Heterogeneous Treatment Effects and Efficient Policy Learning: Evidence from the Oregon Health Insurance Experiment

Shishir Shakya*

August 10, 2019

Abstract

The Medicaid expansion through the Affordable Care Act (ACA) has triggered a national debate among diverse stakeholders on the impacts of insurance coverage on various dimensions of health. Randomized experiments like, the Rand Health Insurance Experiment and the Oregon Health Insurance Experiment, have generated some credible estimates of the average treatment effects. However, identical policy intervention can often distinctly affect different individuals and subpopulations. This paper exploits Oregons health insurance lottery selection to estimate the heterogeneous treatment effects of access to public health insurance on health care use, personal finance, health, and wellbeing. For this, I use the cluster-robust generalized random forest − a causal machine learning approach. I find the federal poverty line, age, household size, and numbers of uninsured months interact on several levels to exhibit heterogeneous treatment effects. My findings are useful for analysts, policymakers, and insurance planners to discover the underlying mechanisms that drive the health outcome results and to design or reform policy.

Keywords: Insurance, causal machine learning, heterogeneous treatment effect, efficient policy learning

JEL Classification:

Note: This version is generated for Economics Graduate Student Conference submission of Washington University in St.Louis.

*Shishir Shakya, John Chambers College of Business and Economics & Regional Research Institute (RRI), West Virginia University, Morgantown, WV, 26506 E-mail: ss0088mix.wvu.edu.

1 Introduction

This research exploits Oregon's health insurance lottery selection as an instrument and contributes to two primary domains that are relevant for policy development. First, unlike the series of papers[1] that have evaluated the average treatment effects of the Oregon Health Insurance Experiments on several outcomes, this paper contributes by estimating the heterogeneous treatment effect of lottery insurance on several issues of interest like health care use, financial strain, and self-reported physical and mental health. Second, this paper contributes possible answers regarding how to target health insurance interventions for effective policymaking. Understanding "who should be treated" with intervention is ubiquitous in policymaking. It can be unfair, unethical, and sometimes illegal to target policy to only a particular subpopulation. Moreover, intervening everyone in the population (a blanket policy) is welfare-maximizing but can be costly.[2]

As of May 13, 2019, 37 states and the District of Columbia have expanded Medicaid coverage for low-income adults to 138% of the federal poverty level through the Affordable Care Act (ACA). This provision to expand[3] the Medicaid program through the Affordable Care Act (ACA) has triggered a substantial nationwide debate among policymakers and diverse stakeholders about what effects - if any - insurance coverage has on the various dimension of health (Baicker, 2019). The findings of this paper are valuable to meet some of the issues forward by the contemporary national debate. The results of this paper can exhibit the diverse impacts on the distinct population strata on health care use, personal finance, and wellbeing regarding the expansion of public access to insurance.

There exists an extensive literature studying the impact of insurance coverage on health outcomes which report average treatment effects. A concern with this literature is that establishing causal effect is challenging due to endogeneity. Endogeneity arises because it is difficult to control for observed and unobserved confounding variables among the insured and uninsured population (Levy and Meltzer, 2008). For example, a comparison of the health between those with and without health insurance, can reveal that insurance is detrimental for one's health (Baicker and Finkelstein, 2011) because people with poor health are more likely to get insurance compared to healthy people.

A Random assignment of insurance can circumvent such confounding problems (Finkelstein et al.,

[1] See Allen et al. (2010); Baicker et al. (2013, 2017, 2014); Baicker and Finkelstein (2011); Finkelstein et al. (2012); Grossman et al. (2016); Taubman et al. (2014); Zhou et al. (2017).

[2] For example, a provision of the Affordable Care Act (ACA) was that the federal government would pay the full cost of coverage expansion through 2016. Moreover, it would reimburse at least 90% of the cost of covering the newly-insured population (Norris, 2018). Oregon responded to this incentive by expanding Medicaid in January 2014 and ensured insurance to everyone with incomes up to 133% of the federal poverty line. When the federal government gradually reduced their payments, the state budget of Oregon (nearly $74 billion for 2017-2019) suffered about $1 billion budget hole due to the cost of health care (Foden-Vencil, 2018).

[3] Following the June 2012 Supreme Court decision, states face a decision about whether to adopt the Medicaid expansion. But, as per the Centers for Medicare and Medicaid Services (CMS) guidance, there is no deadline for states to implement the Medicaid expansion (Kaiser Family Foundation, 2019).

2012), and the Oregon Health Insurance Experiment renders a unique opportunity to test the causal effects of owning health insurance (Baicker and Finkelstein, 2011) on health and personal finance-related outcomes. In early 2008, Oregon's Department of Human Services applied for and received permission from the Centers for Medicare and Medicaid Services to add new members through random lottery draws from a new reservation list (Finkelstein et al., 2012). In the year following the random assignment, the treatment group had higher health care use, lower out-of-pocket medical expenditures and medical debt, and better self-reported physical and mental health than the control group, but it did not have detectable improvements in physical health conditions like high blood pressure − leaving policymakers with tough choices in balancing costs and benefits (Baicker, 2019).

Expensive randomized experiments like the Rand Health Insurance Experiment and the Oregon Health Insurance Experiments have generated some credible average effect estimates of expanding access to public health insurance on a population of interest. However, identical policy intervention can often distinctly affect different individuals and subpopulations in different ways. Along with average treatment effects, policymakers are usually interested in how effects of intervention vary across subpopulations. Identifying such heterogeneous treatment effects accommodate the discovery of underlying mechanisms that drive the results, which allows for more efficient design and reform of policy.

To investigate the heterogeneous treatment effects, one can stratify the data in mutually exclusive groups or include interactions in a regression (Athey and Imbens, 2017a). However, for large-scale investigations of effect heterogeneity, standard $p-$values of standard (single) hypothesis tests are no longer valid because of the multiple hypothesis testing[4] problems (Lan et al., 2016; List et al., 2019). Moreover, performing ad-hoc searches or $p-$hacking[5] to detect the responsive subgroups may lead to false discoveries or may mistake noise for an actual treatment effect (Davis and Heller, 2017). To avoid many of the issues associated with data mining or $p-$hacking, researchers can commit in advance to study only a subgroup by a preregistered analysis plan.[6] However, it may also prevent discovering unanticipated results and developing new hypotheses (Athey and Imbens, 2016).

This paper implements the Athey et al. (2019) cluster-robust version of the generalized random forest methods to explore the heterogeneous treatment effects of the Oregon Health Insurance Experiment. This method re-engineers the strengths and innovations of Breiman (2001) random forest − a predictive

[4]The "multiple hypothesis testing problems" leads to the so-called "ex-post selection problem," which is widely recognized in the program evaluation literature. For example, for fifty single hypotheses tests, the probability that at least one test falsely rejects the null hypotheses at the 5% significance level (assuming independent test statistics as an extreme case) is $1 - 0.95^{50} = 0.92$ or 92%.

[5]The $p-$hacking is an exhaustive search for statistically significant relations from combinations of variables or combinations of interactions of variables or subgroups. The $p-$hacking could lead to discovering the statistically significant relationship, when, in fact, there could have no real underlying effect.

[6]A preregistered analysis plan is sets of analyses plans released in the public domain by the researchers in advance prior they collect the data and learn about outcomes. For example, The American Economic Association's registry for randomized controlled trials is a reputable platform for conducting a preregistered analysis plan.

machine learning method − for causal inference. These modifications allow systematic investigation of the heterogeneous treatment effects that are not prone to data mining and $p-$hacking. Moreover, these methods are especially useful when research includes high-dimensional covariates. In this paper, I show the causal thresholds for distinct subpopulations where the impacts of Medicaid intensify and subdue. These realms have not been explored earlier and are some unique contributions to the literature. I scrutinize these separate subgroup for 36 different outcomes of interest. These outcomes are extensive and intensive margins of health care use, preventive care use, financial strain, mental and physical wellbeing and mechanisms of care, quality, and satisfaction of health care service usages.

"Who should get treatment?" are a widespread issue in policy design. For example, whom to serve in youth employment programs (Davis and Heller, 2017), whom to allocate Medicare funding for hip or knee replacement surgery (Kleinberg et al., 2015), who should get job training, job search, and other assistance (Kitagawa and Tetenov, 2018). This paper implements the efficient policy learning strategies of Athey and Wager (2018) to answer how to set eligibility criteria to intervene with insurance coverage.

This paper implements the efficient policy learning strategies of Athey and Wager (2018) to answer questions of how to set eligibility criteria to intervene with insurance coverage. This paper designs efficient policy rules considering two rationals − first, this paper constraint few observable covariates like race, gender, and residence e.t.c. Constraining specific covariates is essential for ethical, legislative, and political considerations. Second, this paper follows Kitagawa and Tetenov (2018) approach to design policy from an "intention-to-treat" perspective. This approach is crucial because the policy maker's problem is only a choice of the eligibility criteria and not the take-up rate. Individuals decide the take-up rate. I put forward various policies for each outcome of interest along with the cost of each of the policy compared with the random assignment policy. To model some worst-case scenarios of the purpose policies, I also develop lower-bounds for policy effectiveness, fairness, and balance.

In summary, this research uses Oregon Health Insurance Experiment public-use data and contributes to estimating the net impact of expanding access to public health insurance; examining the sources of treatment heterogeneity on such programs and offering an optimal policy rule for such program that could maximize health-related outcomes. The findings of this paper are useful for analysts, policymakers, and insurance designers to discover the underlying mechanisms that drive the health outcome results and to design or reform policy.

ection 2 summarizes the institutional background of the Oregon Health Insurance Experiment. Section 3 summarizes approaches to study health insurance and health outcomes and explains how causal machine learning can help to analyze different research questions. Section 4 lays out identification strategy and empirical methods for the cluster-robust random forest for heterogeneous estimation along with efficient policy learning strategies. Section 5 displays the results and provides discussions on findings.

Section 6 concludes the study.

2 Oregon health insurance experiment

Oregon's Medicaid program, the Oregon health plan (OHP)−created by one of the first federal waivers of traditional Medicaid rules−has two separate parts. First is the "OHP Plus". It serves for the categorically eligible Medicaid population. Low-income children, pregnant women, welfare recipients, and poor elderly and disabled populations groups are categorically eligible Medicaid population in Oregon (Office for Oregon Health Policy and Research, 2009). Second is the "OHP Standard". It servers poor adults who are financially but not categorically eligible for the Plus program. Adults ages 19−64 who are Oregon residents and U.S. citizens or legal immigrants without health insurance for at least six months and are who below the federal poverty level with assets below $2,000) are not categorically eligible Medicaid population in Oregon (Office for Oregon Health Policy and Research, 2009).

Except for vision and non-emergency dental services, the OHP Standard provides relatively comprehensive benefits with no consumer cost-sharing. The OHP Standard coverage includes physician services, prescription drugs, all significant hospital benefits, mental health, and chemical dependency services (including outpatient services), hospice care, and some durable medical equipment (Finkelstein et al., 2012; Baicker and Finkelstein, 2011). In 2001−2004, the average annual Medicaid expenditures for an individual on OHP Standard were about $3000, with monthly premiums that ranged below $20 depending upon income and was $0 for those below 10% of the federal poverty level (Wallace et al., 2008).

In early 2002, OHP Standard covered nearly 110,00 people, but in 2004, a budgetary shortfall halted the new enrollment in the OHP Standard; and by early 2008, attrition had reduced enrollment to about 19,000. However, in early 2008, the Oregon state had the budget to enroll an additional 10,000 adults. However, the demand for the program among eligible individuals would far exceed the 10,000 available slots. Therefore, Oregon's Department of Human Services applied for and received permission from the Centers for Medicare and Medicaid Services to add the new members through random lottery draws from a new reservation list (Finkelstein et al., 2012).

In early 2008, the Oregon state campaign an extensive public awareness program about the lottery opportunity focusing on the group that was not categorically eligible for the Plus program. Any qualified person could sign up from January 28 to February 29, 2008, by telephone, fax, in-person signup, mail, or online by providing very little demographic information. The sign up form required few demographics information like sex, date of birth, address, telephone number, PO box, and preferred language of communication (either English or Spanish) along with the list of names, sex, date of birth of anyone age

nineteen and older in the household whom they wished to add to their signup form.

No attempts were made to verify the information or screen for program eligibility at sign up for the lottery to keep the low barrier to sign up. During a window from January 28 to February 29, 2008, a total of 89,824 individuals signed up. Ineligible individuals for the OHP Standard are excluded before the lottery. The exclusion comprises individual residing outside of Oregon, individuals born before 1944 or after 1989, individuals with the OHP standard plan as of January 2008, individuals with an institutional address and individual who signup by an unrelated third party (Allen et al., 2010).

