The Retaining Employment and Talent After Injury/Illness Network (RETAIN) Demonstration: Evaluation Findings One Year After Enrollment

Supplementary Analysis: Predicting Need for RETAIN Services Based on Control Group Outcomes

February 2026

Yonatan Ben-Shalom, Isabel Musse, Monica Farid, Moriah Bauman, Kara Peterik, Irina Degtiar, Ian Huff, Phil Killewald, and Ankita Patnaik

Submitted to:

Social Security Administration
OAG/DPC
6401 Security Boulevard
1540 Robert M. Ball Bldg.
Baltimore, MD 21235-0001
Project Officer: Taffy McCoy
Contract Number: 28321319C00060001

Submitted by:

Mathematica
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005
Project Director: Yonatan Ben-Shalom
Reference Number: 50751

This page has been left blank for double-sided copying.

Contents

Exhibits

I. Supplementary Analysis: Predicting Need for RETAIN Services Based on Control Group Outcomes

A. Introduction

The Retaining Employment and Talent after Injury/Illness Network (RETAIN) demonstration was a joint initiative of the U.S. Department of Labor (DOL) and the Social Security Administration (SSA). RETAIN aimed to help workers with recently acquired injuries and illnesses remain in the labor force and avoid the need to apply for disability programs such as Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI). In Phase 2 of the RETAIN demonstration, DOL funded programs in Kansas (RETAINWOR*KS*), Kentucky (RETAIN Kentucky), Minnesota (Minnesota RETAIN), Ohio (Ohio RETAIN), and Vermont RETAIN) to identify, enroll, and provide stay-at-work or return-to-work (SAW/RTW) services to people with recent injuries or illnesses at risk of exiting the workforce. The RETAIN programs began enrolling participants in late 2021 and continued enrollment through mid-May 2024. Each program randomly assigned enrollees to a treatment group (eligible to receive services through RETAIN for up to six months) and a control group (ineligible for RETAIN services).

Key findings

- A simple linear regression model predicted return-to-work outcomes as well as more complex machine learning methods.
- The strongest predictors of not returning to work at 2 months were: time since last worked at enrollment, employment status at enrollment, and health status at enrollment.
- About 43 percent of RETAIN participants were predicted to be in high need of RETAIN services, based on their characteristics at the time of enrollment and results of the predictive model.
- RETAINWOR*KS* had favorable impacts on employment, earnings, and disability applications in both the high- and not-high-need groups.
- Minnesota RETAIN reduced disability applications in the not-high-need group and Vermont RETAIN reduced earnings in the high-need group.
- RETAIN Kentucky and Ohio RETAIN did not impact employment, earnings, or disability benefit applications for enrollees in either the high-need or not-high-need groups.
- Predictive models can help programs plan and tailor service intensity but might not be suited to screening participants out of services.

A central challenge in designing effective interventions is determining which participants are most likely to benefit. Because programs typically have limited capacity, it is essential to prioritize those individuals who stand to gain the most from the services offered. In the context of the RETAIN demonstration, many workers return to work following an injury or illness without any intervention beyond standard medical care. For such workers, RETAIN services are unlikely to improve their outcomes. Understanding how baseline characteristics relate to RTW outcomes is critical to effectively directing finite resources to workers who are most likely to benefit from them because they are unlikely to return to work without RETAIN-like services. Improving outcomes for these workers will, in turn, reduce their future need for federal disability benefits.

The five RETAIN programs used different strategies to recruit potential enrollees (Croake et al. 2023). Some programs used a data-driven approach to identify potential enrollees in electronic medical record data. Other programs used a more indirect strategy, relying on referrals from medical providers and other sources to identify potential enrollees. However, given the novelty of RETAIN and limited information on

RTW outcomes for similar populations, the RETAIN programs lacked an established, data-informed way to predict who will not return to work on their own and would therefore benefit from RETAIN services. Ultimately, many RETAIN enrollees in the control group returned to work in the absence of RETAIN services. According to the early follow-up survey of enrollees that Mathematica conducted, about 60 percent of RETAIN enrollees in the control group were working about two months after enrollment (Patnaik at al. 2025).

