects were the 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. Moreover, consumers may simply 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 and end up not demanding expert advice.

6 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 more complex decision environments. This rise of algorithms as a cheap way to scale expert advice could consequently change market allocations across a range of settings. How large these impacts are likely to be in practice remains poorly understood. We also have a poor understating of whether differential exposure to and take-up of the new technologies may have undesirable distributional consequences. While an extremely rich theoretical literature has examined the economics of expert advice, there is much less empirical evidence on how and why decision-makers respond to (either algorithmic or in-person) expert advice. There is even less evidence on what types of consumers demand expertise.

In this paper, we provide novel evidence on both of these issues using data from a randomized-controlled study that offered individualized algorithm-based decision making support to older adults choosing insurance plans. We draw three main conclusions from our experimental results. First, exposure to the decision support tool changed consumer behavior, making consumers more likely to switch their plans and select a plan with lower expected costs. The response to the intervention was more pronounced when consumers were exposed to expert advice in addition to personalized information about product features.

Second, our results suggest that consumers are likely to have noisy information not only about product features, but also about the parameters of their own utility function. As a result, when consumers are exposed to expert advice, they may update their signals either about product features, or about how to value these features, or both. The noise in consumer beliefs most likely leads to relatively small welfare losses, on average;

however, a significant number of consumers could benefit a lot from changing their choices.

The distinction between a consumer's uncertainty about the characteristics of a product versus uncertainty about how product features (should) map into utility, which we have proposed in this paper, is important beyond our specific empirical setting. Allowing for this distinction sheds a different light on numerous empirical findings of consumer mistakes in financial choices. In observational data one cannot distinguish whether suboptimal choices happen because consumers: (i) do not know how to map features into utility (i.e. make "mistakes", which is the most common interpretation in the literature); (ii) have noisy information about product features only; or (iii) both. This distinction is important for policy-making, as interventions that aim to address only (i) or only (ii) may end up having little effect on consumer behavior.

This distinction is also crucial for future 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 preference manipulation by strategic algorithms.

Finally, our results point to a strong selection into who demands (algorithmic) expertise. We document two types of selection: on potential outcomes in the absence of treatment (selection on levels) and on the treatment effects (selection on slopes). We find that individuals who took up our software conditional on having access to it, were inherently more active shoppers. 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 demand for expertise. Further, we find that individuals for whom we predict the largest treatment effects from exposure to expertise, were the least likely to demand such expertise.

A key contribution of our study is thus to demonstrate that the expected benefits of (algorithmic) expert advice are the largest for those consumers who are the least likely to use this advice. In other words, expert advice is complementary to existing knowledge. Scaling expertise through algorithms, then, is likely to accrue to consumers that need this expertise the least. This has important distributional and policy implications. Numerous government policies attempt to provide software-based help for beneficiaries to navigate financial products and enroll in public programs. While our findings do not necessarily invalidate the idea that intuitive tools with clear, simplified, algorithmic recommendations could improve choices if rolled out in a general population, our results suggest that merely offering algorithmic expertise is unlikely to reach those who need most help in navigating these environments.

References

Abaluck, Jason and Abi Adams, "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses," Working Paper 23566, National Bureau of Economic Research June 2017.

_ 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.

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.

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*, 5 2019.

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

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum, "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, October 2013, *103* (6), 2499–2529.

_ , Jeffrey Prince, and Joshua C. Teitelbaum, "Are Risk Preferences Stable across Contexts? Evidence from Insurance Data," *American Economic Review*, April 2011, *101* (2), 591–631.

_ , Maura Coughlin, Francesca Molinari, and Joshua C. Teitelbaum, "Heterogeneous Choice Sets and Preferences," *CEMMAP Working Paper CWP37/19*, July 2019.

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.

Brown, Zach Y. and Jihye Jeon, "Endogenous Information Acquisition and Insurance Choice," *University of Michigan and Boston University Working Paper*, 2019.

Bundorf, Kate, J Levin, and N Mahoney, "Pricing and Welfare in Health Plan Choice," *American Economic Review*, 2012, *102(7)*, 1–38.