This exclusion leads to a sample that comprises 74,922 individuals (representing 66,385 households). After sign up phase, the Oregon state conducts eight lottery drawings that occurred during March through September 2008 and the lottery select 29834 individuals, and the remaining 45,088 individuals are controls.

Lottery selectees were sent a two-page application form[7]. Up to eight supplemental forms, (Allen et al., 2010) could accompany it. The selected individual was eligible to apply for OHP Standard for themselves and their family member (whether listed or not) and was required to submit the paperwork within 45 days. If they met the eligibility requirements, they could enroll in the Oregon Health Plan (OHP) Standard indefinitely. However, they had to verify their status every six months.

About 60% of the people who were selected by lottery send back the application. Half of those applications failed to meet the requirements. The primary reason was the requirement of income in the last quarter, corresponding to annual income below the poverty level. The federal poverty line in 2008 was $10,400 for a single person and $21,200 for a family of four (Allen et al., 2010). Therefore, about 30% of the total selected individuals successfully enrolled in the OHP Standard. Shortly after random assignment of lottery and OHP Standard application form, an "initial survey" was conducted and again after a year "main survey" was performed. These surveys consist of data of 58,405 individual comprising 29,589 individuals in treatment and 28,816 individuals in the control group.

3 Approaches to health insurance & health outcomes

How does health insurance affect health? The answer seems obvious, but Levy and Meltzer (2008) review the literature and draw three conclusions. First, the problem of endogeneity makes causal claims tenuous. Second, the papers that establish causal evidence are, particularly within small subgroup populations. For example, the public health insurance reduces mortality among infants and children (Currie and Gruber, 1996a,b; Hanratty, 1996) while for the elderly, public health insurance improves different outcome

[7] "The main form asked for the names of all household members applying for coverage and inquired about their Oregon residence, U.S. citizenship, insurance coverage over the past six months, household income over the past two months, and assets. Documentation of identity and citizenship and proof of income had to be returned with the completed form" (Allen et al., 2010).

but not mortality (Card and Maestas, 2008; Finkelstein and McKnight, 2008; McWilliams et al., 2007b,a). Third, the nature of studies is not representative of the broader population, which prohibits generalizing for policy purposes.

Allen et al. (2010) point out three practical design for insurance and health outcomes research: observational studies, quasi-experimental studies, and randomized experiments. The first is the observational studies which comprise the most substantial part in the literature. This type of studies typically utilizes the "multivariate regression" approaches. When implemented correctly, these approaches control the observable confounding variables between health insurance & health outcomes. However, these approaches are less likely to address the issues of unobservable confounders between health insurance & health outcomes. Failure to control unobservable differences between the insured and the uninsured may drive the observed differences in health outcomes (Levy and Meltzer, 2004, 2008).

The second set of study exploits natural experiments to evaluate the effect of health insurance on health outcomes. These studies implement techniques like differences-in-differences estimation, regression discontinuity designs, and instrumental variables. These techniques exploit an exogenous event that results in variation within health insurance coverage − changes that are plausibly unrelated to health and other underlying determinants of health insurance coverage (Levy and Meltzer, 2008). Exploiting an exogenous events makes the variation of the health insurance coverage take-up as good as random. In other words, health insurance coverage varies in a way that is unrelated to the unobservable factor. Thus a comparison of various outcomes between insured and uninsured are likely to support a causal interpretation.

However, the results of natural experiments are valid for only specific population groups and therefore, cannot be generalized to the broader population. For example, several studies show that public health insurance reduces mortality among infants and children (Currie and Gruber, 1996a,b; Hanratty, 1996) while for the elderly, public health insurance does not reduce mortality (Card and Maestas, 2008; Finkelstein and McKnight, 2008; McWilliams et al., 2007b,a). These "one size fit all" policy approach is unlikely to be useful for the broader population. For example, the channels or mechanisms through which having insurance affects health outcomes may be different for infants and children than they are for elderly adults.

The third set of studies are social experiments, which are the gold standard for establishing causality. The RAND health insurance experiments and the Oregon health insurance experiment are only two of such kind in the United States. Newhouse (1994) provides details on the RAND experiment while Finkelstein et al. (2012) provide details on Oregon experiments. Using RAND experiment data, Newhouse (1994) and Brook et al. (1983) find no significant effect of insurance on the health status of an average adult. Levy and Meltzer (2008) suggest the fact that the RAND experiment did not randomize people

to receive any health insurance. Instead, random individuals have treated health insurance with varying degrees of generosity. Finkelstein et al. (2012) study the Oregon health insurance experiment data. They find statistically significant higher health care utilization, lower out-of-pocket medical expenditures and medical debt, and better self-reported physical and mental health among the treatment group.

The observational studies, quasi-experimental studies, and randomized experiments often focus on causal inference and have been dominant in empirical policy research in health economics and economics in general. However, recently, due to availability of big-data and computing powers, machine learning approaches are gaining momentum among the researchers and policymakers. Several scholars like Varian (2014), Mullainathan and Spiess (2017), and Athey (2018) have briefed the utilities of the big-data and machine learning method in the field of economics. Within the domain of machine learning in economics, two strands of literature are gaining momentum: machine learning for policy prediction problems; and machine learning for causal inference problems.

The machine learning algorithms behave well for out-of-sample prediction as it utilizes flexible model selection, model ensembles, high dimensional data environment, and cross-validations. Therefore these algorithms are useful in many policy applications where the causal inference is not central or maybe not necessary. For example, Kleinberg et al. (2015) consider a resource allocation problem in health policy in which a policymaker needs to decide which otherwise-eligible patients should not be given hip replacement surgery through Medicare. They predict the probability that a candidate for a joint replacement would die within a year from other causes. Then they identify patients who are at particularly high risk and should not receive joint replacement surgery.

Similarly, Henderson et al. (2012) use satellite data on night lights to predict economic growth, and Glaeser et al. (2018) use Google Street View images to predict income in New York City. Glaeser et al. (2016) develop a system for allocating health inspectors to restaurants in Boston, and Naik et al. (2016) quantify the "urban appearance" from street-level imagery for 19 American cities and establish an empirical connection between the physical appearance of a city and the behavior and health of its inhabitants.

The machine learning algorithms are not well suited for causal inference. Establishing causal effect relates to understanding the counterfactual— what would happen with and without a policy— rather than just correctly predicting out-of-sample (Athey, 2018). However, some slight modifications of "off-the-shelf" or readily-available machine learning algorithm can utilize the strengths and innovations of machine learning algorithms for causal inference. The predictive machine learning algorithms are readily available with the open-source routines for the statistical software like Python and R.

The approaches that use machine learning methods for causal inference focus on estimating the average treatment effect, heterogeneous treatment effects, and optimal policies Athey (2018). This paper

implements causal machine learning approach mainly the "generalized random forest" of Athey et al. (2019) to explore the heterogeneous treatment effects of expanding access to public health insurance on various dimensions of healthcare utilization, personal finance, health, and wellbeing. Then, utilize efficient policy learning strategies of Athey and Wager (2018) for exploring some strategies that can help to refor or redesign access to public health insurance programs.

4 Empirical strategy

4.1 Identification

Finkelstein et al. (2012) provides the most detailed explanations and analyses of the Oregon Health Insurance Experiment. They give the intent-to-treat (ITT), and local average treatment (LATE) estimates for various outcome variables using the data from the "main survey" along with several other data sources. Though analyses in this paper consider similar outcome variables as Finkelstein et al. (2012), the interpretations are very distinct compared to their approach. This paper contemplates a situation where an analyst knows her outcome variable (Y) at post-treatment and has data of observables (X) at the pre-treatment period. Therefore, the sample in this study is not independent because the covariates are all drawn from the "initial sample" and merged to the outcome variables that are from the "main survey" sample. For this reason, this paper analyzes the data as an observational rather than a genuinely randomized study. This paper assumes unconfoundedness to identify causal effects. Unconfoundedness means that treatment assignment is as good as random conditionally on covariates Rosenbaum and Rubin (1983).

Consider $i \in \{1, \ldots, N\}$ observations with potential outcomes for each unit is either $\{Y_i(0), Y_i(1)\}$. Following Rosenbaum and Rubin (1983), the unit level causal effect is the difference in potential outcomes $\tau_i = Y_i(1) - Y_i(0)$, where, $W_i \in \{0, 1\}$ is a binary indicator for the treatment with $W_i = 0$ indicating that unit i received the control and $W_i = 1$ indicating that unit$i$ received the treatment. The X_i be a k-component vector of features or covariates not to be affected by the treatment. The data consist of triple $\left(Y_i^{obs}, W_i, X_i\right), \ \forall \ i = 1, \ldots, N$. The realized outcome for unit i is the potential outcomes corresponding to the treatment i.e. Y_i^{obs} is

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) \ if \ W_i = 0, \\ Y_i(1) \ if \ W_i = 1. \end{cases}$$

then, unconfoundedness can be formalize as:

$$\{Y_i(0), Y_i(1)\} \perp W_i | X_i.$$

4.2 Mean comparison of demographics

In this study, the outcome variables are health care utilization, preventive care utilization, financial strain, and health after a year of lottery assignment. The treatment variable is lottery selection, and observable covariates comprise pre-treatment demographics. This paper begins the analyses by comparing the mean of control and treatment group demographics.

$$X_{i,h} = \gamma_0 + \gamma_1 W_{i,h} + \eta_{ih} \tag{1}$$

where the X is observable demographics in the pre-treatment period, the γ_0 is mean of the control group and, the γ_1 is the mean difference between the control and treatment group. One should expect that the γ_1 to be statistically zero for comparable control and treatment groups. The selected individuals were eligible to apply for OHP Standard for themselves and their family member (whether listed or not); therefore, standard errors are household-level clustered and heteroscedasticity-consisted. Table 1 exhibits the results.

4.3 Intent to treatment effect of lottery

Secondly, this paper estimate the "intent-to-treat" (ITT) estimates of winning the lottery (i.e., the difference between treatment and controls). The ITT provides an assessment of the net impact of expanding access to public health insurance and causal inference to estimate ITT.

This paper utilizes Belloni et al. (2014b) double-selection post-LASSO approach. This method is based on the"LASSO"[8]. Under the assumption of sparsity[9], the double-selection post-LASSO approach select the observable confounders and covariates properly. After the double-selection post-LASSO method to select variables, the ordinary least squares (OLS) provides the ITT estimates.

The double-selection post-LASSO procedure comprises the following steps (Belloni et al., 2014a).

[8]The Least Absolute Shrinkage and Selection Operator (LASSO) is an appealing method to estimate the sparse parameter from a high-dimensional linear model is introduce by Frank and Friedman (1993) and Tibshirani (1996). LASSO simultaneously performs model selection and coefficient estimation by minimizing the sum of squared residuals plus a penalty term. The penalty term penalizes the size of the model through the sum of absolute values of coefficients. Consider a following linear model $\tilde{y}_i = \Theta_i \beta_1 + \varepsilon_i$, where Θ is high-dimensional covariates, the LASSO estimator is defined as the solution to $\min_{\beta_1 \in \mathbb{R}^p} E_n \left[(\tilde{y}_i - {}_i\Theta\beta_1)^2 \right] + \frac{\lambda}{n} \|\beta_1\|_1$, the penalty level λ is a tuning parameter to regularize/controls the degree of penalization and to guard against overfitting. The cross-validation technique chooses the best λ in prediction models and $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. The kinked nature of penalty function induces $\hat{\beta}$ to have many zeros; thus LASSO solution feasible for model selection.

[9]The "sparse" outcome model means a model with a few meaningful covariates affect the average outcome.

First, run LASSO of dependent variables on a large inventory of potential covariates to select a set of predictors for the dependent variable. Second, run LASSO of treatment variable (lottery) on an extensive list of potential covariates to choose a set of predictors for treatment. If the treatment is genuinely exogenous, one should expect this second step should not select any variables. Third, run OLS regression of dependent variable on treatment variable, and the union of the sets of regressors chosen in the two LASSO runs to estimate the effect of treatment on the dependent variable then correct the inference with usual heteroscedasticity robust OLS standard error.

$$Y_{i,h} = \beta_0 + \beta_1 W_{i,h} + x_{ih}\beta_2 + \varepsilon_{it} \tag{2}$$

where, the β_1 is the main coefficient of interest and gives the average difference in (adjusted) means between the treatment group (the lottery winners) and the control group (those not selected by the lottery). The β_1 is the impact of being able to apply for OHP Standard through the Oregon lottery (Finkelstein et al., 2012). The x_{ih} are selected from the X_{it} implementing the double-selection post-LASSO. The x_{ih} include the set of confounding variables that correlate with treatment probability (and potentially with the outcome) along with covariates that explain treatment and outcome. Therefore, controlling these covariates helps to estimate the "unbiased" relationship between winning the lottery and the outcome.

