# Predicting Disability Enrollment Using Machine Learning[*]

Timothy J. Layton[†] Helge Liebert[‡] Nicole Maestas[§] Daniel Prinz[¶]
Boris Vabson[‖]

November 21, 2019

**Abstract**

We use data on enrollment in the Supplemental Security Income (SSI) and Social Security Disability Insurance (SSDI) program and data on health care spending by Medicaid beneficiaries to analyze the extent to which Medicaid spending is predictive of future disability insurance receipt among non-disabled teenagers and future disability insurance disenrollment among disabled teenagers. In our first set of analyses, we find that we currently do not have enough data to predict future SSI and SSDI enrollment among non-disabled teenagers. In our second set of analyses, we find that observed Medicaid spending among disabled teenagers can be used to predict SSI disenrollment. Our results indicate that machine learning models using information on healthcare spending may be useful for identifying current teenage SSI recipients who are more or less likely to be removed from SSI.

[*] We thank Julia Yates and Elena Stacy for excellent research assistance. We gratefully acknowledge financial support from the Social Security Administration through grant DRC12000002-06 to the National Bureau of Economic Research as part of the SSA Disability Research Consortium, and the Agency for Healthcare Research and Quality (K01-HS25786-01). The findings and conclusions expressed are solely those of the authors and do not represent the opinions or policy of SSA, any agency of the Federal Government, or the National Bureau of Economic Research.

[†] Harvard University and NBER. Email: layton@hcp.med.harvard.edu

[‡] Harvard University. Email: liebert@hcp.med.harvard.edu

[§] Harvard University and NBER. Email: maestas@hcp.med.harvard.edu

[¶] Harvard University. Email: dprinz@g.harvard.edu

[‖] Harvard University. Email: vabson@hcp.med.harvard.edu

# 1 Introduction

Despite recent declines, enrollment levels as well as total spending in U.S. disability insurance (DI) programs have risen in recent decades. The Social Security Disability Insurance (SSDI) program had 11.7 million beneficiaries in 2017 (Social Security Administration, 2018a), two and half times more beneficiaries than 30 years earlier. Similarly, the number of beneficiaries in the Supplemental Security Income (SSI) program doubled during the same period (Social Security Administration, 2018b). With almost \$200 billion in total federal spending on cash benefits (Social Security Administration, 2018a,b) and an additional \$270 billion in total federal spending on health insurance benefits (Kaiser Family Foundation, 2014; Congressional Budget Office, 2016), these programs account for 2.5% of GDP.

However, beyond their fiscal importance, DI programs impact the lives of millions of Americans in a variety of ways. Recent research has provided evidence on the effects of these programs on beneficiaries, including effects on labor supply and earnings (Maestas, Mullen and Strand, 2013; French and Song, 2014; Moore, 2015; Deshpande, 2016*a*; Gelber, Moore and Strand, 2017), financial well-being (Deshpande, Gross and Su, 2019), and health (Gelber, Moore and Strand, 2018). Disability programs have also been shown to impact entire households beyond individual recipients (Duggan and Kearney, 2007; Dahl, Kostol and Mogstad, 2014; Deshpande, 2016*b*; Autor et al., 2019). Despite extensive research into the effects of enrollment in these programs, we still know relatively little about the factors that lead beneficiaries to enroll, with existing research on determinants of program take-up largely limited to economic factors (Autor and Duggan, 2003; Maestas, Mullen and Strand, 2015, 2018). While the extent to which overall changes in health and aging have impacted enrollment is debated in the literature (Autor and Duggan, 2006; Autor, 2011; Liebman, 2015), evidence on the importance of non-economic factors is scarce.[1]

In this paper, we take a first step toward understanding the role of an understudied but potentially important factor in the process determining enrollment in DI programs: healthcare. Many disabling conditions can be prevented or managed given appropriate health care treatment. For example, mental conditions such as severe depression could influence labor force participation, but at the same time could be treatable via psychotherapy and/or pharmaceutical therapies such as antidepressants and anti-psychotics. Heart attacks that may inhibit workers from returning to certain

[1] Exceptions include (Cutler, Meara and Richards-Shubik, 2011) who tie disability enrollment to health shocks and recent work by Park and Powell (2019) tying enrollment in SSI and SSDI to the reformulation of OxyContin and the subsequent shift from prescription opioids to heroin and fentanyl.

jobs may be prevented by pharmaceuticals such as statins. Physical therapy may lessen the debilitating effects of back or knee injuries.

To make progress on this question, we leverage rich administrative data including information on both enrollment in disability programs and utilization of healthcare. Specifically, we use data from state Medicaid programs that provides information on each Medicaid beneficiary's disability status, and whether they qualified for Medicaid on the basis of disability or an alternative channel. Medicaid eligibility due to disability typically implies that a beneficiary is simultaneously in the SSI program. These data also include each beneficiary's full set of health insurance claims, providing a complete view of each individual's healthcare utilization. We link these data at the individual level to enrollment data from the Medi*care* program. Consistent with the Medicaid data, the Medicare enrollment data also tracks whether a beneficiary is enrolled in Medicare due to disability as opposed to another reason. Medicare enrollment due to disability definitively implies concurrent enrollment in the SSDI program.

Using this data, we focus on teenagers enrolled in the Medicaid program in the early 2000's. We look separately at teenagers who are eligible for Medicaid due to disability vs. some other eligibility pathway. For both groups, we first identify whether they are enrolled in SSDI anytime prior to 2010, at most 13 years after we observe their healthcare utilization in Medicaid. We find that almost no individual in our sample is enrolled in SSDI at that time, implying that there is little possibility of estimating the role of teenage health care use in determining SSDI enrollment as of age 23. We then focus on individuals who were at one time enrolled in Medicaid but not disabled, and then examine whether they enroll in Medicaid on the basis of disability (and SSI) at a later date, prior to 2010. Again, we find that very few of these individuals enrolled in SSI by that time, implying that there again is little possibility of estimating the role of health care use in determining SSI enrollment for this group.[2] We then turn to the group of teenagers who are originally eligible for Medicaid due to disability. For this group, we ask a different question: Did they leave SSI at age 18? We find that a substantial portion leave, but a substantial portion stay, providing a promising setting for assessing the effects of healthcare on SSI *dis*enrollment. As a start toward this causal question, we perform an exploratory analysis of the association between healthcare utilization and disenrollment at age 18.

[2] In both cases, it is likely that age 23 is not old enough for much enrollment in SSI or SSDI to have occurred. This data limitation also rules out analyses of the effects of specific treatments, such as post-surgical opioids, on later-life DI program enrollment. Future work should focus on these important questions once the data become available.