Bundorf, M. Kate, Maria Polyakova, Cheryl Stults, Amy Meehan, Roman Klimke, Ting Pun, Albert Solomon Chan, and Ming Tai-Seale, "Machine-Based Expert Recommendations And Insurance Choices Among Medicare Part D Enrollees," *Health Affairs*, 2019, *38* (3), 482–490. PMID: 30830808.

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

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin, "Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism"," *University of Pennsylvania Law Review*, 04 2003, *151.*

Caplin, Andrew, Mark Dean, and John Leahy, "Rational Inattention, Optimal Consideration Sets, and Stochastic Choice," *The Review of Economic Studies*, 07 2018, *86* (3), 1061–1094.

Cardon, James H. and Igal Hendel, "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey," *The RAND Journal of Economics*, 2001, *32* (3), 408–427.

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.

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 Ivan Fernandez-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," Working Paper 24678, National Bureau of Economic Research June 2018.

Cohen, Alma and Liran Einav, "Estimating Risk Preferences from Deductible Choice," *American Economic Review*, June 2007, *97* (3), 745–788.

Coughlin, Maura, "Insurance Choice with Non-Monetary Plan Attributes: Limited Consideration in Medicare Part D," *Cornell University Working Paper*, November 2019.

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.

Decarolis, Francesco, Maria Polyakova, and Stephen P Ryan, "Subsidy Design in Privately-Provided Social Insurance:Lessons from Medicare Part D," *Journal of Political Economy*, forthcoming.

DellaVigna, Stefano and Ethan Kaplan, "The Fox News Effect: Media Bias and Voting*," *The Quarterly Journal of Economics*, 08 2007, *122* (3), 1187–1234.

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

_ **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.

Fadlon, Itzik and David Laibson, "Paternalism and Pseudo-Rationality," Working Paper 23620, National Bureau of Economic Research July 2017.

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.

Harsanyi, John C., "Games with Incomplete Information Played by "Bayesian" Players, I–III Part I. The Basic Model," *Manage. Sci.*, November 1967, *14* (3), 159–182.

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, 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.

Lightle, John, "The Paternalistic Bias of Expert Advice," *Journal of Economics & Management Strategy*, 2014, *23* (4), 876–898.

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.

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.

Ottaviani, Marco, "The economics of advice," *Universita Bocconi Working Paper*, 2000.

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.

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

Starc, Amanda and Robert J Town, "Externalities and Benefit Design in Health Insurance," *The Review of Economic Studies*, 09 2019. rdz052.

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.

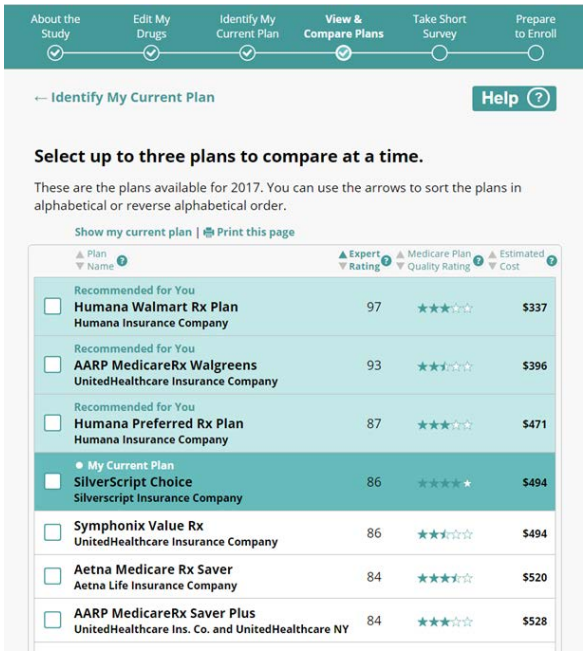
Tsai, George, "Rational Persuasion as Paternalism," *Philosophy & Public Affairs*, 2014, *42* (1), 78–112.

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.