4.4 Local average treatment effect of lottery

The ITT estimates from equation 2 provides the causal effect of winning the lottery to apply for OHP standard. Another interesting causal parameter would be the impact of actual OHP standard insurance coverage rather than just the impact of winning the lottery to be eligible for OHP standard (ITT). In other word, policymakers are interested in the causal effect of compliance to the lottery and not just winning the lottery. The "complier" is the subset[10] of individuals who obtain insurance on winning the lottery and who would not obtain insurance without winning the lottery. One way to retrieve this parameter is to utilize lottery selection as an instrument and perform a two-stage least square (2SLS).

[10] Imbens and Angrist (1994) point out that there exist four possible groups of individuals based upon the compliance types: complier, always-taker, never-taker, and defier. The "complier" is the subset of individuals who obtain insurance on winning the lottery and who would not obtain insurance without winning the lottery. Never takers are a subset of individuals who never get insurance even after winning the lottery. Always takers will get insurance regardless of the lottery. The defier insured themselves when they are in the control group, and don't take insurance when they are in the treatment group. So, always taker and defier have insurance though they are in the control group. The never taker and defier won't take insurance though they win the lottery.

Equation 3 represents first stage equation and second stage equation respectively.

$$Z_{i,h} = \delta_0 + \delta_1 W_{i,h} + x_{ih}\delta_2 + \mu_{it}$$
$$Y_{i,h} = \phi_0 + \phi_1 \widehat{Z}_{i,h} + x_{ih}\phi_2 + \nu_{it} \tag{3}$$

where, $W_{i,h}$ is lottery assignment and it is an instrumental variable; $Z_{i,h}$ is an endogenous binary variable that takes a value of 1 if an individual is "ever in Medicaid" during the study period (from initial notification period until September 2009) or 0 otherwise. The first stage equation provides $\widehat{Z}_{i,h}$, which is the predicted value of "ever in Medicaid". The main coefficient of interest is ϕ_1 and is interpreted as a local average treatment effect (LATE) of Medicaid insurance Imbens and Angrist (1994) and identifies the causal impact of insurance among the "compliers". For just identified model, the LATE estimates ϕ_1 is the ratio of ITT estimates from equation 2 and the first-stage coefficient on wining the lottery from equation 3 or $\phi_1 = \frac{\beta_1}{\delta_1}$ (Finkelstein et al., 2012). Relative to the study population, "compliers" are somewhat older, more likely white, in worse health, and in lower socioeconomic status (Finkelstein et al., 2012).

4.5 Heterogeneous treatment effects of lottery

Numerous study examines the population marginal average treatment effect of having an insurance can be formalize using potential outcome framework as $\tau = E\left[Y_i(1) - Y_i(0)\right]$. However, this paper's main contribution is to examine the heterogeneous treatment effect of insurance on several health and personal finance related outcomes. The treatment heterogeneity can be expressed as the conditional average treatment effect (CATE) and can be formalized as $\tau(x) \equiv E\left[Y_i(1) - Y_i(0)|X_i = x\right]$.

This paper employs Athey and Wager (2019) cluster-robust random forest approach to access the treatment heterogeneity. This approach is based on the "causal tree" (Athey and Imbens, 2016), "causal forest" (Wager and Athey, 2018) and the "generalized random forest" (Athey et al., 2019) methods. The "causal tree" approach re-engineers Breiman et al. (1984) classification and regression tree (CART)[11] −a machine learning algorithms− for causal inference. The remaining methods extend the "causal tree" approach utilizing Breiman (2001) random forest[12] machine learning algorithm for causal inference.

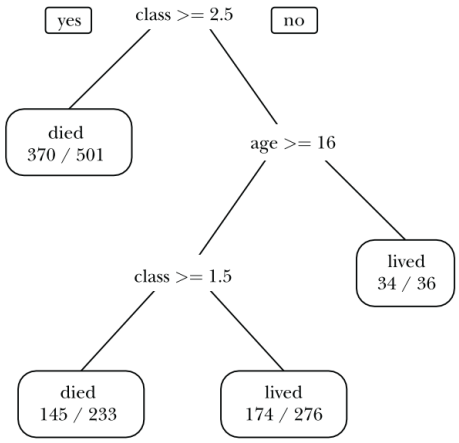
In a nutshell, the CART recursively filters and partitions the large data-set into binary sub-groups

[11] In simplest, the CART algorithm chooses a variable and split that variable above or below a certain level (which forms two mutually exclusive subgroups or leaves) such that the sum of squared residuals is minimized. This splitting process is repeated for each leave until the reduction in the sum of squared residuals is below a certain level (as defined by users), thus resulting a tree format (Athey and Imbens, 2017b).

[12] The Breiman (2001) random forest ensembles or bootstrap and aggregate many CART and report the average.

(nodes) such that the samples within each subset become more homogeneous that fit the response variable, thus resulting in a tree-like format. Figure 1 shows an example of the Titanic survivors features using the CART method, as shown by Varian (2014).

Figure 1: A Classification Tree for Survivors of the Titanic



Source: Varian (2014).*Interpretation:* If the class of travel is more than 2.5 (a third-class accommodation) 370 out of 501 died. Out of 36 people of the age-cohort below 17 who were in a second-class accommodation, 34 survived. Those who were age-cohort more than 16, if they were in second-class accommodation, 145 died out of 233, while 174 out of 276 died if they were in the first-class accommodation. These rules fit the data reasonably well, misclassifying about 30 percent of the observations in the testing set.

The CART minimizes the mean-squared error of the prediction of outcomes to capture heterogeneity in outcomes. However, the "causal tree" minimizes the mean-squared error of treatment effects to capture treatment effect heterogeneity. The approach to estimate the "causal tree" is similar to Imai and Ratkovic (2013) approach. A sample is split into two halves. One half is used to determine the optimal partition of covariates space. The other half is used to estimate treatment effects within the leave based on the optimal partition of covariates selected from the first partition (Athey and Imbens, 2016). This sample-splitting approach is known as "honest" estimation because model training and model estimation are independent. This approach leads to loss of precision as only half of the data is used to estimate the effect. However, this approach generates a treatment effect and a confidence interval for each subgroup that is valid no matter how many covariates are used in estimation. To prevent the loss of precision, this paper employs Chernozhukov et al. (2018a) cross-fitting approach (explained later in this section).

One caveat of the causal tree is that it does not provide personalized estimates. Wager and Athey (2018) utilize "random forest" machine learning approach and propose a "causal forest" method, where many different causal trees are generated and averaged, thus, can provide personalized estimates. This method provides causal effects that change more smoothly with covariates and provides distinct individualized estimates and confidence intervals. Wager and Athey (2018) also provide an important finding

that the predictions from causal forests are asymptotically normal and centered on the true conditional average treatment effect for each individual. Athey et al. (2016) extend the approach to other models for causal effects, such as instrumental variables, or other models that can be estimated using the generalized method of moments (GMM). In each case, the goal is to estimate how a causal parameter of interest varies with covariates.

4.6 Cluster-robust random forest

In a nutshell, random forest makes prediction as an average of b trees, as follow: (1) for each tree $b = 1, \ldots, B$, draw a subsample $S_b \subseteq \{1, \ldots, n\}$; (2) grow a tree via recursive partitioning on each such subsample of the data; and (3) make prediction by averaging the prediction made by individual tree as:

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \sum_{n=1}^{n} \frac{Y_i \mathbf{1}\left(\{X_i \in L_b(x), i \in S_b\}\right)}{|\{i : X_i \in L_b(x), i \in S_b\}|} \tag{4}$$

where, $L_b(x)$ denotes the leaf of the b^{th} tree containing the training sample x. For out-of-bag prediction, one can estimate the average as $\hat{\mu}^{(-i)}(x)$ by only considering those trees b for which $i \notin S_b$. Athey et al. (2019) show that a random forest can be equivalent as an adaptive kernel method and re-express the random forest from equation 4 as

$$\hat{\mu}(x) = \sum_{i=1}^{n} a_i(x) Y_i; \quad a_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{Y_i \mathbf{1}\left(\{X_i \in L_b(x), i \in S_b\}\right)}{|\{i : X_i \in L_b(x), i \in S_b\}|} \tag{5}$$

where, $a_i(x)$ is a data-adaptive kernel or simply these are weights that measure how often the i^{th} training example appears in the same leaf as the test point x. The treatment effect estimation is

$$\hat{\tau} = \frac{\sum_{i=1}^{n} a_i(x_i)\left(Y_i - \hat{m}^{(-i)}(X_i)\right)\left(W_i - \hat{e}^{(-i)}(X_i)\right)}{\sum_{i=1}^{n} a_i(x_i)\left(W_i - \hat{e}^{(-i)}(X_i)\right)} \tag{6}$$

where, $e(x) = P[W_i | X_i = x]$ is the propensity score or probability of being treated; $m(x) = P[Y_i | X_i = x]$ is expected outcomes marginalizing over treatment; $^{(-i)}$-superscript denote "out-of-bag" or "out-of-fold" prediction. Causal forest has several tuning parameters[13] and the cross-validation on "$R-$learner" objective function helps to select these tuning parameters. Nie and Wager (2017) showed that "$R-$learner" objective function for heterogeneous treatment effect estimation as

$$\hat{\tau}(\cdot) = \arg\min_{\tau} \left\{ \sum_{i=1}^{n} \left(\left(Y_i - \hat{m}^{(-i)}(X_i)\right) - \tau(X_i)\left(W_i - \hat{e}^{(-i)}(X_i)\right)\right)^2 + \lambda_n(\tau(\cdot)) \right\} \tag{7}$$

[13] These tuning parameters include the number of variables to try for each split, number of trees grown in the forest, a target for the minimum number of observations in each tree leaf, number of minimum node size for tree.

where, $\lambda_n(\tau(\cdot))$ is a regularize that controls the complexity of the learned $\hat{\tau}(\cdot)$ function.

At the implementation level, the causal forest starts by fitting two separate regression forests to estimate $\hat{m}(\cdot)$ and $\hat{e}(\cdot)$ and making out-of-bag prediction using these two first-stage forest. Then use these out-of-bag predictions as inputs to the causal forest where cross-validation on the "$R-$learner" objective function as given in equation 7 chooses the tuning parameters for causal forest.

The random forests in this paper employs the "honest" estimation as in Wager and Athey (2018). Furthermore, the lottery assignment was to the household rather than to an individual. Therefore, this paper grows random forests by drawing a subsample at household level rather than individual level. Similarly, the out-of-bag predictions are made using the household that was not in the training sample. Equation 8 exhibits effectiveness of intervention in individual, household and global levels.

$$\hat{\tau}_h = \frac{1}{n_h} \sum_{\{i:H_i=h\}} \hat{\Gamma}_i, \quad \hat{\tau} = \frac{1}{H} \sum_{h=1}^{H} \hat{\tau}_h, \quad \hat{\sigma}^2 \frac{1}{H(H-1)} \sum_{h=1}^{H} (\hat{\tau}_h - \hat{\tau})^2,$$
$$\widehat{\Gamma}_i = \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1 - \hat{e}^{(-i)}(X_i)\right)} \left(Y_i - \hat{m}^{(-i)}(X_i) - \left(W_i - \hat{e}^{(-i)}(X_i)\right)\hat{\tau}^{(-i)}(X_i)\right) \tag{8}$$

where, for the individual with household index $A_i \in \{1, \ldots, H\}$, the individual level effectiveness of lottery intervention is $\hat{\Gamma}_i$ and estimated based the "doubly-robust" estimator with cross-fitting Chernozhukov et al. (2018a). The household-level effectiveness of lottery intervention is $\hat{\tau}_j$. The global effectiveness of lottery intervention is $\hat{\tau}$ with standard error of $\hat{\sigma}^2$. The "doubly-robust" estimator is a variant of the augmented inverse-probability weighting. The name "doubly robust" means in the sense that estimates are consistent whenever either the propensity fit $\hat{e}(\cdot)$ or the outcome fit $\hat{m}(\cdot)$ is consistent, and are asymptotically efficient in a semiparametric specifications. The cross-fitting as suggested by Chernozhukov et al. (2018a) is similar to Athey and Imbens (2016) "honest" estimation. A sample is split into two halves. The first half (main sample) is used to determine the optimal partition of covariates space. The Second half (auxiliary sample) is used to estimate treatment effects within the leave based on the optimal partition of covariates selected from the first partition. Then flip the role of the main and auxiliary samples. Each of the estimates is "honest" or the two estimators will be approximately independent, so simply averaging them offers an efficient procedure Chernozhukov et al. (2018a). In Section 5, column (3) of Table 2, 3, 4 and 5 exhibits the estimates of $\hat{\tau}_j$.