We first show that, intuitively, teenagers with higher healthcare spending have a lower probability of exiting the SSI program. Nearly half (48%) of teenagers in the lowest quartile of Medicaid spending exit SSI, while only 23% in the highest spending quartile do, a 25 percentage point difference. We then use a series of machine learning algorithms to assess the extent to which healthcare utilization predicts disenrollment from SSI. The models correctly predict whether an individual will be disenrolled from SSI about 80% of the time. They perform better for individuals who stay on SSI vs. individuals who are disenrolled. We show that overall the actual disenrollment rate for the top quartile of predicted disenrollment is 71-82 percentage points higher than the actual disenrollment rate for the bottom quartile of predicted disenrollment. Indeed, individuals in the bottom quartile of predicted disenrollment (top quartile of predicted enrollment in SSI at age 23) have an actual disenrollment rate of just 1% (rate of SSI enrollment at age 23 of 99%). We then condition on all non-healthcare-related variables in the prediction models (demographics, time, geography) and assess the extent to which actual disenrollment differs across quartiles of conditional predicted disenrollment, finding that the disenrollment rate for the top quartile of conditional predicted disenrollment is 38-54% higher than the disenrollment rate for the bottom quartile of conditional predicted disenrollment, implying that information about health care spending during teenage years is predictive of disenrollment at age 18.

Our results suggest that observed healthcare spending (as well as other information found in Medicaid data) can be used to predict future disability insurance receipt status among current disability insurance recipients. Our results also indicate that machine learning models using information on predicted healthcare spending may be useful for identifying current teenage SSI recipients who are more or less likely to be removed from the program. In theory, one possible use of these models could be to target redetermination efforts to the groups of recipients most likely to be removed, possibly saving resources by putting off redeterminations of the groups least likely to be removed. Indeed, the models allow for easy and transparent identification of *ex ante* identifiable groups of individuals for whom the *ex post* probability of disenrollment is a mere 1%, indicating that redetermination for large groups of teenage SSI recipients could be avoided with almost no real effect on who is actually removed from the program. These models may also be useful for helping recipients understand the likelihood they will be removed at age 18, helping those who are likely to be removed to more accurately plan for their future.

This paper contributes to the literature on take-up of disability insurance programs. Most of this literature has focused on the economic sources of disability insurance take-up. For example, in recent work (Maestas, Mullen and Strand, 2015, 2018) show that when economic conditions are worse, workers are more likely to apply for and be awarded SSDI payments. Other existing works considers a variety of factors, including aging, demographic change, and program rules (Autor and Duggan, 2006; Autor, 2011; Liebman, 2015). We complement this literature by performing an exploratory descriptive analysis of the role of healthcare in disability insurance take-up.

The paper proceeds as follows. Section 2 provides background on SSDI and SSI and age-18 redetermination in the SSI program. Section 3 describes the Medicaid data we use. Section 4 describes our machine learning methods. Section 5 summarizes our empirical findings. Section 6 concludes.

## 2 Background

In the U.S., there are two major federal disability insurance programs: the Social Security Disability Insurance (SSDI) program and the Supplemental Security Income (SSI) program. Both of these programs provide monthly cash benefits to individuals classified as disabled, where disability is federally defined as the "inability to engage in substantial gainful activity (SGA) by reason of any medically determinable physical or mental impairment(s) which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months." SSDI benefits are awarded to individuals with disabilities who have sufficient work histories, as well as certain dependent family members. SSI benefits are awarded to individuals with disabilities who have limited income and resources. The SSI program covers low-income children with disabilities under age 18.

When children enrolled in the SSI program turn 18, they must go through a redetermination process to remain eligible for SSI payments. While, under age 18 the income and assets of the parents are used to determine eligibility and benefit amounts, after age 18, the child's own income and assets are considered. Recipients also go through a medical review and they must qualify under the medical eligibility criteria for adults, which differ from the eligibility criteria for children.[3] Specifically, the

[3] Separately, disabled children may become eligible for the SSDI program at age 18, based on the earnings record of a parent in the event the parent is already receiving old age/disability SS benefits or has died. For SSDI, children at 18 are also subject to the stricter adult medical criteria, identical to requirements for SSI. In the event that children qualify for sufficiently high SSDI benefits, they may no longer qualify for SSI.

disability criteria for adults is defined as an inability to work, while for children disability is defined as "marked and severe functional limitation."

Because of the difference between the definition of disability for children vs. adults, approximately 40 percent of individuals who received SSI as children are removed from the program during this redetermination. Approximately three-quarters of program removals following redetermination come from failure to satisfy the new adult medical criteria. For example, two-thirds of children with mental health conditions (other than intellectual disability) and muskuloskeletal disabilities are denied benefits during the initial redetermination. SSI children who first became eligible at a later age are more likely to get an initial cessation decision: 23% of those whose initial eligibility was at an age younger than 5 are get denied benefits vs. 49% of those with initial eligibility between ages 13 and 17. (Hemmeter and Gilby, 2009; Deshpande, 2016*a*).

The age 18 redetermination process is time-consuming, expensive, and mandatory. It may be the case that the large number of age 18 redeterminations take SSA resources away from conducting redeterminations for other groups such as adult SSDI recipients and children under age 18. It is possible that targeting of age-18 redetermination resources to the children most likely to be disenrolled, while subjecting others to lower-cost, simpler reviews would optimize resource use while also better serving program recipients.

While the disabling condition and the age of initial eligibility are clearly predictive for which children will ultimately be disenrolled from the program at age 18, information about health care utilization may also have predictive power. If it does, this information could be used to target redetermination resources toward the children most likely to be disenrolled from the program, rather than toward children who have little chance of being removed.

## 3 Data and Sample

We use several administrative datasets from the Centers for Medicare and Medicaid Services (CMS) for the states of Texas and New York, and covering both the Medicaid and Medicare programs. Administrative data pertaining to the Medicaid program extend from 1999 to 2010, while those pertaining to Medicare extend from 1999 to 2015. These data longitudinally track beneficiaries' program enrollment status, disability and SSI/SSDI program status, medical care utilization and spending, as well as prescription drug use. Uniquely, the data allow us to track individuals' enrollment and utilization over time not just for those staying in Medicaid or Medicare, but also those beneficiaries transitioning from one program to another.

Using these data, we can precisely identify the cohorts of interest in our analyses. Specifically, we restrict our analysis samples to teenagers enrolled in either Texas or New York Medicaid, for a minimum of 9 continuous months, while they were between the ages of 13 and 18. We further restrict to individuals who would have turned 23 by 2010 (the final year for which Medicaid data is available that can be linked to Medicare data due to issues with the Medicaid-Medicare crosswalk in later years), so that we have an extended view of Medicaid utilization and SSI status for every individual in our sample. For the set of analyses focused on SSI attrition among previously enrolled teenagers enrolled in SSI, we further restrict just to those who were also on SSI between the ages of 13 and 18.