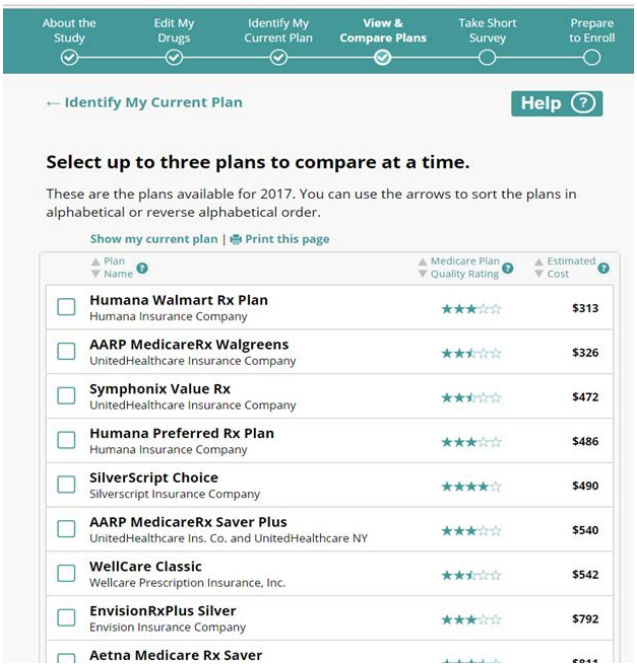
Figures and Tables

Figure 1: User Interface by Experimental Arm

A. Information + Expert Arm



B. Information Only Arm