A heuristic approach to gain qualitative insights about the strength of heterogeneity is to see how different are the doubly robust average treatment effects for the subgroup whose out-of-bag CATE estimates are below or above median Athey and Wager (2019). Davis and Heller (2017) also uses this

approach to test for heterogeneity. However, another for formal test is based on "best linear predictor" or BLP method of Chernozhukov et al. (2018b). The main idea is to compute the best linear fit of the target estimand using the forest prediction (on held-out data) as well as the mean forest prediction as the sole two regressors. A coefficient of one for mean forest prediction (MFP) suggests that the mean forest prediction is correct, whereas a coefficient of one for differential forest prediction (DFP) additionally suggests that the forest has captured heterogeneity in the underlying signal. The p-value of the DFP coefficient also acts as an omnibus test for the presence of heterogeneity: if the coefficient is significantly greater than 0, then one can reject the null of no heterogeneity. However, asymptotic results justifying such inference are not presently available.

4.7 Estimation of treatment policies

Once, policymaker understand the heterogeneity effect, she would like to assign the correct treatment to each individual or subpopulation. For that, I implement Athey and Wager (2018) strategy to find the policy function π that can map the observable characteristic of individuals X_i to an available set of treatment W_i.

$$\pi : X_i \rightarrow W_i \in \{+1, -1\}$$

Note, $W_i \in \{1, 0\}$ is reindexed as $W_i \in \{+1, -1\}$ which will help to formulate optimal policy assignment strategy later. Then an optimal treatment assignment policy can be given as π^* that maximizes expected utility, in our case, the health outcomes.

$$\pi^* \in \arg \max_{\pi \in \prod} E\left[Y_i\left(\pi\left(X_i\right)\right)\right]$$

Alternatively, any other non-optimal policy experiences the regret of $R(\pi)$, and we would like to minimize the regret function:

$$R(\pi) = E\left[Y_i\left(\pi^*\left(X_i\right)\right)\right] - E\left[Y_i\left(\pi\left(X_i\right)\right)\right]$$

Under unconfoundedness and the overlap assumption and binary treatment assignment Athey and Wager (2018) purpose a technique to estimate the regret, regret convergence and bound of the regret. They first determine the treatment effect $\hat{\Gamma}_i$ for each i using the double robust estimation technique called double machine learning of Chernozhukov et al. (2018a) and given as:

$$\hat{\Gamma}_i = \hat{\mu}_{+1}^{-k(i)}\left(X_i\right) - \hat{\mu}_{-1}^{-k(i)}\left(X_i\right) + W_i \frac{Y_i - \hat{\mu}_{W_i}^{-k(i)}\left(X_i\right)}{\hat{e}_{W_i}^{-k(i)}\left(X_i\right)}$$

This is doubly-robust estimator because only one of $\hat{\mu}$ or $\hat{e}$ needs to be correctly specified, and the term double machine learning is used because $\hat{\mu}$ and $\hat{e}$ can be semi- or non-parametric estimators. I use L_1-penalty logistic regression to estimate $\hat{\mu}$ and $\hat{e}$ with $k(i)$ fold cross-validation. If estimate is a positive treatment effect $\hat{\Gamma}_i$ I assign individual to treatment $(\pi(X_i) = 1)$ and if not then I assign individual to control $(\pi(X_i) = 0)$ and penalize for mismatch and maximize the following Q function to assess the effective policy:

$$\hat{Q}(\pi) = n^{-1} \sum_i \pi(X_i) |\hat{\Gamma}_i| sign\left(\hat{\Gamma}_i\right)$$

Further, Athey and Wager (2018) show that the regret has $\sqrt{n}\left(\hat{R}_{DML}(\pi) - R(\pi)\right) \xrightarrow{d} N\left(0, \sigma^2(\pi)\right)$ convergence and is bounded with the order of $\sqrt{VC(\Pi)/n}$ where $\hat{R}_{DML}(\pi)$ is the double machine learning estimates of regret. The bound provides a robust theoretical prediction that test error on any out of sample data is upper bounded with the sum of training error and $\sqrt{VC(\Pi)/n}$.

5 Results

The analysis presented in this paper utilizes data from the "initial survey" and the "main survey." The "initial survey" (administered shortly after random assignment of lottery and mailing of the OHP Standard application form to the lottery selectee) and the "main survey" (conducted after a year from the random assignment of the lottery) collect virtually similar data from 58,405 individual comprising 29,589 individuals in treatment and 28,816 individuals in the control group. Each of these individuals is adults of ages 19−64 who are Oregon residents and the U.S. citizens or legal immigrants without health insurance for at least six months and are below the federal poverty level have assets below $2,000.

5.0.1 Pre-treatment comparison of demographic characteristics

Employing the equation 1, Table 1 begins the analysis by presenting how different are treatment and control groups in their demographics in the pre-treatment period. These demographics are retrieved from the lottery list data and the initial survey data. Table 1 shows the mean of the control group and difference of means between treatment group and control group. Given the random assignment of insurance, one should expect that the mean of the treatment and control group should be statistically similar. Except for a few variables, the differences in the means between treatment and control group are statistically zero. There exist some anomalies that mean difference of few demographics is statistically non zero, but close to zero, which could be due to the large sample size.

Table 1: Pre-treatment comparison of demographic characteristics

Variable	Control mean	Mean diff	Variable	Control mean	Mean diff
% Female §	0.600	-0.015***	% dont currently work	0.527	-0.007
		(0.006)			(0.008)
% English preferred §	0.921	-0.009**	% work below 20 hours/week	0.096	-0.002
		(0.004)			(0.005)
% Self signup §	0.880	-0.045***	% work 20–29 hours/week	0.111	-0.003
		(0.004)			(0.005)
% Signed up on first day §	0.102	0.004	% work 30+ hrs/week	0.266	0.012*
		(0.004)			(0.007)
% PO Box address §	0.127	0.000	%FPL below 50%	0.436	-0.029***
		(0.005)			(0.009)
% MSA §	0.750	-0.004	%FPL 50–75%	0.125	0.005
		(0.006)			(0.006)
Age (as of 2008) §	42.33	-0.108	%FPL 75–100%	0.154	0.000
		(0.169)			(0.006)
% Race as White	0.838	-0.009	%FPL 100–150%	0.171	0.012*
		(0.006)			(0.007)
% Race as Black	0.031	-0.001	%FPL above 150%	0.114	0.011*
		(0.003)			(0.006)
% Race as Spanish/Hispanic/Latino	0.100	0.009*	% Insurance	0.293	0.145***
		(0.005)			(0.008)
% 4-year college degree or more	0.113	0.000	% OHP	0.067	0.158***
		(0.005)			(0.006)
% High school disploma or GED	0.506	-0.007	% Private insurance	0.028	-0.002
		(0.008)			(0.003)
% Less than high school	0.168	0.002	% Other insurance	0.055	0.00
		(0.006)			(0.004)
% Vocational training or 2-year degree	0.212	0.004	Household size	2.884	0.094***
		(0.007)			(0.029)

Notes: The initial survey consists of data of 58,405 individual comprising 29,589 individuals in the treatment group and 28,816 individuals in the control group. The variables collected from the lottery list for the population that appeared in the "initial survey" are marked with §. Enclosed in the parenthesis are household-level clustered heteroscedasticity-consistent standard errors. The ***, **, and * represent 1%, 5%, and 10% level of significance, respectively. The FPL represents the federal poverty line; in 2008, it was $10,400 for a single person and $21,200 for a family of four Allen et al. (2010). The variables presented in this table are similar to Finkelstein et al. (2012) paper. However, these estimates are different from theirs. They compare the means of treatment and control group using lottery list data (marked as §) for the observation of $n = 74922$ and the "main survey" data while this table utilizes "initial survey" data.

5.1 ITT, LATE and Heterogeneous treatment effects

The treatment effect often varies with individual's observable characteristics. For example, if the treatment is costly and less accessible, then only those who are likely to benefit most will take up the treatment. In this case, the availability of the treatment may reduce the average effect among the treatment recipients. While, on the other hand, if the treatment provided to the individuals who are less likely to benefit, then the availability of the treatment may increase the average effect among the treatment recipients. Therefore understanding the heterogeneity in treatment effects has important implications for policymakers mainly to yield valuable insights about how to distribute scarce social resources in an unequal society (Xie et al., 2012) by balancing the competing policy objectives, such as reducing cost, maximizing average outcomes, and reducing variance in outcomes within a given population (Manski, 2009).

As noted earlier, this paper contemplates a situation where an analyst knows her outcome variable (Y) at post-treatment and has data of observables (X) at the pre-treatment period. For this reason, this paper analyzes the data as an observational rather than a genuinely randomized study. Thus,

there is likely to be a treatment heterogeneity. Such a situation could also arise if there are unobserved household-level features that are an important treatment effect modifiers. For example, some household may have better access to care and probably implement the intervention better than others or may have a culture that is more receptive to the treatment.

To generalize the results outside the sample size, one needs to robustly accounts for the sampling variability of potentially unexplained household-level effects. This study takes a conservative approach and assumes that the outcome variables of an individual within the same household may be arbitrarily correlated within a household (or "cluster"). Therefore, utilize the cluster-robust analysis. Furthermore, to generalize beyond the household given in the data, each household is equally weighted such that, the model allows to predict the effect on a new individual from a new household.

Table 2, 3, 4 and 5 comprises various estimates for health care and preventive utilization, financial strain, self-reported health and potential mechanisms respectively. These outcome variables are taken from the "main survey" and proxy the causal effects after one year of insurance experiment experiences. Each of these table has several estimates. The estimates in column (1) shows "intent-to-treat" effect implementing double-selection post-LASSO method. The column (3) shows a local average treatment effect which can be interpreted as the impact of insurance coverage. The column (3) presents the doubly-robust average treatment effect which shows the average effectiveness of the lottery intervention on the outcomes.

Column (4), (5) and (6) explores the treatment heterogeneity. Column (4) provides a heuristic or qualitative insights about the strength of heterogeneity, it groups the out-of-bag CATE estimates to above or below the median CATE estimate then estimates average treatment effects in these two subgroups separately using the doubly robust approach to test if those average treatment effects are statistically similar or not. Column (5) and (6) provides test calibration for causal forest or the omnibus evaluation of the quality of the random forest based on the "best linear predictor" method of Chernozhukov et al. (2018b). It computes the best linear fit of the target estimand using the forest prediction (on held-out data) as well as the mean forest prediction as the sole two regressors. A coefficient of one for mean forest prediction (MFP) suggests that the mean forest prediction is correct, whereas a coefficient of one for differential forest prediction (DFP) additionally suggests that the forest has captured heterogeneity in the underlying signal. The $p-$value of the DFP coefficient also acts as an omnibus test for the presence of heterogeneity: If the coefficient is significantly greater than 0, then we can reject the null of no heterogeneity. Though the treatment heterogeneity is not detected, necessarily doesnot exclusively means non existence of treatment heterogeneity therefore a heat map plot is provided for a closer look for the location of heterogeneity when the ATE coefficient in the tables are reported significant.

5.1.1 Health care utilization

Table 2 panel A shows health care utilization on extensive and intensive margins. The extensive margin relates with if individual is currently taking any medication, has any outpatient visits, has any emergency visits, has any inpatient hospital admission in last six months while the intensive margins quantifies on the basis of numbers.