### 3.1 Beneficiary Characteristics and Enrollment Information

Leveraging a combination of Medicaid and Medicare administrative data, we collect information on beneficiary characteristics, Medicaid and Medicare program enrollment status, as well as their SSI/SSDI status. For Medicaid beneficiaries, we obtain this information from the CMS Medicaid Analytic eXtract (MAX) Personal Summary (PS) files, while for Medicare beneficiaries we obtain it from CMS's Medicare Beneficiary Summary Files (MBSF). These files track each individual's Medicaid and Medicare enrollment status at a monthly frequency. These files also track the basis for each beneficiary's ongoing eligibility for Medicaid or Medicare, such as through qualification for SSI, SSDI, or alternative channels. Given that enrollment in Medicaid is automatic for SSI beneficiaries in New York and Texas, these states' Medicaid enrollment files can be used to track everyone in that state who receives SSI, based on the disability indicator contained in those files. Meanwhile, given that Medicare enrollment is automatic for SSDI beneficiaries following a 2 year waiting period, Medicare enrollment files and the disability indicator contained in them can likewise be used to track everyone receiving SSDI in these states.

The key outcome variables for this study of SSI and SSDI enrollment status can be derived directly from these files. To identify the impact of childhood health care utilization under Medicaid on subsequent SSDI takeup, Medicaid and Medicare administrative data need to be linked. For linking Medicare and Medicaid (claims and enrollment) files, we employ a special crosswalk obtained from CMS that tracks the Medicare-specific beneficiary ID corresponding to each Medicaid-specific beneficiary ID. This crosswalk allows us to identify childhood Medicaid enrollees taking up SSDI and Medicare later in life and to examine the childhood medical and demographic factors that were predictive of later take-up.

### 3.2 Medical Utilization and Expenditures Data

We track childhood medical utilization, expenditures, and chronic condition presence using claims data from Medicaid. These claims data cover the full range of inpatient as well as outpatient care, and are contained in the MAX inpatient (IP) and other therapy (OT) files. These data track claims paid by the public Medicaid program as well as those paid by private Medicaid plans. We use payment amount information in claims to construct annual measures of expenditures for different service types, such as inpatient and outpatient, and diagnosis information to construct beneficiary-level indicators for different chronic conditions. We then examine these measures as potential predictors of disability related outcomes, both pertaining to attrition from SSI as well as to possible future takeup of SSDI.

### 3.3 Prescription Drug Utilization and Expenditures Data

We track childhood prescription drug utilization and expenditures using administrative claims data from Medicaid, covering both New York and Texas. This prescription drug data covers beneficiaries under public as well as private Medicaid. The drug data tracks each prescription's cost, the date when the prescription was filled, the days' supply associated with the fill, and the drug identifier (NDC code), which we link to external data in order to group drugs by therapeutic class. This data thereby enables us to examine overall prescription drug use as well as drug-specific use as potential predictors of disability-related outcomes.

### 3.4 Sample

To construct our main analytic sample, we limit to Texas and New York Medicaid enrollees who we observe in our data between ages 13 and 18 and at age 23. In practice, this means that we limit to individuals who turn 23 by 2010 and who are 13 to 18 between 2000 and 2005. We use health care utilization information from their oldest year that keeps them in our baseline teenager sample and limit to enrollees who are in Medicaid for at least 9 months in this baseline year.

## 4 Methods

Our objective is to predict SSI disenrollment by age 23 for children ages 18 and below who are currently in SSI using the data available in the Medicaid records. We aim to build a model that performs well in predicting disenrollment on out-of-sample data. To do so, we train and compare three different models for binary classification, $\ell_1$-penalized logistic regression, random forests and gradient boosted trees. To compare the performance of these models, we initially reserve 20% of our data as a test set. This data is randomly selected and never used to train any of the models. This guarantees an unbiased comparison of the predictive performance of the different models. To avoid overfitting and guarantee good out-of-sample performance, we use 10-fold cross-validation to train the models and select the optimal tuning parameters. We split the remaining data in 10 equal sized folds, estimate our model on 9 of these and predict its performance on the left-out validation fold. We then repeat this for all possible splits, and average the classification error rate on the validation sample to get a good measure of out-of-sample performance.

This process is repeated for all combinations of tuning parameters on a chosen model-specific grid. When training the models, we always use classification accuracy as our performance metric. Using AUC or other alternative performance metrics leads to very similar final models. We then pick the best models in each class and compare their performance on the 20% hold-out sample. Finally, we re-estimate the models using the optimal tuning parameters on the full sample to obtain predictions and predicted probabilities for all individuals in the data.

We choose the three different machine learning classifiers due to their popularity in applications, and because from a modelling perspective, they span the range of modelling options. Logistic Lasso ($\ell_1$-penalized logistic regression; Tibshirani, 1996; Hastie, Tibshirani and Friedman, 2001) is conceptually similar to a simple linear index model and classical regression. Lasso additionally applies a penalty to the absolute size of the coefficients, shrinking them. Since the penalty function is based on the $\ell_1$ norm, some coefficients are shrunk exactly to zero, leading to a more parsimonious model. To achieve good performance for this type of model, we include as model inputs additional non-linear transformations of the base input variables. Specifically, we also include up to a third-order polynomial of all continuous variables, and all second-order interactions of these and the original variables. The tuning parameter for penalized regression is the penalty factor $\lambda$.

The other two methods we consider are ensemble methods that combine a set of base learners.

Random forests rely on the ensemble algorithm *bagging* (bootstrap aggregating, Breiman, 1996, 2001). In bagging, a base learner is fit on a with-replacement bootstrap sample of the original sample. This process is repeated multiple times, and the predictions of the base learner across the different bootstrap samples are then aggregated. Typically decision trees are chosen as base learners (Breiman et al., 1984; Hastie, Tibshirani and Friedman, 2001). Since individual tree estimators tend to overfit, averaging their predictions substantially reduces variance at a negligible cost of bias. Random forests are essentially bagged trees, with the additional modification that at each split point in the tree, only a random subset of the available predictor variables is considered for partitioning the sample. The size of this random subset is the major tuning parameter for random forests. We set the number of bootstrap resamples to 1,000, i.e. our forests consist of 1,000 individual trees, as this appears to stabilize estimates sufficiently.