C. Control Arm

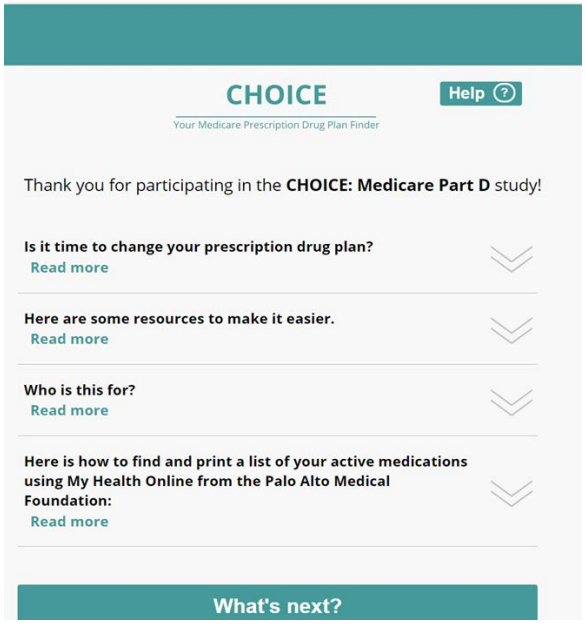


Figure 2: Experimental Design

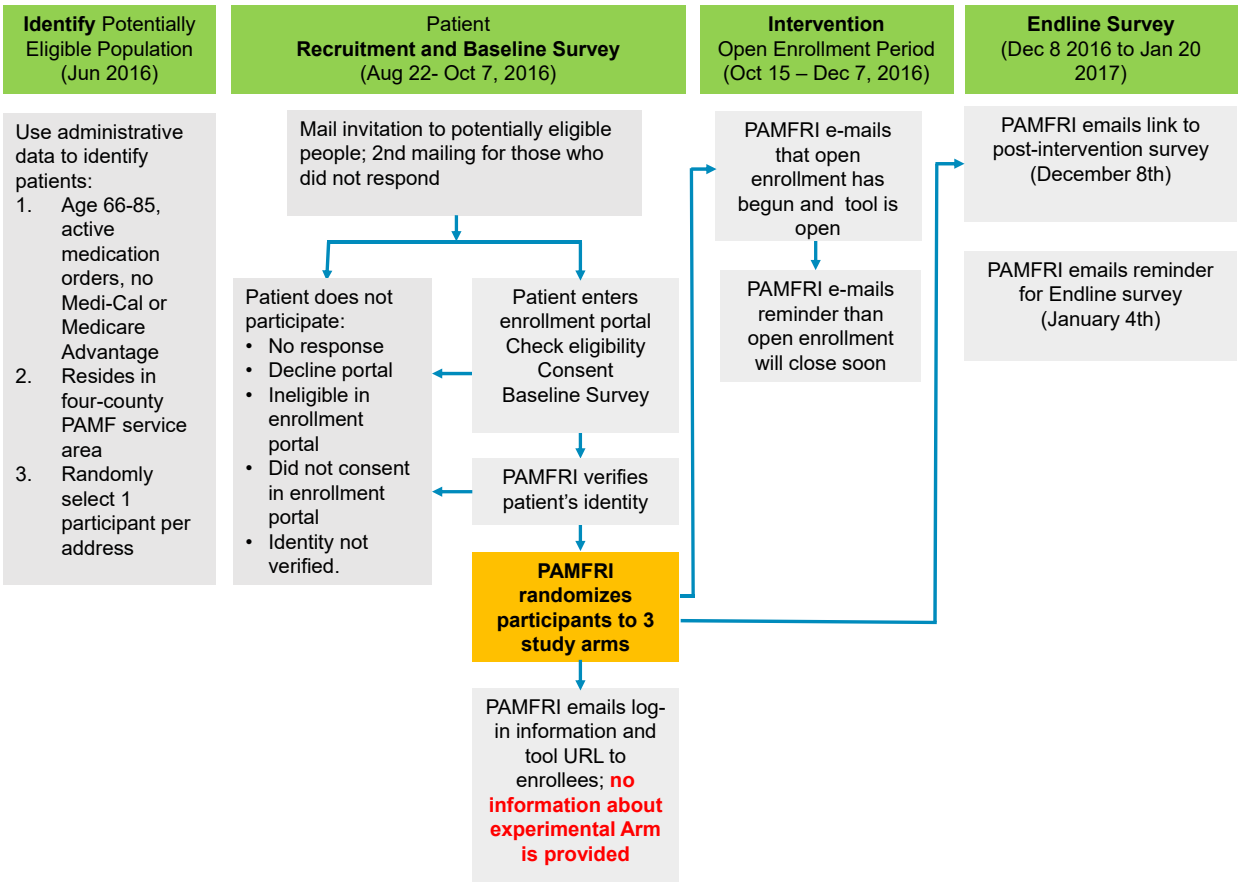
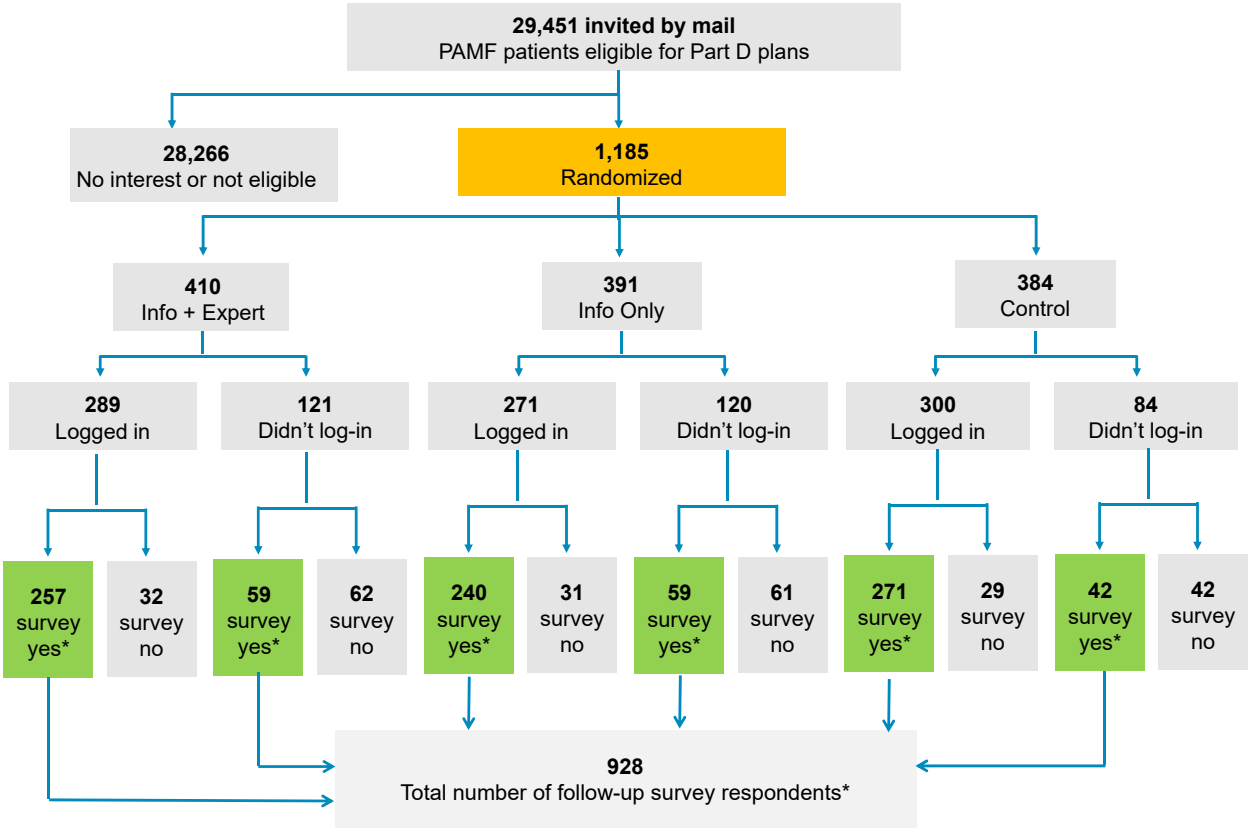