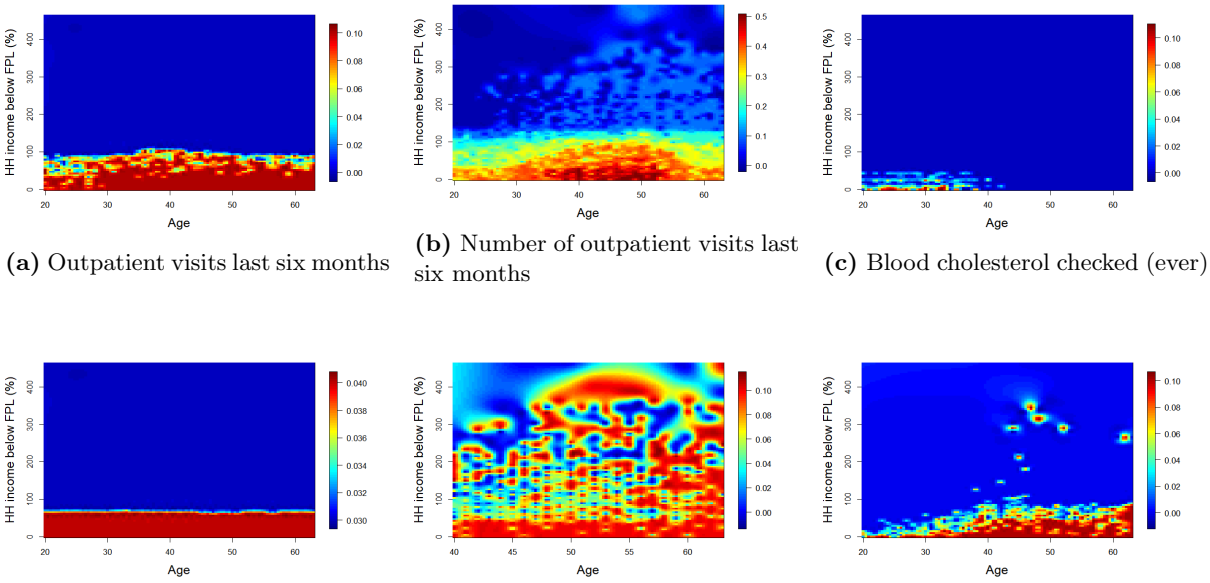
Table 2: Health care utilization

Outcome variables	ITT (1)	LATE (2)	ATE (3)	Heuristic (4)	MFP (5)	DFP (6)
Panel A: Health care utilization						
Extensive margins						
Currently taking any prescription medications	0.021**	0.067**	0.007	-0.018	0.801	-0.494
	(0.009)	(0.03)	(0.009)	(0.018)	(1.015)	(0.734)
Outpatient visits last six months	0.07***	0.224***	0.062***	0.055***	1.028***	1.316***
	(0.009)	(0.027)	(0.009)	(0.017)	(0.145)	(0.312)
ER visits last six months	0.009	0.029	0.005	-0.014	0.696	-3.331
	(0.008)	(0.024)	(0.008)	(0.015)	(1.172)	(1.816)
Inpatient hospital admissions last six months	0.002	0.005	0.001	-0.006	0.272	-0.626
	(0.004)	(0.014)	(0.005)	(0.009)	(2.322)	(1.4)
Intensive margins						
Number of prescription medications currently taking	0.104*	0.342*	0.042	-0.119	0.899	-0.383
	(0.055)	(0.177)	(0.055)	(0.109)	(1.219)	(1.005)
Number of Outpatient visits last six months	0.335***	1.087***	0.304***	0.426***	1.037***	1.502***
	(0.052)	(0.166)	(0.055)	(0.11)	(0.188)	(0.373)
Number of ER visits last six months	0.006	0.018	-0.003	-0.115***	1.97	-10.89
	(0.016)	(0.053)	(0.017)	(0.035)	(14.846)	(2.98)
Number Inpatient hospital admissions last six months	0.007	0.024	0.007	0.008	0.713	-2.071
	(0.007)	(0.021)	(0.007)	(0.014)	(0.661)	(1.974)
Panel B: Preventive care utilization						
Blood cholesterol checked (ever)	0.036***	0.116***	0.035***	0	1.043***	1.022*
	(0.008)	(0.026)	(0.008)	(0.016)	(0.236)	(0.73)
Blood tested for high blood sugar/diabetes (ever)	0.038***	0.121***	0.035***	0.003	0.982***	-1.588
	(0.008)	(0.025)	(0.008)	(0.017)	(0.235)	(1.618)
Mammogram within last 12 months (women 40)	0.078***	0.249***	0.063***	0.048*	0.992***	2.036***
	(0.013)	(0.039)	(0.014)	(0.027)	(0.213)	(0.697)
Pap test within last 12 months (women)	0.053***	0.18***	0.047***	0.037*	1.003***	2.159***
	(0.01)	(0.034)	(0.011)	(0.022)	(0.23)	(0.671)

Notes: The ***, **, and * represent 1%, 5%, and 10% level of significance, respectively. Enclosed in the parenthesis are household-level clustered heteroscedasticity-consistent standard errors. The regressions in columns (1) and (2) include household size dummies, survey wave dummies and survey wave interacted with household size dummies. For the LATE estimates in column (2), instrumental variables is lottery assignment and endogenous variables is "Ever in Medicaid". The ITT and LATE estimates are base on the double-selection post-LASSO.

The ITT and LATE estimates in Table 2 panel A shows that on both margins there are substantial and (mostly) statistically significant increases in prescription drugs and outpatient use. However, the doubly robust ATE estimator shows significant effect for the outpatient usages only. The average treatment

Figure 2: Health care and preventive care utilization and financial strain



(a) Outpatient visits last six months

(b) Number of outpatient visits last six months

(c) Blood cholesterol checked (ever)

(d) Blood tested for high blood sugar/diabetes (ever)

(e) Mammogram within last 12 months (women 40+)

(f) Pap test within last 12 months (women)

effect of wining the lottery is associated with about 0.30 (std. err. = 0.06) increase in the number of outpatient usages among. Table 2 panel B shows the preventive care utilization. The ITT and ATE estimates are about similar and statistically significant suggesting winning lottery increases the likelihood for preventive cares like blood test cholesterol, blood test for diabetes, Mammogram test (for women of age 40+) and Pap test (for women).

Table 2 column (4) shows the heuristic approach to test the treatment heterogeneity. Some qualitative evidences of treatment heterogeneity are explored for the outpatient usages and preventive care utilization. Table 2 column (5) shows the MFP and column (6) shows the DFP. The MFP and DFP are close to unit and statistically non zero suggesting treatment heterogeneity among these variables.

Note that 2000 causal trees were ensemble to develop cluster-robust random forest. Among these 2000 causal trees, algorithm always select the age and the household income below federal poverty along with household size and other variables like education, employment. Appendix A provides variable importance table for all the outcome variables. It list the variables which were split (more than average) by the random forest. Therefore, for the illustration purpose, I develop a heat map by grouping age and percentage of household income below FPL and average the out-of-bag conditional average treatment. The heat map has age in the x−axis and percentage of household income below FPL in the y−axis. The treatment heterogeneity for the outpatient usages and preventive cares are presented in Figure 2 panel (a) to (e).

Figure 2 panel (a) and (b) show an insight of outpatient usage CATE over age and household income below FPL. It appears that outpatient usage CATE (in extensive margin) for lottery winners is high and positive for those who belongs to the household below 100% of FPL regardless of age cohorts. Similar is true for intensive margin of outpatient usage CATE, however, some additional heterogeneity can bee seen for age-cohorts.

Figure 2 panel (c) exhibits treatment heterogeneity if the blood test for cholesterol level were ever done within the study period. Mostly age cohorts between 20 to 40 who belong to household close to the FPL have higher likelihood for this preventive test. Figure 2 panel (e) shows treatment subgroup who are in household below 80% of the FPL are more likely to blood test for diabetes. Figure 2 panel (e) and (f) illuminates CATE for the Mammogram test (for women of age above 40) and the Pap test (for women). It appears that women of age 40 years and above who belongs below 50% of FPL household are highly likely to Mammogram test. Post 5o years, women are likely to Mammogram test regardless of the household income below the FPL. The CATE of the Pap test shows, women who belongs to household close to FPL or 100% below the FPL are likely to test for the PAP.

5.1.2 Financial Strain

Table 3 displays extensive margins and intensive margins of the financial strains. Wining lottery is associated with lower financial strains both in extensive and intensive margins. The ITT and ATE estimates for financial strains in intensive margins quantifies the results in dollars terms as the net effect of winning lottery. The ITT and ATE ranges shows winning lottery relates with reduction of various types of out of pocket cost for past six months. The ITT and ATE estimates ranges shows on average $20 reduction on out of pocket costs for doctors visits, clinics or health centers ; nearly $40 to $49 reduction on out of pocket costs for emergency room or overnight hospital care, about; $13 to $15 reduction on out of pocket costs for medical care and nearly about $ 50 reduction on the total out of pocket cost for medical care. Other than these financial strains, the winner group also has nearly $450 to $500 on average reduction of the medical debts.

The "best linear prediction" (BPL) model shows the treatment heterogeneity in the out of pocket expenses (last 6 months) only. Again, this does not necessarily mean that there are no heterogeneity because the BPL acts as an omnibus test for the presence of heterogeneity. A closer look at the heat map presented in Figure 3 illuminates the some sources of treatment heterogeneity.

Figure 3 panel (a) shows the reduction on the CATE for the extensive margin on the out of pocket medical expenses (last 6 months) suggesting lower financial strain for all age group and for all household and less pronounced for household that range between 100% to 200% below the federal poverty line. Figure 3 panel (b) exhibits a sharp discontinuity of owing money for medical expenses for household

Table 3: Financial strain

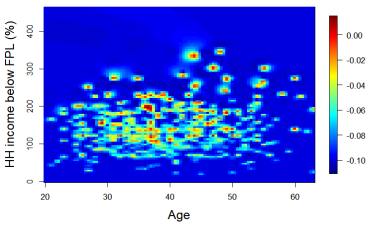
Outcome variables	ITT (1)	LATE (2)	ATE (3)	Heuristic (4)	MFP (5)	DFP (6)
Extensive margins						
Any out of pocket medical expenses, last six months	-0.073*** (0.009)	-0.238*** (0.029)	-0.073*** (0.009)	0.028 (0.018)	1.021*** (0.125)	1.449*** (0.562)
Owe money for medical expenses currently	-0.053*** (0.009)	-0.17*** (0.027)	-0.058*** (0.009)	0.038** (0.018)	1.076*** (0.169)	0.87 (1.253)
Borrowed money or skipped other bills to pay medical bills, last six months	-0.057*** (0.009)	-0.184*** (0.028)	-0.064*** (0.009)	0.008 (0.017)	1.061*** (0.145)	0.473 (1.323)
Refused treatment because of medical debt, last six months	-0.012** (0.005)	-0.037** (0.015)	-0.013*** (0.005)	0.006 (0.009)	1.054*** (0.387)	-3.706 (2.121)
Intensive margins						
Out of pocket costs for doctors visits, clinics or health centers, past 6 months	-19.308*** (3.46)	-61.429*** (10.919)	-20.175*** (3.594)	-8.47 (7.192)	0.999*** (0.179)	0.371 (0.664)
Out of pocket costs for emergency room or overnight hospital care, past 6 months	-49.519** (21.611)	-157.71** (67.674)	-40.73** (18.46)	14.213 (36.89)	1.035** (0.468)	0.211 (0.689)
Out of pocket costs for prescription medicine, past 6 months	-15.042** (6.941)	-45.756** (22.054)	-12.747** (6.012)	2.234 (12.067)	0.889** (0.403)	-1.116 (1.405)
Out of pocket costs for other medical care, past 6 months	-3.431 (2.088)	-10.577 (6.55)	-3.052 (2.083)	-7.223* (4.188)	0.894* (0.617)	-3.693 (1.492)
Total out of pocket costs for medical care, last 6 months	-48.203*** (9.552)	-152.815*** (30.393)	-53.793*** (9.751)	13.3 (19.707)	1.034*** (0.188)	0.489 (0.732)
Total amount currently owed for medical expenses	-442.39*** (96.744)	-1447.906*** (318.1)	-496.084*** (105.023)	167.277 (208.674)	1.038*** (0.223)	-0.298 (1.125)

Notes: The ***, **, and * represent 1%, 5%, and 10% level of significance, respectively. Enclosed in the parenthesis are household-level clustered heteroscedasticity-consistent standard errors. The regressions in columns (1) and (2) include household size dummies, survey wave dummies and survey wave interacted with household size dummies. For the LATE estimates in column (2), instrumental variables is lottery assignment and endogenous variables is "Ever in Medicaid". The ITT and LATE estimates are base on the double-selection post-LASSO.

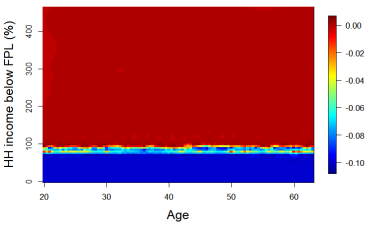
those are below 100% FPL. The effect is virtually zero for rest of the household for all age groups. Figure 3 panel (c) shows that lottery winners are not likely to be refused for treatment because of the medical debt. Figure 3 panel (d) as the CATE of borrowing money or skipped other bills to pay medical bills shows a reduction except for household that range between 100% to 200% below the federal poverty line.