The other ensemble method we apply is gradient *boosting* (Friedman, 2001, 2002; Hastie, Tibshirani and Friedman, 2001). In contrast to bagging, which aggregates a series of high-variance models, adaptive boosting algorithms rely on sequentially fitting a series of low-variance models. Specifically, we rely on the gradient boosting algorithm and choose trees as our individual base learners (Chen and Guestrin, 2016). Gradient boosting works by sequentially adding shallow tree classifiers to the ensemble. Each new tree is fit to the residuals of the previous one, partially correcting the predecessor's errors and improving overall predictive performance. By sequentially combining models, boosting can substantially improve upon the prediction of the simple base model and explain large parts of the residual error. Gradient boosting features a large set of tuning parameters—specifically, the number of boosting iterations, the weight shrinkage factor, the maximum tree depth, the minimum tree split loss reduction, the subsample and column subsample ratio.

# 5 Results

## 5.1 Summary Statistics

We start by describing the New York and Texas samples. As discussed in Section 3, we split our samples into two groups: disabled teenage Medicaid enrollees (i.e., enrolled in SSI) and non-disabled teenage Medicaid enrollees (i.e., not enrolled in SSI). In Table 1, we report average annual healthcare spending by type of service for these two groups in each state. As expected, the table shows that the disabled Medicaid enrollees had significantly higher Medicaid spending vs. the non-disabled enrollees. These dissagregated healthcare spending measures will be the key inputs to our predictive models.

**Table 1:** Annual Healthcare Spending

**(a)** New York

| | Inpatient Spending | Outpatient Spending | Rx Spending | Long-Term Care Spending |
|---|---|---|---|---|
| Not Initially Disabled | $375 | $594 | $252 | $149 |
| Initially Disabled | $1,615 | $6,443 | $1,640 | $6,693 |

**(b)** Texas

| | Inpatient Spending | Outpatient Spending | Rx Spending | Long-Term Care Spending |
|---|---|---|---|---|
| Not Initially Disabled | $581 | $1,172 | $288 | $29 |
| Initially Disabled | $1,484 | $3,230 | $3,551 | $2,150 |

**Note:** Table shows mean healthcare spending by spending category for individuals in New York and Texas.

We now present overall statistics describing what portion of each group is enrolled in SSI at age 23 in each state.[4] In results not reported here, we also analyzed SSDI enrollment at any age but found that very few individuals enrolled in SSDI during our sample period (<0.1%), making prediction of this outcome infeasible.[5]

Table 2 shows SSI enrollment at age 23 for the disabled and non-disabled teenagers in each state. The table shows initial teenager SSI enrollment status (in rows) vs. SSI enrollment status at age 23 (columns). The first item of note from this table is that less than 2% (e.g., 4,223/343,382) of those not initially on SSI are on SSI at age 23. Given the small number of individuals in this category enrolling in SSI, prediction of this outcome is likely to be difficult. Indeed, our attempts to estimate predictive

[4] We use SSI enrollment at age 23 as our outcome of interest rather than enrollment at age 18 due to the fact that, for some teenagers, the redetermination and disenrollment process take several years.

[5] In future versions of this paper, we hope to extend the period during which we observe SSDI enrollment all the way up through 2018, meaning that we will observe a larger share of our original sample eventually enrolling in SSDI. We are currently impeded from doing so due to issues with the Medicaid-Medicare crosswalk provided to us by CMS, where a valid crosswalk is necessary for linking members of our original Medicaid sample to their eventual SSDI outcomes in the Medicare data. These issues affect data from 2011 on, preventing us from observing outcomes beyond 2010.

models for this outcome largely failed, leading us to focus instead on predicting disenrollment from the initially disabled group.

When considering SSI enrollment rates among the initially disabled group, we find that about 23% of initial enrollees in New York and 46% of initial enrollees in Texas are removed from SSI by age 23, in line with the national averages discussed in Section 2. For this group, there is a sufficiently large number of individuals enrolled in SSI at age 23 and a sufficiently large number of individuals not enrolled in SSI at age 23 for us to estimate our predictive models. For all remaining results, we focus on predicting this outcome for this group.

**Table 2:** Sample

**(a)** New York

| | Not Disabled at 23 | Disabled at 23 |
|---|---|---|
| Not Initially Disabled | 343,382 | 4,223 |
| Initially Disabled | 17,745 | 20,737 |

**(b)** Texas

| | Not Disabled at 23 | Disabled at 23 |
|---|---|---|
| Not Initially Disabled | 157,496 | 2,967 |
| Initially Disabled | 5,667 | 19,428 |

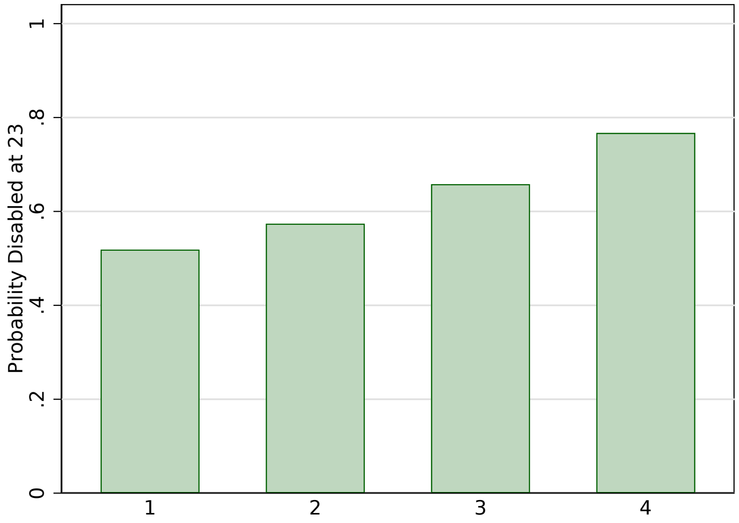
**Note:** Table shows our baseline sample for New York and Texas.

### 5.2 Relationship between Teenage Healthcare Utilization and SSI Disenrollment

We now use the linked health care and SSI enrollment data, consisting specifically of teenage Medicaid healthcare utilization data connected to SSI enrollment data, to determine whether teenage healthcare utilization is predictive of disenrollment from SSI. We start by comparing SSI enrollment rates at age 23 by quartile of healthcare spending for individuals who were enrolled in SSI as teenagers. These rates are presented in Figure 1. The figure shows a strong monotonic relationship between teenage healthcare spending and SSI enrollment at age 23. While only 52% of those in the lowest quartile of spending remain enrolled in SSI as adults, 58% of those in the second quartile, 66% of those in the third quartile, and 77% of those in the top quartile remain in SSI at age 23.