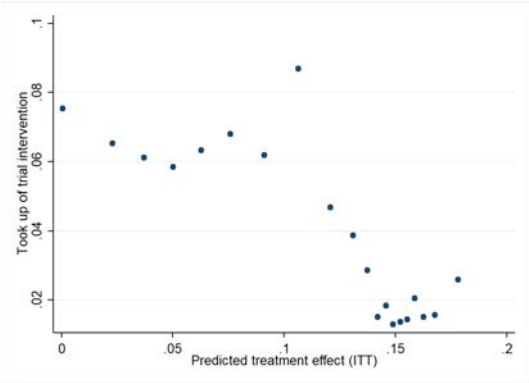
Figure 3: Enrollment Flow



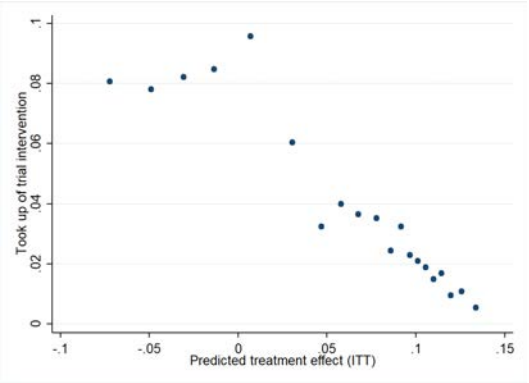
* Number of participants that responded to at least one survey question by pre-specified cutoff date

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

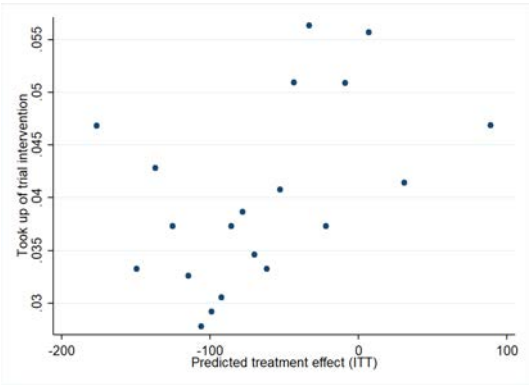
A. Switched Plans: Info + Expert



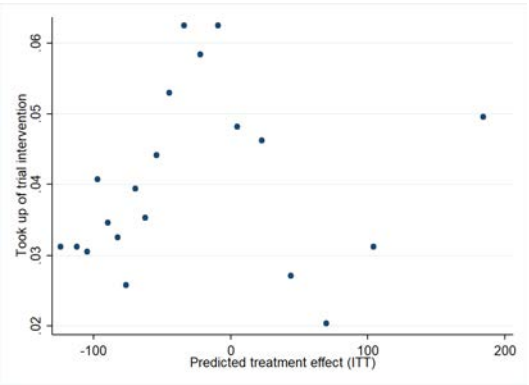
B. Switched Plans: Info Only



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.

‡ 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.

‡ 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.

‡ 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.