Figure 3 panel (e) shows that more than $25 up to $30 reduction of out of pocket costs for doctors visits, clinics or health centers in past 6 months for age group 50 plus who belongs to the household that are below 80% of the FPL. The 40 below age group of the household within the range of 80% to 200% of the FPL less than about $15 reduction of such cost. The rest of the subgroup has about average of $20 reduction of such cost. Figure 3 panel (f) shows about $60 to $70 or little more reduction in the out of pocket costs for emergency room or overnight hospital care in past 6 months for 40 below age group for the household those are below 100% below the FPL. The reduction of such cost is less than $20 for

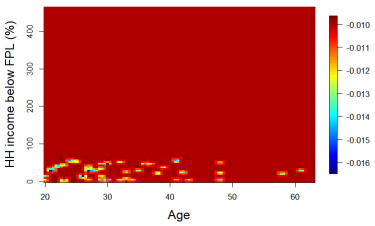
Figure 3: Financial strain



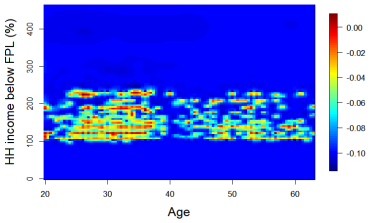
(a) Any out of pocket medical expenses, last 6 months



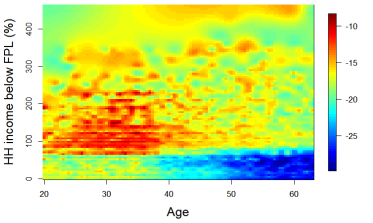
(b) Owe money for medical expenses currently



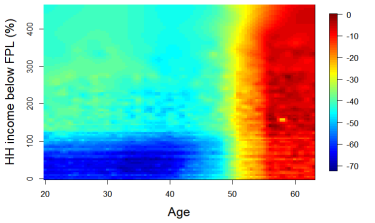
(c) Refused treatment because of medical debt, last six months



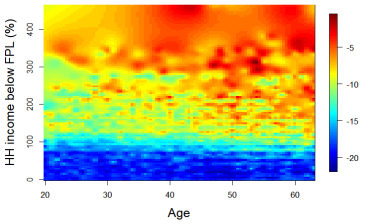
(d) Borrowed money or skipped other bills to pay medical bills, last 6 months



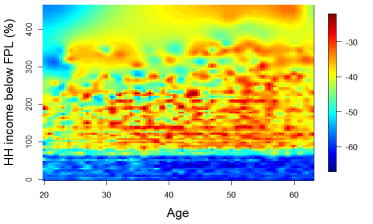
(e) Out of pocket costs for doctors visits, clinics or health centers, past 6 months



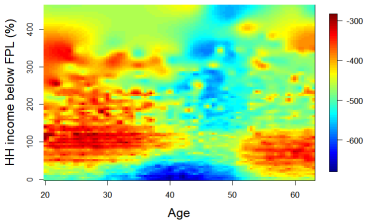
(f) Out of pocket costs for emergency room or overnight hospital care, past 6 months



(g) Out of pocket costs for prescription medicine, past 6 months



(h) Total out of pocket costs for medical care, last 6 months



(i) Total amount currently owed for medical expenses

above 50 years regardless of their household level income status. The remaining sub group of age below 50 who belongs to household with more than 100% FPL enjoys about $30 to $50 reduction on the cost of the out of pocket costs for emergency room or overnight hospital. Figure 3 panel (g) shows that the lottery winners who belong to the household below 100% FPL (regardless of their age) report more than $15 reduction in the out of pocket costs for prescription medicine in past 6 months. Figure 3 panel (h) shows that the lottery winners who belong to the household below 100% FPL (regardless of their age) have more than $50 reduction in the total out of pocket cost for medical care in last 6 months.

Figure 3 panel (i) exhibits the reduction of the total amount currently owned for the medical expenses. Compare to control group, the treatment group with age of 35 to 50 have medical debt reduction, such medical debt reduction is more pronounced (more than $600) if the person belongs to the household of 50% below the FPL. The medical debt reduction is less than $400 for the younger ages (regardless of the

household income) and for age above 50 years within the household income below 100% of the FPL.

5.2 Self-reported health

Table 4 show the effectiveness of the Oregon Health Insurance Experiment in various dimension of the perceived physical and mental health outcomes after a year. The ITT and ATE are similar and positive suggesting lottery winners on average reported higher self-reported health. The LATE shows the effective is even higher for compliance subgroup. There exist detectable quantities of the treatment heterogeneity as per the DFP.

Table 4: Self-reported health

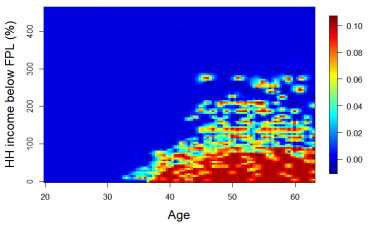
Variables	ITT	LATE	ATE	Heuristic	MFP	DFP
Self-reported health good/very good/excellent (not fair or poor)	0.046*** (0.009)	0.15*** (0.028)	0.046*** (0.009)	0.032* (0.017)	0.984*** (0.19)	1.485*** (0.431)
Self-reported health not poor (fair, good, very good, or excellent)	0.033*** (0.006)	0.107*** (0.019)	0.033*** (0.006)	0.044*** (0.012)	1.036*** (0.188)	1.085*** (0.316)
Health about the same or gotten better over last six months	0.035*** (0.008)	0.115*** (0.026)	0.039*** (0.008)	0.078*** (0.016)	1.086*** (0.223)	1.748*** (0.437)
Number of days physical health good, past 30 days	0.557*** (0.182)	1.796*** (0.587)	0.602*** (0.183)	0.431 (0.364)	1.037*** (0.312)	1.011*** (0.4)
Number days poor physical or mental health did not impair usual activity, past 30 days	0.432** (0.198)	1.397** (0.641)	0.454** (0.197)	1.333*** (0.392)	1.157** (0.511)	1.286*** (0.421)
Number of days mental health good, past 30 days	0.741*** (0.209)	2.479*** (0.675)	0.806*** (0.207)	0.807** (0.411)	1.041*** (0.27)	0.815*** (0.311)
Did not screen positive for depression, last two weeks	0.024*** (0.008)	0.079*** (0.027)	0.027*** (0.008)	0.023 (0.017)	1.055*** (0.338)	0.657 (0.81)

Notes: The ***, **, and * represent 1%, 5%, and 10% level of significance, respectively. Enclosed in the parenthesis are household-level clustered heteroscedasticity-consistent standard errors. The regressions in columns (1) and (2) include household size dummies, survey wave dummies and survey wave interacted with household size dummies. For the LATE estimates in column (2), instrumental variables is lottery assignment and endogenous variables is "Ever in Medicaid". The ITT and LATE estimates are base on the double-selection post-LASSO.

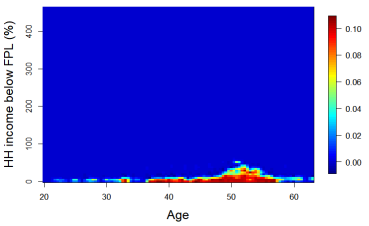
The survey has the self-reported health section. The subjects had five options to choose (Excellent, very good, good, fair and poor) to report their health for different time frames. These are ordinal questions in nature and there is no doubt that responder may have different point of reference to what does a good health represent for each individual. These options are re coded as binary for the self-reported health good/very good/excellent to 1 and not fair or poor to 0. Figure 4 panel (a) shows that the subgroup of age above 40 who are below 100% of the federal poverty line are more likely to report better health. Again, the health choice options are re coded 1 for the the self-reported health not poor (fair, good, very good, or excellent) and 0 for poor. Only the small subgroup of age around 50 reported at least not poor health as shown in Figure 4 panel (b).

Figure 4 panel (c) shows heterogeneity for another question regarding if responder perceived the

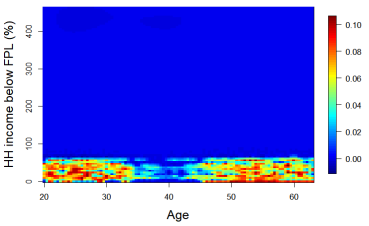
Figure 4: Self-reported health



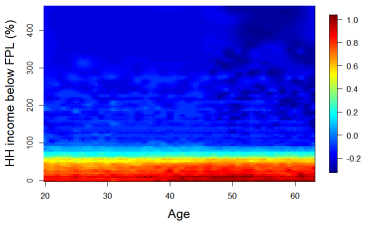
(a) Self-reported health good/very good/excellent (not fair or poor)



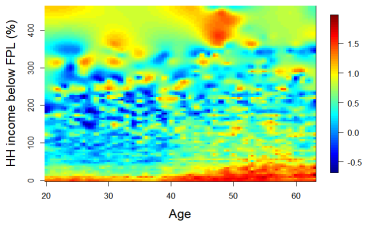
(b) Self-reported health not poor (fair, good, very good, or excellent)



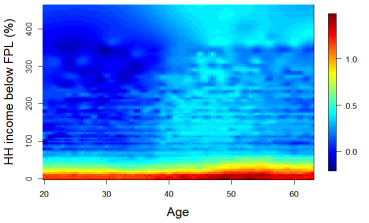
(c) Health about the same or gotten better over last six months



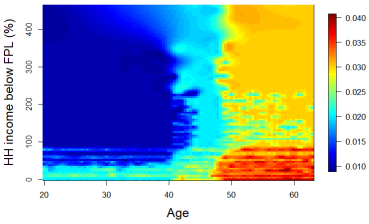
(d) Number of days physical health good, past 30 days



(e) Number of days mental health good, past 30 days



(f) Number days poor physical or mental health did not impair usual activity, past 30 days



(g) Did not screen positive for depression, last two weeks

health better or worse health in comparison to last six months, the responder from the household below 70% of the FPL expect the age group 35 to 45 report a comparatively better health.

When asked to quantify the number of good physical health days in past 30 days, lottery winning household closer to FPL report higher numbers. See Figure 4 panel (d). However, in Figure 4 panel (e), the number of good mental health days in past 30 days is reported to be higher for age group above 40 from the lottery winning household closer to FPL. The severity of mental and physical health is captured from the question to quantify the number of poor physical or mental health days did not impair usual activity, past 30 days. Again household closer to FPL report higher numbers of days that were not impaired by the poor physical and mental health. See Figure 4 panel (f). See Figure 4 panel (g) shows the age group 50 above who are from household below 100% of FPL are more likely not to be detected positive for depression (in last 2 weeks).

5.3 Potential mechanism for improved health

Table 5: Potential mechanism for improved health

Variables	ITT	LATE	ATE	Heuristic	MFP	DFP
Have usual place	0.087***	0.274***	0.086***	0.041**	1.012***	2.185***
of clinic-based care	(0.009)	(0.029)	(0.009)	(0.018)	(0.109)	(0.736)
Have personal doctor	0.073***	0.235***	0.072***	0.101***	1.031***	1.329***
	(0.009)	(0.029)	(0.009)	(0.018)	(0.127)	(0.202)
Got all needed medical	0.085***	0.274***	0.085***	0.095***	1.019***	1.985***
	(0.009)	(0.028)	(0.009)	(0.017)	(0.106)	(0.332)
Got all needed drugs,	0.07***	0.227***	0.073***	0.058***	1.016***	1.733***
last six months	(0.008)	(0.026)	(0.008)	(0.016)	(0.112)	(0.416)
Didn't use ER for n	0.00	0.00	0.003	-0.04***	1.163	-4.168
on emergency, last six months	(0.005)	(0.015)	(0.005)	(0.01)	(1.469)	(2.29)
Quality of care received last six months	0.049***	0.15***	0.053***	-0.312***	1.028***	-402.796
good/very good/excellent (conditional on any)	(0.01)	(0.03)	(0.01)	(0.019)	(0.179)	(19.252)
Happiness, very happy or pretty	0.062***	0.202***	0.069***	0.057***	1.049***	1.551***
happy (vs. not too happy)	(0.009)	(0.029)	(0.009)	(0.017)	(0.134)	(0.379)

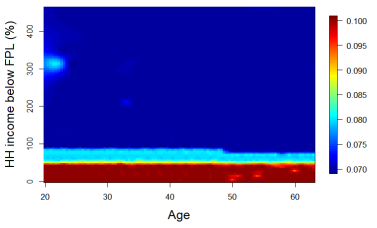
Notes: The ***, **, and * represent 1%, 5%, and 10% level of significance, respectively. Enclosed in the parenthesis are household-level clustered heteroscedasticity-consistent standard errors. The regressions in columns (1) and (2) include household size dummies, survey wave dummies and survey wave interacted with household size dummies. For the LATE estimates in column (2), instrumental variables is lottery assignment and endogenous variables is "Ever in Medicaid". The ITT and LATE estimates are base on the double-selection post-LASSO.