The gradient of age-23 SSI enrollment in teenage healthcare spending implies that taking this one variable and using a relatively coarse classification of the variable (quartiles), we can identify a group that has an 80% chance of remaining enrolled in SSI and a group that has about a 50% chance of remaining enrolled, a difference of 30 percentage points. Such a difference in the probability of disenrollment suggests that disabled teenagers in the top quartile of healthcare spending may not be particularly good targets for redetermination efforts, given that so few of them will eventually be removed from the program.

**Figure 1:** SSI Enrollment at Age 23 By Level of Medicaid Spending



**Note:** Figure shows SSI Enrollment at age 23 by quartile of teenage Medicaid spending.

### 5.3 Machine Learning

To more fully examine how teenage healthcare utilization is predictive of adult SSI enrollment among individuals who were disabled as teenagers, we now turn to advanced machine learning methods. These methods allow us to optimally use the predictive information embedded in the healthcare utilization data, maximizing explanatory power by exploring different functional forms (polynomials, interactions, etc.) while also using separate testing and estimation samples to avoid over-fitting.

We start by describing diagnostics for our predictive models and then use our predictions to stratify the sample and show how well the models allow us to classify individuals into groups that differ in their probability of being enrolled in SSI as adults.

**Table 3:** Predicting SSI receipt, Full Sample

| | $\ell_1$ penalized regression | | | | Random forest | | | | Gradient boosting | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Actual SSI receipt | | | | Actual SSI receipt | | | | Actual SSI receipt | |
| | | | 0 | 1 | | | 0 | 1 | | | 0 | 1 |
| | Prediction | 0 | 12858 | 2677 | Prediction | 0 | 13144 | 2549 | Prediction | 0 | 13574 | 2434 |
| | | 1 | 10554 | 37488 | | 1 | 10317 | 37687 | | 1 | 9838 | 37731 |
| Accuracy | 0.7919 | | | | 0.798 | | | | 0.807 | | | |
| 95% CI | (0.7887, 0.795) | | | | (0.7949, 0.8011) | | | | (0.8039, 0.81) | | | |
| Sensitivity | 0.9333 | | | | 0.9366 | | | | 0.9394 | | | |
| Specificity | 0.5492 | | | | 0.5602 | | | | 0.5798 | | | |
| Pos Pred Value | 0.7803 | | | | 0.7851 | | | | 0.7932 | | | |
| Neg Pred Value | 0.8277 | | | | 0.8376 | | | | 0.8480 | | | |
| Kappa | 0.519 | | | | 0.5337 | | | | 0.5558 | | | |
| Mcnemar's Test P-Value | $< 2.2\times10^{-16}$ | | | | $< 2.2\times10^{-16}$ | | | | $< 2.2\times10^{-16}$ | | | |

Note: Tuning parameters chosen via 10-fold cross-validation on a random 80% training sample. The table shows the results for the final models re-estimated on the full sample. The original model comparison on the 20% hold-out test sample is given in Appendix Table A1. As described in section 4, the input variables for the penalized regression model include a set of non-linear transformations of the original variables. We use trees as the base learner for the gradient boosting model.

**Estimation Results and Diagnostics** Table 3 shows the results for our prediction models. We report the results for the final models re-estimated on the full sample. Results for the hold-out sample are reported in Appendix Table A1.

Overall, the best option from each class of models achieves about 80% accuracy. This means that we correctly classify SSI receipt at age 23 for four out of every five people who are receiving SSI at age 18. We find that the different models show very similar performance and the ensemble methods only have a slight edge over penalized regression approaches. The accuracy of the logistic lasso is about 0.79, relative to which the random forest and boosted trees achieve about one percentage point better accuracy. Predictive performance across both SSI receipt and non-receipt states is also similar. For positive predicted value, about 78% to 79% of those predicted to remain on SSI actually remain on SSI. Negative predicted value is slightly larger, between 82% and 85% of those predicted to be disenrolled are actually disenrolled.

However, we find that the models do not perform equally well in identifying the true persistence of SSI and in identifying true disenrollment. For all three models, the true positive rate (sensitivity) is very high. About 93% of those remaining on SSI are correctly identified as such. In contrast to this, the true negative rate (specificity) of the models is lower. Between 55% and 58% percent of individuals disenrolled from SSI are correctly identified. This is also what distinguishes the different models. While sensitivity is similar for all three models, specificity is larger for the ensemble methods. The gradient boosting model is three percentage points more likely to correctly identify those who drop out.

At about 0.52 to 0.56, the Kappa statistics are moderately good. The Kappa statistic expresses the accuracy of our classifier compared to a random classifier given the unconditional class frequencies in the data. Again, the gradient boosting model achieves the best value for the statistic. The results for McNemar's test indicate that sensitivity and specificity for all models differ and the marginal propensities in the data are not equal.

One insight gained from training and model development is that comparable performance between the penalized regression and ensemble methods can only be achieved if the inputs for the penalized regression model include non-linear transformations of the original variables (higher order polynomials of continuous variables, second order interactions of all variables and polynomials); otherwise, performance is significantly worse (only about 0.7 accuracy or less). This allows the penalized regression model to account for more complicated non-linearities, which tree-based methods incorporate naturally. In general, random forests and boosted trees perform better for a wider range of tuning parameters in training.

**Identification of Most Predictive Variables** In the next step, we examine which predictor variables most enhance predictive power. In Table 4 we rank the variables by descending influence using model-specific variable importance computed during training. For the logistic lasso, we rank the penalized coefficients of the input variables and their transformations by coefficient size in descending order. For the tree based models, we compute variable importance as the number of times a variable is split on weighted by the depth of the split.

Looking at the 20 most influential variables, we find that similar variables are important. Across all models, indicators of the baseline age are important (age_18, age_17, etc.). Recall that these variables indicate the last year we observe the individual enrolled in Medicaid for at least 9 months. It is not surprising that this is predictive of SSI enrollment at age 23, as individuals who drop out of Medicaid at younger ages are likely the healthiest disabled teenagers and much less likely to re-enroll in SSI later in life. Gender, race, and whether the individual was on Medicaid in the year prior are also common to all models. In addition, indicators for the state, the baseline year, and certain counties can be found on both lists. All of these variables indicate that non-healthcare information in the Medicaid data seems to be highly predictive of SSI enrollment at age 23.