‡ 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.

‡ 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.

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)		(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.

† 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: Utility Model

	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)
p-value of χ^2-statistc for equality of λ vs. η	0.04	0.01	0.00	0.44	0.25
Panel B - implied utility weights					
Info Only Arm	-0.21	1.56	2.99	2.15	1.09
Info + Expert Arm	-0.17	0.80	2.08	1.81	0.84
Panel C - implied willigness to pay, $					
Info Only Arm	1	740	1416	1017	515
Info + Expert Arm	1	483	1253	1090	508

Tables reports estimates of empirical utility model. Panel A reports coefficient point estimates. 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 a random coefficient on the OOP Cost parameter. Standard errors are reported in parentheses. Panels B reports implied estimates of utility weights, summaing the main effect - ψ - with the interaction effects λ or η. Panel C reports implied willigness to pay amounts. To compute these we divide utility weights for each feature by the coefficent on the annual cost of plan, which measures the marginal utility of money.

Table 9: Distribution of Potential Benefits from Expertise

	Mean	5th percentile‡	25th percentile	50th percentile	75th percentile	95th percentile
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A - cost difference between top expert plan and simulated plan choice ($/year)						
Under "Control" preferences	537.2	0	27.9	204.1	835.6	1,499.3
as % of $ in expert plan	*64%*	*0*	*10%*	*52%*	*88%*	*53%*
Under "Info Only" preferences	427.6	0	0	204.3	782.1	1,146.5
as % of $ in expert plan	*51%*	*0*	*0%*	*52%*	*82%*	*40%*
Under "Info+Expert" preferences	403.2	0	0	36.2	623.2	1,270.1
as % of $ in expert plan	*48%*	*0*	*0%*	*9%*	*65%*	*45%*
Panel B - probability of trial take-up‡‡						
Probability of trial participation	0.040	0.042	0.040	0.037	0.041	0.040

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 simulate which plans would have been chosen under four scenarios: (1) using preferences as estimated for the experimental control group (2) using preferences as estimated under the "Information Only" treatment; (3) using preferences as estmated under the "Information + Expert" treatment; (4) the plan with the highest expert score. To simulate choices in scenarios (1) to (3), we first compute the "observable" part of utility, which includes five product features; then, for each individual we take one draw of the 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. The plan with the highest rank is then the simulated choice. For each simulation we record the expected out of pocket cost of the chosen plan. The table reports moments of the distribution of the difference (in levels and % terms) in the expected out of pocket cost between plans simulated under scenarios (1) to (3) and the cost in scenario (4). Pancel B reports the rate of trial participation for the whole sample (column 1), as well as within percentiles of the distribution of differences in expected cost between the expert plan and the plan chosen under scenario (1).

‡ Percentiles computed across 29, 451 individuals that were invited to participate in the trial
‡‡ We use the distribution of cost differences between the top expert plan and simulated plan choice under "control" group preferences

Table 10: 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 11: 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.

Table 12: 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	0.91	-0.02	-6.60	0.01
	(0.00)	(0.00)	(0.05)	(0.00)	(2.00)	(0.00)
Mean among randomized	0.09	0.05	0.89	0.08	-56.64	0.02
Std. dev. among randomized	0.05	0.04	1.61	0.07	67.86	0.06
No. of Obs.	29451	29451	29451	29451	29451	29451
Mean of Dep. Var.	0.11	0.05	1.76	0.07	-62.98	0.02
Std. Dev. Of Dep. Var.	0.05	0.03	1.61	0.07	62.08	0.05
Panel B: Information Only Treatment Effects						
Not randomized	0.04	0.02	0.08	-0.03	-3.25	0.01
	(0.00)	(0.00)	(0.08)	(0.00)	(2.42)	(0.00)
Mean among randomized	0.02	0.05	-1.36	0.07	-22.40	0.04
Std. dev. among randomized	0.06	0.07	2.88	0.07	81.74	0.04
No. of Obs.	29451	29451	29451	29451	29451	29451
Mean of Dep. Var.	0.06	0.06	-1.28	0.04	-25.51	0.05
Std. Dev. Of Dep. Var.	0.06	0.06	2.60	0.07	82.83	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.