6 Conclusion

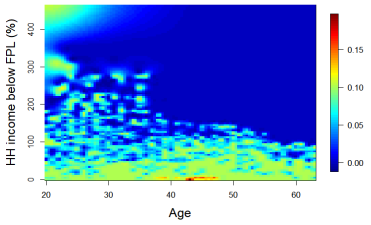
Extensive literature studying the impact of insurance coverage on health outcomes report average treatment effects. However, establishing causal effects is challenging due to endogeneity. Endogeneity arises because it is difficult to control for observed and unobserved confounding variables among the insured and uninsured population (Levy and Meltzer, 2008). For example, a comparison of the health between those with and without health insurance, can exhibit that insurance is detrimental for one's health (Baicker and Finkelstein, 2011) because people with poor health are more likely to get insurance compared to healthy people.

A Random assignment of insurance can circumvent such confounding problems (Finkelstein et al., 2012), and the Oregon Health Insurance Experiment renders a unique opportunity to evaluate the causal effects of owning health insurance (Baicker and Finkelstein, 2011) on health and personal finance-related outcomes. In early 2008, Oregon's Department of Human Services applied for and received permission from the Centers for Medicare and Medicaid Services to add new members through random lottery draws from a new reservation list (Finkelstein et al., 2012). In the year following the random assignment, the treatment group had higher health care utilization, lower out-of-pocket medical expenditures and medical

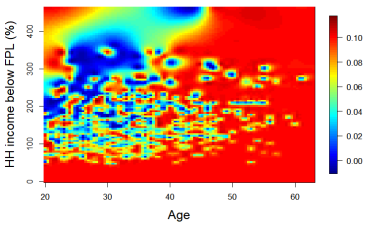
Figure 5: Potential mechanism for improved health



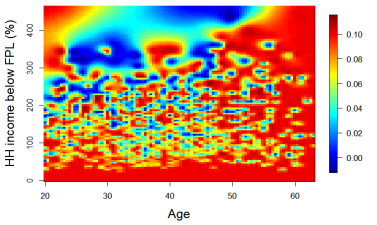
(a) Have usual place of clinic-based care



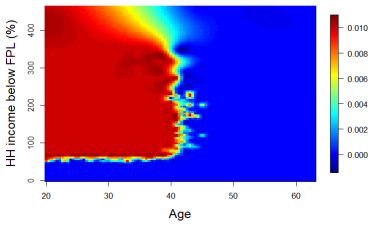
(b) Have personal doctor



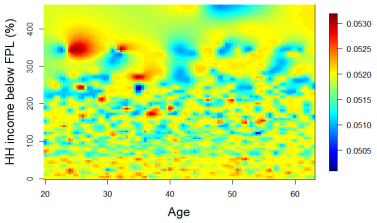
(c) Got all needed medical care, last six months



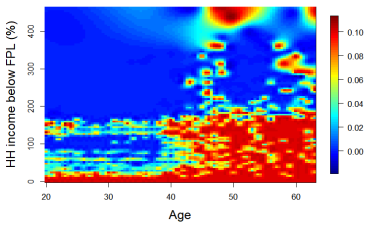
(d) Got all needed drugs, last six months



(e) Didn't use ER for non-emergency, last six months



(f) Quality of care received last six months good/very good/excellent (conditional on any)



(g) Happiness, very happy or pretty happy (vs. not too happy)

debt, and better self-reported physical and mental health than the control group, but it did not have detectable improvements in physical health conditions like high blood pressure (Finkelstein et al., 2012) – leaving policymakers with tough choices in balancing costs and benefits (Baicker, 2019).

References

Allen, H., Baicker, K., Finkelstein, A., Taubman, S., and Wright, B. J. (2010). What the Oregon Health Study can Tell us about Expanding Medicaid. *Health Affairs*, 29(8):1498–1506.

Athey, S. (2018). The Impact of Machine Learning on Economics. In *The Economics of Artificial Intelligence: An Agenda*, number Ml, pages 1–27. University of Chicago Press.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. (2017a). Chapter 3 - the econometrics of randomized experiments. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 73 – 140. North-Holland.

Athey, S. and Imbens, G. W. (2017b). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

Athey, S., Tibshirani, J., and Wager, S. (2016). Solving Heterogeneous Estimating Equations with Gradient Forests. Research Papers 3475, Stanford University, Graduate School of Business.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2018). Efficient Policy Learning. pages 1–37.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*.

Baicker, K. (2019). The effect of health insurance on spending, health, and well-being − evidence and implications for reform.

Baicker, K., Allen, H. L., Wright, B. J., and Finkelstein, A. N. (2017). The Effect Of Medicaid On Medication Use Among Poor Adults: Evidence from Oregon. *Health Affairs*, 36(12):2110–2114.

Baicker, K. and Finkelstein, A. (2011). The Effects of Medicaid Coverage Learning from the Oregon Experiment. *New England Journal of Medicine*, 365(8):683–685.

Baicker, K., Finkelstein, A., Song, J., and Taubman, S. (2014). The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment. *American Economic Review*, 104(5):322–328.

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., and Finkelstein, A. N. (2013). The Oregon Experiment Effects of Medicaid on Clinical Outcomes. *New England Journal of Medicine*, 368(18):1713–1722.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). Classification and regression trees Regression trees. *Wadsworth: Belmont, CA*, (June):358.

Brook, R. H., Ware, J. E., Rogers, W. H., Keeler, E. B., Davies, A. R., Donald, C. A., Goldberg, G. A., Lohr, K. N., Masthay, P. C., and Newhouse, J. P. (1983). Does Free Care Improve Adults' Health? *New England Journal of Medicine*.

Card, D. and Maestas, N. (2008). Care Utilization : Evidence from Medicare. *American Economic Review*, 98(5):2242–2258.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018b). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research.

Currie, J. and Gruber, J. (1996a). Health Insurance Eligibility, Utilization of Medical Care, and Child Health. *The Quarterly Journal of Economics*, 111(2):431–466.

Currie, J. and Gruber, J. (1996b). Saving Babies : The Efficacy and Cost of Recent Changes in the Medicaid Eligibility of Pregnant Women. *Journal of Political Economy*, 104(6):1263–1296.

Davis, J. M. and Heller, S. B. (2017). Using Causal Forests to Predict Treatment Heterogeneity: An application to Summer Jobs. *American Economic Review*, 107(5):546–550.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *CoRR*, abs/1103.4601.

Dudk, M., Erhan, D., Langford, J., and Li, L. (2014). Doubly robust policy evaluation and optimization. *Statist. Sci.*, 29(4):485–511.

Finkelstein, A. and McKnight, R. (2008). What Did Medicare Do? The Initial Impact of Medicare on Mortality and Out of Pocket Medical Spending. *Journal of Public Economics*.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H. L., Baicker, K., and Oregon Health Study Group, . (2012). The Oregon Health Insurance Experiment: Evidence From The First Year. *Quarterly Journal of Economics*, 127(August (3)):1057–1106.

Foden-Vencil, K. (2018). Oregon Measure 101: What You Need To Know . News — OPB.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*.

Glaeser, E. L., Hillis, A., Kominers, S. D., and Luca, M. (2016). Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. *American Economic Review*, 106(5):114–118.

Glaeser, E. L., Kominers, S. D., Luca, M., and Naik, N. (2018). Big Data and Big Cities: the Promises and Limitations of Improved Measures of Urban Life. *Economic Inquiry*, 56(1):114–137.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., and Staudt, L. M. (2016). Effect of Medicaid Coverage on ED Use Further Evidence from Oregon's Experiment. *New England Journal of Medicine*, 363(1):1–3.

Hanratty, M. J. (1996). American Economic Association Canadian National Health Insurance and Infant Health. *The American Economic Review*, 86(1):276–284.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *Source: The American Economic Review American Economic Review*, 102(1022):994–1028.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Jiang, N. and Li, L. (2015). Doubly robust off-policy evaluation for reinforcement learning. *CoRR*, abs/1511.03722.

Kaiser Family Foundation (2019). Status of state action on the medicaid expansion decision.

Kallus, N. (2018). Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8895–8906.

Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2017). Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach. Economics Working Paper Series 1711, University of St. Gallen, School of Economics and Political Science.

Lan, W., Zhong, P.-S., Li, R., Wang, H., and Tsai, C.-L. (2016). Testing a single regression coefficient in high dimensional linear models. *Journal of Econometrics*, 195(1):154 – 168.

Levy, H. and Meltzer, D. (2004). What Do We Really Know About Whether Health Insurance Affects Health? *Health policy and the uninsured*, pages 179–204.

Levy, H. and Meltzer, D. (2008). The Impact of Health Insurance on Health. *Annual Review of Public Health*, 29(1):399–409.

Li, L., Chu, W., Langford, J., Moon, T., and Wang, X. (2012). An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In Glowacka, D., Dorard, L., and Shawe-Taylor, J., editors, *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pages 19–36, Bellevue, Washington, USA. PMLR.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*.

Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

McWilliams, J. M., Meara, E., Zaslavsky, A. M., and Ayanian, J. Z. (2007a). Health of Previously Uninsured Adults after Acquiring Medicare Coverage. *JAMA - Journal of the American Medical Association*.

McWilliams, J. M., Meara, E., Zaslavsky, A. M., and Ayanian, J. Z. (2007b). Use of Health Services by Previously Uninsured Medicare Beneficiaries. *New England Journal of Medicine*.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic PerspectivesVolume*, 31(2Spring):87–106.

Naik, N., Raskar, R., and Hidalgo, C. A. (2016). Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance. *American Economic Review*, 106(5):128–132.

Newhouse, J. P. (1994). Free for All: Lessons from the RAND Health Insurance Experiment. *BMJ*.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

Norris, L. (2018). Oregon and the ACA's Medicaid expansion: eligibility, enrollment and benefits — healthinsurance.org.

Office for Oregon Health Policy and Research (2009). Trends in oregons healthcare market and the oregon health plan: A report to the 75th legislative assembly.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Strehl, A. L., Langford, J., and Kakade, S. M. (2010). Learning from logged implicit exploration data. *CoRR*, abs/1003.0120.

Swaminathan, A. and Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755.

Taubman, S. L., Allen, H. L., Wright, B. J., and Baicker, K. (2014). Oregon's Health Insurance Experiment. *Science*, 343(6168):263–268.

Thomas, P. S. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. *CoRR*, abs/1604.00923.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wallace, N. T., McConnell, K. J., Gallia, C. A., and Smith, J. A. (2008). How effective are copayments in reducing expenditures for low-income adult medicaid beneficiaries? Experience from the Oregon Health Plan. *Health Services Research*, 43.

Xie, Y., Brand, J. E., and Jann, B. (2012). Estimating Heterogeneous Treatment Effects with Observational Data. *Sociological methodology*, 42(1):314–347.

Zhou, R. A., Baicker, K., Taubman, S., and Finkelstein, A. N. (2017). The uninsured do not use the emergency department more- they use other care less. *Health Affairs*, 36(12):2115–2122.

A Variable importance

Table 6: Variable importance

Variables	FPL	Age	HHS	INS	Other variables
Currently taking any prescription medications	✓	✓	✓		% MSA
Outpatient visits last six months	✓	✓		✓	
ER visits last six months	✓	✓	✓		
Inpatient hospital admissions last six months	✓	✓	✓	✓	
Number of prescription medications currently taking	✓	✓	✓	✓	
Number of Outpatient visits last six months	✓	✓		✓	
Number of ER visits last six months	✓	✓	✓	✓	% High school disploma or GED
Number Inpatient hospital admissions last six months	✓	✓	✓		% Self signup
Any out of pocket medical expenses, last six months	✓	✓	✓	✓	% MSA
Owe money for medical expenses currently	✓	✓	✓		
Borrowed money or skipped other bills to pay medical bills, last six months	✓	✓	✓		
Refused treatment because of medical debt, last six months	✓	✓	✓		
Out of pocket costs for doctors visits, clinics or health centers, past 6 months	✓	✓			% work 30+ hrs/week
Out of pocket costs for emergency room or overnight hospital care, past 6 months	✓	✓	✓		
Out of pocket costs for prescription medicine, past 6 months	✓	✓	✓	✓	
Out of pocket costs for other medical care, past 6 months	✓	✓		✓	
Total out of pocket costs for medical care, last 6 months	✓	✓	✓	✓	% work 30+ hrs/week
Total amount currently owed for medical expenses	✓	✓	✓	✓	
Have usual place of clinic-based care	✓	✓			
Have personal doctor	✓	✓		✓	% work 30+ hrs/week
Got all needed medical care, last six months	✓	✓	✓	✓	% work 30+ hrs/week
Got all needed drugs, last six months	✓	✓	✓		% dont currently work
Didn't use ER for non emergency, last six months	✓	✓	✓		% work 30+ hrs/week
Quality of care received last six months good/very good/excellent (conditional on any)	✓	✓	✓		% MSA
Happiness, very happy or pretty happy (vs. not too happy)	✓	✓	✓	✓	
Blood cholesterol checked (ever)	✓	✓	✓	✓	
Blood tested for high blood sugar/diabetes (ever)	✓	✓	✓		
Mammogram within last 12 months (women 40)	✓	✓	✓	✓	% work 30+ hrs/week
Pap test within last 12 months (women)	✓	✓	✓		% work 30+ hrs/week
Self-reported health good/very good/excellent (not fair or poor)	✓	✓	✓	✓	
Self-reported health not poor (fair, good, very good, or excellent)	✓	✓		✓	
Health about the same or gotten better over last six months	✓	✓	✓		% High school diploma or GED
Number of days physical health good, past 30 days	✓	✓	✓		
Number days poor physical or mental health did not impair usual activity, past 30 days	✓	✓	✓	✓	
Number of days mental health good, past 30 days	✓	✓	✓	✓	% Female
Did not screen positive for depression, last two weeks	✓	✓	✓		

Notes: FPL represents household below the federal poverty line (in %), HHS represents household size, INS represents the nummber of non insurance months in last six months. The random forest model always splits on FPL and Age along with HHS and INS. Along with these variables the random forest also splits on different variables included in the last column. For example, consider the model called "Currently taking any prescription medications", the random forest splits (more than average) the data on FPL, Age, HHS and % MSA. Therefore, the treatment heterogeneity are likely within these variables.