**Table 4:** Important Predictors

| $\ell_1$ penalized regression | | Random forest | | Gradient boosting | |
|---|---|---|---|---|---|
| Variable | Rank | Variable | Relative importance | Variable | Relative importance |
| cz999 | 20 | age_18 | 100.00 | age_18 | 100.000 |
| age_18 | 19 | cz134 | 36.34 | elg_cd22 | 74.786 |
| elg_cd12 * age_18 | 18 | mdcd_spend | 33.26 | age_17 | 55.397 |
| elg_cd22 | 17 | rx_spend | 30.45 | mdcd_spend | 48.808 |
| ot_spend | 16 | ot_spend | 30.03 | ot_spend | 45.829 |
| age_18 * mdcd_mo_12 | 15 | mdcd_mo_12 | 29.23 | rx_spend | 35.925 |
| medicaid_last_yr * cz134 | 14 | elg_cd22 | 28.36 | state | 24.327 |
| age_14 | 13 | elg_cd12 | 28.35 | age_16 | 23.320 |
| medicaid_last_yr | 12 | medicaid_last_yr | 22.23 | cz134 | 18.582 |
| year_2000 * state | 11 | state | 21.33 | elg_cd12 | 18.527 |
| mdcd_spend | 10 | ip_spend | 20.71 | mdcd_mo_12 | 15.094 |
| elg_cd42 | 9 | lt_spend | 20.53 | ip_spend | 13.882 |
| mdcd_spend * ot_spend | 8 | el_prvt_insrnc_mo_cnt | 20.12 | age_15 | 12.636 |
| ot_spend_2 | 7 | yr_2005 | 16.84 | age_14 | 9.442 |
| yr_2001 * state | 6 | tot_ltc_cvr_day_cnt_psych | 16.45 | lt_spend | 7.955 |
| cz102 | 5 | yr_2000 | 15.72 | medicaid_last_yr | 7.832 |
| yr_2002*state | 4 | female | 14.81 | el_prvt_insrnc_mo_cnt | 7.566 |
| elg_cd12*cz134 | 3 | race9 | 12.62 | yr_2005 | 5.903 |
| el_prvt_insrnc_mo_cnt | 2 | elg_cd42 | 11.67 | yr_2000 | 5.105 |
| cz96 | 1 | yr_2002 | 11.30 | female | 4.501 |

Note: The variable importance ranking for the Lasso is based on the size of the penalized model coefficients. Variable importance for the tree-based ensemble methods is computed as the number of times the variable is split on weighted by the depth of the split.

Importantly, we also find that variables related to healthcare utilization appear in the lists of most important variables. Overall Medicaid spending (mdcd_spend), outpatient spending (ot_spend), inpatient spending (ip_spend), long-term care spending (lt_spend), and drug spending (rx_spend) all appear in the top 20 variables for at least one of the three models. Inpatient and long-term care spending seem to appear higher (i.e., more important) on the list of important variables, indicating that these types of healthcare utilization are more predictive of SSI enrollment at age 23. Overall, these results suggest that even conditional on all of the non-healthcare information found in the Medicaid data, information about healthcare utilization remains important and highly predictive.

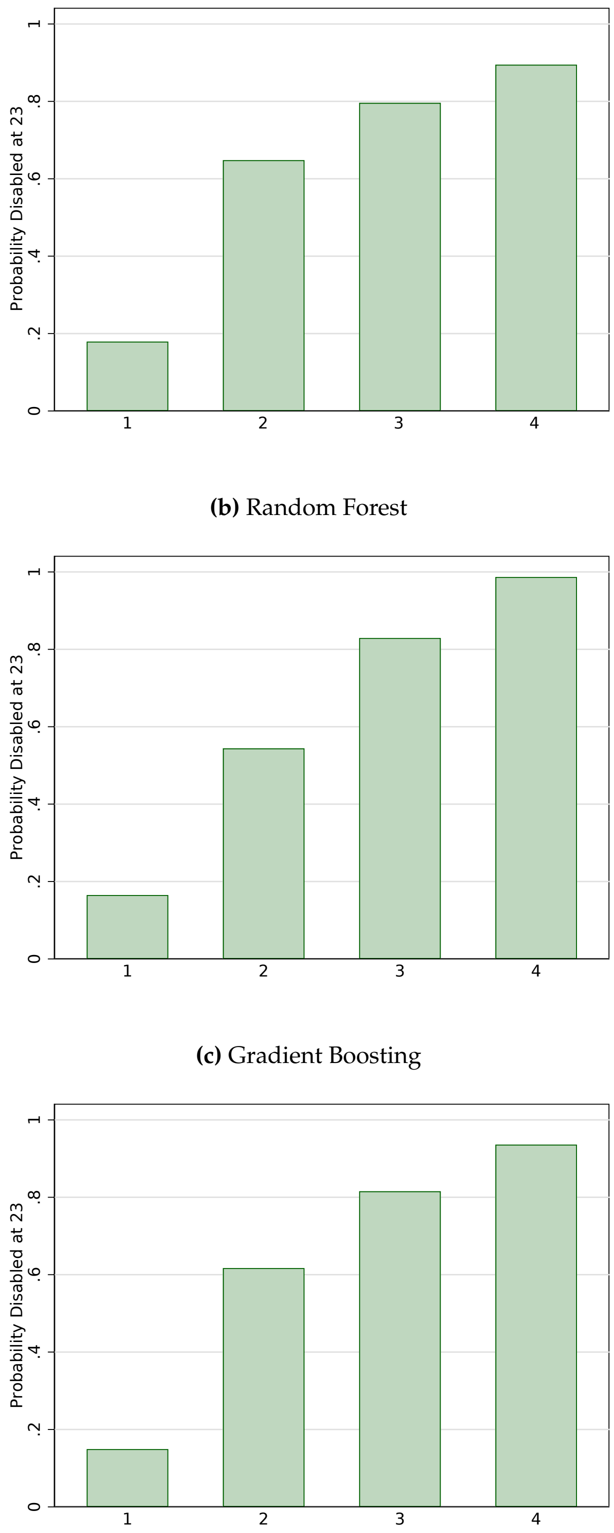
**Using Predictive Models to Identify Groups with High and Low Probability of Adult SSI Enrollment** Next, we repeat out previous quartile comparison by calculating average enrollment by quartiles of the predicted enrollment probability. Our purpose in doing this is to illustrate in a simple and transparent way the power of the predictive model for identifying groups of individuals who are *ex ante* likely to be disenrolled from SSI at age 18 vs. individuals who are *ex ante* unlikely to be disenrolled from SSI.

Figure 2 shows the actual likelihood of remaining on SSI by quartiles of the predicted probability of remaining on SSI. Teenagers who are predicted to remain on SSI with a higher likelihood are indeed more likely to be on SSI at age 23. When using Lasso, 18% of teenagers in the bottom quartile of the predicted probability distribution remain on SSI vs 90% for the top quartile. When using Random Forest, 17% in the bottom quartile and 99% in the top quartile of the predicted probability distribution are still on SSI at age 23. When using Gradient Boosting, 15% in the bottom quartile and 94% in the top quartile of the predicted probability distribution are still on SSI at age 23.