B Causal machine learning approaches

The observational studies, quasi-experimental studies, and randomized experiments often focus on causal inference and have been dominant in empirical policy research in health economics and economics in general. However, recently, due to availability of big-data and computing powers, machine learning approaches are gaining momentum among the researchers and policymakers. Varian (2014), Mullainathan and Spiess (2017) and Athey (2018) brief the utilities of the big-data and machine learning method in the field of economics. Within the domain of machine learning in economics, two strands of literature are gaining momentum: machine learning for policy prediction problems; and machine learning for causal inference problems.

The machine learning algorithms behave well for out-of-sample prediction. Therefore these algorithms are useful in many policy applications where the causal inference is not central or maybe not necessary. For example, Kleinberg et al. (2015) consider a resource allocation problem in health policy in which a policymaker needs to decide which otherwise-eligible patients should not be given hip replacement surgery through Medicare. They predict the probability that a candidate for a joint replacement would die within a year from other causes. Then they identify patients who are at particularly high risk and should not receive joint replacement surgery. Similarly, Henderson et al. (2012) use satellite data on night lights to predict economic growth; Glaeser et al. (2018) use Google Street View images to predict income in New York City; Glaeser et al. (2016) develop system for allocating health inspectors to restaurants in Boston; and Naik et al. (2016) quantify urban appearance from street-level imagery for 19 American cities and establish an empirical connection between the physical appearance of a city and the behavior and health of its inhabitants.

The machine learning algorithms are not well suited for causal inference (Athey, 2018) because establishing causal effect relates to understanding the counterfactual− what would happen with and without a policy− rather than just correctly predicting out-of-sample. However, some slight modifications of "off-the-shelf" or readily-available[14] machine learning algorithm can utilize the strengths and innovations of machine learning algorithms for causal inference. The approaches that utilize machine learning methods for causal inference are termed as causal machine learning. These methods usually focus on estimating average treatment effect, heterogeneous treatment effects, and optimal policies in the environment which are otherwise infeasible to estimate using standard econometric approaches.

B.1 Average treatment effect

In this paragraph, I show a few examples of causal machine learning approach to estimate the average treatment effect. For example, Belloni et al. (2014b) and Belloni et al. (2014a) utilize "off-the-shelf" or readily available predictive machine learning algorithm called the "LASSO"[15] method and purpose a

[14]The predictive machine learning algorithms are easily available with the open-source routines for the statistical software like Python and R.

[15]The Least Absolute Shrinkage and Selection Operator (LASSO) is an appealing method to estimate the sparse parameter from a high-dimensional linear model is introduce by Frank and Friedman (1993) and Tibshirani (1996). The LASSO simultaneously performs model selection and coefficient estimation by minimizing the sum of squared residuals plus a penalty term. The penalty term penalizes the size of the model through the sum of absolute values of coefficients. Consider a following linear model $\tilde{y}_i = \Theta_i\beta_1 + \varepsilon_i$, where Θ is high-dimensional covariates, the LASSO estimator is defined as the solution to $\min_{\beta_1 \in \mathbb{R}^p} E_n\left[(\tilde{y}_i - {}_i\Theta\beta_1)^2\right] + \frac{\lambda}{n}\|\beta_1\|_1$, the penalty level λ is a tuning parameter to regularize/controls the degree of penalization and to guard against over-fitting. The cross-validation technique chooses the best λ in prediction models and $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. The kinked nature of penalty function induces $\hat{\beta}$ to have many zeros; thus LASSO solution feasible for model selection.

correction[16] called the "double-selection post-LASSO"[17] method. This method is useful for estimating average treatment effect when the analyst is required to select a "sparse" outcome model[18] from high-dimensional observables when some covariates correlate with treatment and outcome, and the analyst does not know which ones are important. Similarly, Athey et al. (2018) utilize "doubly-robust"[19] method and LASSO method and purpose "residual balancing"[20] approach for estimating average treatment effect under the assumption of unconfoundedness[21] and the assumption of the outcome model is linear and sparse. Similarly, Chernozhukov et al. (2018a) purpose "double machine learning" for estimating the average treatment effect under unconfoundedness. The idea is to first run any feasible machine learning methods of outcomes on covariates, and then second run another feasible machine learning methods of the treatment indicator on covariates; then, the residuals from the first machine learning are regressed on the residuals from the second machine learning to estimate the average treatment effect. This idea is similar to Frish-Waugh-Lovell theorem[22] and close to the concept of Robinson (1988) residual-on-residual regression approaches where the estimator was a kernel regression.

B.1.1 Heterogeneous treatment effects

Along with the average treatment effect, heterogeneous treatment effects estimation interests policy-makers because it helps to quantify the sizes of effects on different subpopulations, which is valuable to improve program targeting and to understand the underlying mechanisms driving the results. Usually, data are stratified in mutually exclusive groups or include interactions in a regression to explore heterogeneous treatment effects. However, ad-hoc searches for the responsive subgroups may lead to false

[16] When LASSO of outcome variable is implemented to select the covariates while always restricting the treatment indicator, the estimated treatment effect is biased because LASSO's sole objective is to select variables that predict outcome thus LASSO fails to select confounders that are also strong predictor of treatment assignment.

[17] Belloni et al. (2014a) simplify the double-selection post-LASSO procedure as following. First, run LASSO of outcome variables on a large list of potential covariates to select a set of predictors for the outcome variable. Second, run LASSO of treatment variable on a large list of potential covariates to select a set of predictors for treatment. If the treatment is truly exogenous, we should expect this second step should not select any variables. Third, run OLS regression of outcome variable on treatment variable, and the union of the sets of regressors selected in the two LASSO runs to estimate the effect of treatment on the outcome variable then correct the inference with usual heteroscedasticity robust OLS standard error.

[18] The "sparse" outcome model means a model with a few meaningful covariates affect the average outcome. These few meaningful covariates are selected from a given list of many observables covariates, and potentially a situate when numbers of observables k are greater than numbers of observations n, i.e., $k > n$. When $k > n$ an estimation based on the least-squares estimation is infeasible. However, traditionally, the principal component analysis (PCA) is commonly used to reduce dimension when the likelihood function is normal. The PCA creates principal components using linear combinations of a much larger set of variables from a multivariate data-set. Interpreting the coefficients on the principal components requires the researcher first to interpret the principal components, which can prove a challenge as all variables have non-zero loadings.

[19] The "doubly-robust" estimator proceeds by taking the average of the efficient score, which involves the estimation of conditional mean of outcomes given covariates as well as the inverse propensity score Athey (2018).

[20] The "residual balancing" replaces inverse propensity score weights with weights obtained using quadratic programming, where the weights are designed to achieve balance between the treatment and control group. The conditional mean of outcomes is estimated using LASSO Athey (2018).

[21] The unconfoundedness assumption implies treatment is randomly assigned and knowing observable characteristics of an individual, and their treatment status gives no additional information on the potential outcomes. This means the treatment assignment is independent of the outcome variable.

[22] The Frisch-Waugh-Lovell theorem is that estimating a parameter in a multiple regression is equivalent to estimating the same parameter in a simple regression of the residual of the regress and regressed on all other predictors on the residual of the regressor regressed on all other predictors.

discoveries or may mistake noise for a true treatment effect (Davis and Heller, 2017). Knaus et al. (2017) point out that for large-scale investigations of effect heterogeneity, standard $p-$values of standard (single) hypothesis tests are no longer valid because of the multiple hypothesis testing problems (Lan et al., 2016; List et al., 2019) and leads to so-called "ex-post selection" problem which is widely recognized in the program evaluation literature. For example, for fifty single hypotheses tests, the probability that at least one test falsely rejects the null hypotheses at the 5% significance level (assuming independent test statistics as an extreme case) is $1 - 0.95^{50} = 0.92$ or 92%.

The recent avenue of causal machine learning provides a better systematic approach to search the groups with heterogeneous treatment effects. One intuitive approach proposed by Imai and Ratkovic (2013) is to sample-split and use the first sample to run LASSO regression model with the treatment indicator interacted with covariates and perform variables selections then use the selected model with the second sample to perform an ordinary least squares regression to guard against over-fitting. While Athey and Imbens (2016) utilize the Breiman et al. (1984) classification and regression tree (CART)[23] machine learning algorithms and purpose "causal tree" method. The CART recursively filters and partitions the large data-set into binary sub-groups (nodes) such that the samples within each subset become more homogeneous that fit the response variable. Unlike the CART that minimizes the mean-squared error of the prediction of outcomes to capture heterogeneity in outcomes, the "causal" tree minimizes the mean-squared error of treatment effects to capture treatment effect heterogeneity. The approach to estimate the "causal" tree is similar to Imai and Ratkovic (2013) approach, in which half of the sample is used to determine the optimal partition of covariates space, while the other half is used to estimate treatment effects within the leave based on the optimal partition of covariates selected from the first partition (Athey and Imbens, 2016). The sample-splitting approach also known as "honest" estimation lead to loss of precision as only half of the data is used to estimate the effect, but generates a treatment effect and a confidence interval for each subgroup that is valid no matter how many covariates are used in estimation. Athey and Imbens (2017b) points out that the researcher is free to estimate a more complex model in the second part of the data, for example, if the researcher wishes to include fixed effects in the model, or model different types of correlation in the error structure.

The causal tree doesn't provide personalized estimates, Wager and Athey (2018) utilize "random forest" machine learning approach and propose a "causal forest" method, where many different causal trees are generated and averaged. This method provides causal effects that change more smoothly with covariates and provides distinct individualized estimates and confidence intervals. Wager and Athey

[23] In simplest, the CART algorithm chooses a variable and split that variable above or below a certain level (which forms two mutually exclusive subgroups or leaves) such that the sum of squared residuals is minimized. This splitting process is repeated for each leave until the reduction in the sum of squared residuals is below a certain level as defined by users, thus resulting a tree format (Athey and Imbens, 2017b).

(2018) also show that the predictions from causal forests are asymptotically normal and centered on the true conditional average treatment effect for each individual. Athey et al. (2016) extend the approach to other models for causal effects, such as instrumental variables, or other models that can be estimated using the generalized method of moments (GMM). In each case, the goal is to estimate how a causal parameter of interest varies with covariates.

B.1.2 Optimal policy estimation

The optimal policy estimation have received greater attention in the machine learning literature[24] (Athey, 2018). The optimal policy function map the observable characteristics of an individual to a policy or treatment assignment. In simplest, the main goal of optimal policy estimation is to answer—- "who should be treated?" or optimal treatment allocation. The understanding of optimal policy is essential in policymaking because an ad-hoc targeting a specific subpopulation with positive interventions can be unfair, unethical, illegal and unpolitical to some other subpopulations while intervening everyone in the population (a blanket policy) is welfare-maximizing but can be extremely costly.

The optimal policy estimation or optimal treatment allocation, has been recently studied in using causal machine learning in economics, mainly by Kitagawa and Tetenov (2018) and **?**. The main idea is to select a policy function that minimizes the loss from failing to use the ideal policy, referred to as the regret of the policy. Note that estimating conditional average treatment effect or heterogeneous treatment effect focus on the squared-error loss while the optimal policy estimation focuses on utilitarian regret **?**.

[24] See Strehl et al. (2010); Dudík et al. (2011); Li et al. (2012); Dudk et al. (2014); Swaminathan and Joachims (2015); Jiang and Li (2015); Thomas and Brunskill (2016) and Kallus (2018).