These differences in *ex post* probabilities of adult SSI enrollment based on *ex ante* information found in the Medicaid data imply that given this information and our predictive models, one can identify a (large) group of disabled teenagers with a lowered chance of being disenrolled from the SSI program at age 18. This suggests that targeting redetermination resources away from individuals in the top quartile of *ex ante* probability of persistence in the program may have little effect on actual disenrollment of individuals from SSI.

Recall, however, that many of the most important variables in the predictive model were not related to healthcare utilization. To understand the independent predictive power of information about healthcare utilization, we examine the predicted probabilities conditional on non-healthcare variables. Figure 3 shows the distribution of the likelihood of remaining on SSI by predicted probabilities residualized on non-healthcare variables (so that the remaining variation comes from healthcare variables). The difference between the likelihood of remaining on SSI between top and the bottom quartile of the residual predicted probabilities is 38 percentage points for Lasso, 54 percentage points for Random Forest, and 52 percentage points for Gradient Boosting. Again, for individuals in the top quartile of residual predicted probabilities, the vast majority (over 80%) of individuals stay on SSI. While the healthcare variables alone do not perform as well as when combined with other information (demographics, age, geography, etc.) found in the Medicaid data, these variables do appear to have substantial predictive power. Importantly, the machine learning methods seem to add value here, with the difference in *ex post* probabilities of SSI enrollment between the first and fourth quartiles of *ex ante* residual predicted probabilities being significantly larger than the same difference when just grouping individuals by overall Medicaid spending as presented in Figure 1.

**Figure 2:** SSI Enrollment at Age 23 By Predicted Probability



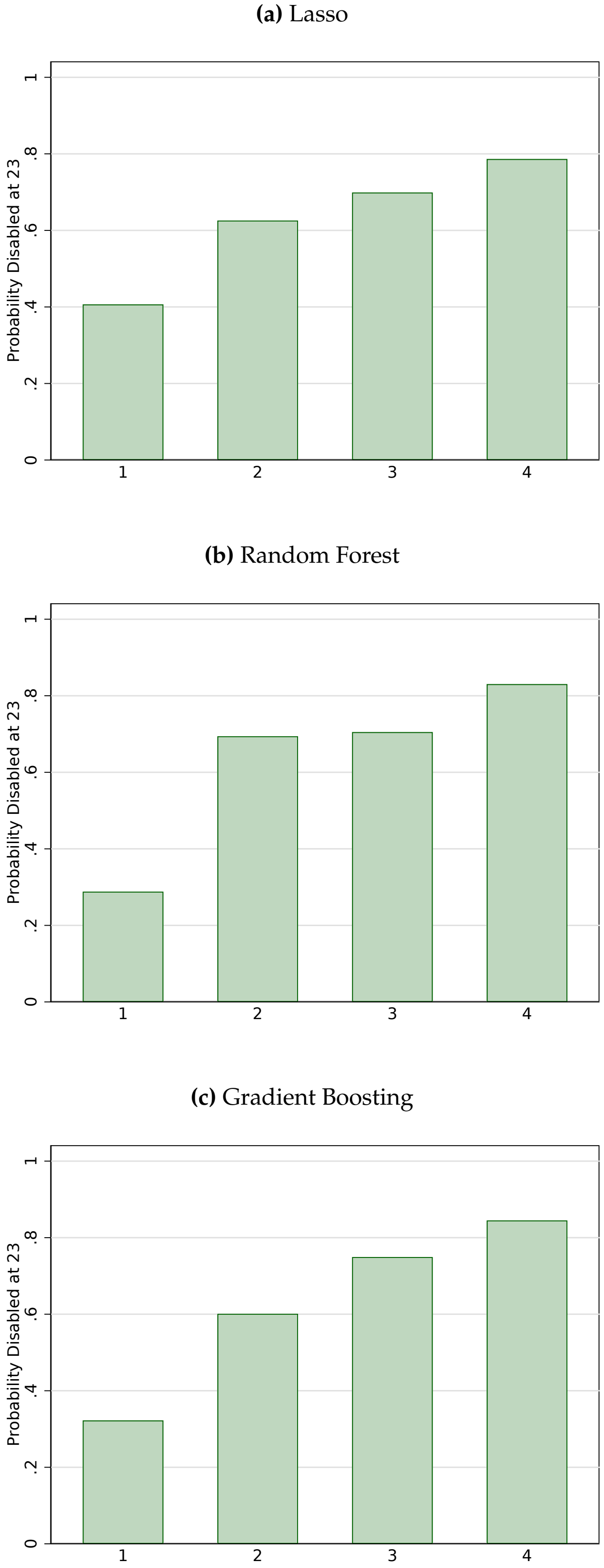
**(a)** Lasso

**(b)** Random Forest

**(c)** Gradient Boosting

**Note:** Figure shows SSI Enrollment at age 23 by quartile of predicted enrollment probability.

**Figure 3:** SSI Enrollment at Age 23 By Predicted Probability Conditional on Non-Healthcare Variables



**(a)** Lasso

**(b)** Random Forest

**(c)** Gradient Boosting

**Note:** Figure shows SSI Enrollment at age 23 by quartile of predicted enrollment probability.

# 6 Conclusion

In this paper, we used machine learning techniques on detailed Medicaid claims data to predict SSI enrollment among teenagers reaching adulthood. Our results suggest that relying on Medicaid data and machine learning techniques can be useful in predicting continuing disability insurance enrollment.

We find that both healthcare and non-healthcare information found in Medicaid data is predictive of eventual disenrollment from SSI. Using both sets of information combined, the *ex post* probability of disenrollment from SSI for the top quartile of individuals by *ex ante* predicted probability of adult enrollment in SSI is a mere 1%. This indicates that our predictive models are powerful for identifying groups for which persistence in the program (i.e., non-removal at adulthood) is a near certainty. The ability to do this could be valuable for targeting redetermination resources to groups where the removal is less certain.

We further investigated the independent role of healthcare information for predicting SSI enrollment at age 23. We found that healthcare has important predictive power, and that the machine learning methods we use are important for fully harnessing that power. The implications of these results are unclear. It could be the case that healthcare utilization helps identify groups with more vs. less severe disabilities. However, it could also be the case that use of healthcare services has a causal impact on whether a teenager's condition improves and he or she exits the SSI program. Disentangling these selection and treatment effects represents a key area of future research.

# References

**Autor, David, Andreas Kostøl, Magne Mogstad, and Bradley Setzler.** 2019. "Disability Benefits, Consumption Insurance, and Household Labor Supply." *American Economic Review*, 109(7): 2613–2654.

**Autor, David H.** 2011. "The Unsustainable Rise of the Disability Rolls in the United States: Causes, Consequences, and Policy Options." National Bureau of Economic Research Working Paper 17697.

**Autor, David H., and Mark G. Duggan.** 2003. "The Rise in the Disability Rolls and the Decline in Unemployment." *Quarterly Journal of Economics*, 118(1): 157âĂŞ206.

**Autor, David H., and Mark G. Duggan.** 2006. "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding." *Journal of Economic Perspectives*, 20(3): 71–96.

**Breiman, Leo.** 1996. "Bagging Predictors." *Machine Learning*, 24(2): 123–140.

**Breiman, Leo.** 2001. "Random Forests." *Machine Learning*, 45(1): 5–32.

**Breiman, Leo, Jerome Friedman, Richard A Olshen, and Charles J Stone.** 1984. *Classification and Regression Trees. The Wadsworth statistics / probability series*, CRC.

**Chen, Tianqi, and Carlos Guestrin.** 2016. "XGBoost: A Scalable Tree Boosting System." *CoRR*, abs/1603.02754.

**Congressional Budget Office.** 2016. "Social Security Disability Insurance: Participation and Spending." https://www.cbo.gov/publication/51443.

**Cutler, David M., Ellen Meara, and Seth Richards-Shubik.** 2011. "Health Shocks and Disability Transitions Among Near-elderly Workers." National Bureau of Economic Research Retirement Research Center Paper NB11-08.

**Dahl, Gordon B., Andreas Ravndal Kostol, and Magne Mogstad.** 2014. "Family Welfare Cultures." *Quarterly Journal of Economics*, 129(4): 1711–1752.

**Deshpande, Manasi.** 2016*a*. "Does Welfare Inhibit Success? The Long-Term Effects of Removing Low-Income Youth from the Disability Rolls." *American Economic Review*, 106(11): 3300–3330.

**Deshpande, Manasi.** 2016*b*. "The Effect of Disability Payments on Household Earnings and Income: Evidence from the SSI Children's Program." *Review of Economics and Statistics*, 98(4): 638–654.

**Deshpande, Manasi, Tal Gross, and Yalun Su.** 2019. "Disability and Distress: The Effect of Disability Programs on Financial Outcomes." National Bureau of Economic Research Working Paper 25642.

**Duggan, Mark, and Melissa Schettini Kearney.** 2007. "The Impact of Child SSI Enrollment on Household Outcomes." *Journal of Policy Analysis and Management*, 26(4): 861–885.

**French, Eric, and Jae Song.** 2014. "The Effect of Disability Insurance Receipt on Labor Supply." *American Economic Journal: Economic Policy*, 6(2): 291–337.

**Friedman, Jerome H.** 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, 29(5): 1189–1232.

**Friedman, Jerome H.** 2002. "Stochastic Gradient Boosting." *Computational Statistics and Data Analysis*, 38(4): 367–378.

**Gelber, Alexander, Timothy J. Moore, and Alexander Strand.** 2017. "The Effect of Disability Insurance Payments on Beneficiaries' Earnings." *American Economic Journal: Economic Policy*, 9(3): 229–261.

**Gelber, Alexander, Timothy J. Moore, and Alexander Strand.** 2018. "Disability Insurance Income Saves Lives." *Mimeo*.

**Hastie, Trevor, Robert Tibshirani, and Jerome Friedman.** 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York.

**Hemmeter, Jeffrey, and Elaine Gilby.** 2009. "The Age-18 Redetermination and Postredetermination Participation in SSI." *Social Security Bulletin*, 69(4): 1–25.

**Kaiser Family Foundation.** 2014. "Medicaid Spending by Enrollment Group." https://www.kff.org/medicaid/state-indicator/medicaid-spending-by-enrollment-group/.

**Liebman, Jeffrey B.** 2015. "Understanding the Increase in Disability Insurance Benefit Receipt in the United States." *Journal of Economic Perspectives*, 29(2): 123–50.

**Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand.** 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review*, 103(5): 1797–1829.

**Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand.** 2015. "Disability Insurance and the Great Recession." *American Economic Review*, 105(5): 177–182.

**Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand.** 2018. "The Effect of Economic Conditions on the Disability Insurance Program: Evidence from the Great Recession." National Bureau of Economic Research Working Paper 25338.

**Moore, Timothy.** 2015. "The Employment Effects of Terminating Disability Benefits." *Journal of Public Economics*, 124(1): 30–43.

**Park, Sujeong, and David Powell.** 2019. "Is the Rise in Illicit Opioids Affecting Labor Supply and Disability Claiming Rates?" *Mimeo*.

**Social Security Administration.** 2018a. *Annual Statistical Report on the Social Security Disability Insurance Program, 2017.* Washington, DC.

**Social Security Administration.** 2018b. *Annual Report of the Supplemental Security Income Program, 2018.* Baltimore, MD.

**Tibshirani, Robert.** 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288.

## **Table A1:** Predicting SSI receipt, Hold-out Sample

| | $\ell_1$ penalized regression | | | | Random forest | | | | Gradient boosting | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Actual SSI receipt | | | | Actual SSI receipt | | | | Actual SSI receipt | | | |
| | | | 0 | 1 | | | 0 | 1 | | | 0 | 1 |
| | Prediction | 0 | 2543 | 575 | Prediction | 0 | 2585 | 555 | Prediction | 0 | 2603 | 581 |
| | | 1 | 2127 | 7494 | | 1 | 2103 | 7496 | | 1 | 2074 | 7457 |
| Accuracy | 0.7879 | | | | 0.7913 | | | | 0.7912 | | | |
| 95% CI | (0.7807, 0.795) | | | | (0.7842, 0.7984) | | | | (0.784, 0.7982) | | | |
| Sensitivity | 0.9287 | | | | 0.9311 | | | | 0.9277 | | | |
| Specificity | 0.5445 | | | | 0.5514 | | | | 0.5566 | | | |
| Pos Pred Value | 0.7789 | | | | 0.7809 | | | | 0.7824 | | | |
| Neg Pred Value | 0.8156 | | | | 0.8232 | | | | 0.8175 | | | |
| Kappa | 0.5089 | | | | 0.5182 | | | | 0.5189 | | | |
| Mcnemar's Test P-Value | $< 2.2\times10^{-16}$ | | | | $< 2.2\times10^{-16}$ | | | | $< 2.2\times10^{-16}$ | | | |

Note: Tuning parameters chosen via 10-fold cross-validation on a random 80% training sample. The table shows the results for the final models re-estimated on the full sample. The re-estimated models for the full sample are given in Table 3. As described in section 4, the input variables for the penalized regression model include a set of non-linear transformations of the original variables. We use trees as the base learner for the gradient boosting model.