In this appendix, we analyze the risk factors for *not* returning to work and applying for SSDI or SSI in the absence of RETAIN services. Our predictive analysis used control group data to identify the baseline characteristics most predictive of these outcomes in the absence of RETAIN services. We then used our preferred predictive model to classify all RETAIN enrollees into "high need" and "not high need" groups based on their predicted probability of not returning to work within two months after enrollment. Then, for each state, we compared impact estimates between the two groups for the RETAIN evaluation's three primary outcomes: employment in the fourth quarter after enrollment, earnings in the four quarters after enrollment, and having applied for SSDI or SSI in the 12 months after enrollment.

Across the five RETAIN programs, the preferred predictive model classified between 36 and 50 percent of participants as high need—those least likely to return to work within two months in the absence of services. High-need participants differed substantially from others at baseline, with lower labor force attachment, poorer self-rated health, and lower earnings prior to enrollment. In RETAINWOR*KS*, the only program that generated broad positive impacts, both high- and not–high-need participants experienced gains in employment and earnings, and reductions in applications for SSDI and SSI. The point estimates of impacts on employment and disability applications were larger for the high-need group; however, they were not statistically significantly different from those for the not-high-need group.

In what follows, we first provide background information on how predictive analytics can support programs aimed at helping people stay at or return to work after an injury or illness. We then describe the data and methods we used, how we selected the preferred predictive model, and our findings based on that model. We conclude with a discussion of the findings.

B. Background

Below, we provide context for understanding how predictive analytics can support programs aimed at helping people stay at or return to work after an injury or illness. We first review individual-level factors that are associated with longer-term work absence or eventual application for federal disability benefits. We then examine how predictive analytics models can be used to identify people who might benefit most from early intervention and program support. Finally, we highlight key concerns related to the accuracy and fairness of predictive models, particularly in the context of public programs, and underscore the importance of transparency and thoughtful model development.

1. Individual factors linked to long-term work absence or application for federal disability benefits after injury or illness

Previous research has identified individual factors associated with risk of prolonged absence from work or application for federal disability benefits (SSDI or SSI) following an injury or illness. These factors include

receipt of workers' compensation (WC) or employer- or state-provided short-term disability insurance (STDI) benefits, type of injury or illness, and sociodemographic characteristics.

Before turning to federal disability benefits, and depending on access to coverage and other circumstances, people with a work-limiting injury or illness might pursue WC or STDI benefits that their employer or state provide. For work-related injuries or illness, WC provides injured employees with medical care and partial wage replacement while they recover.[1] STDI provides partial wage replacement (but not medical care) to employees who are temporarily unable to work due to non-work-related injury or illness.[2] Several studies have suggested that receipt of WC or STDI benefits might be associated with subsequent application to federal disability benefits. For example, Contreary and Honeycutt (2020) found relatively high rates of SSDI benefit application among new recipients of WC and employer-sponsored disability benefits. However, other research has suggested that workers at risk of exiting the workforce due to injury or illness might be less likely to apply for SSDI if they live in a state with mandatory STDI coverage (Ben-Shalom et al. 2021).

Analyses of WC and STDI claims have found certain diagnoses to be associated with longer benefit duration or benefit exhaustion, indicating prolonged absences from work following an injury or illness. For example, Neuhauser et al. (2018) found that for both WC and STDI, claims related to musculoskeletal injuries or illness and those related to mental disorders tended to last longer compared to claims for other types of injuries or illnesses. Contreary et al. (2018) found that cancer, intervertebral disc disorders, back disease, and mental health disorders were more prevalent among STDI claimants who exhausted their benefits compared to claimants who did not exhaust their benefits.

Researchers have also identified personal characteristics such as age, sex, and race that might increase risk of prolonged absence from work or application for SSDI following injury or illness. Older age is associated with increased risk of exhaustion of STDI claims (Contreary et al. 2018), long-duration WC and STDI claims (Neuhauser et al. 2018) and exiting the workforce due to injury or illness (Ben-Shalom et al. 2021). SSDI beneficiaries also tend to be significantly older than the general population (Livermore et al. 2010). Claimants' sex might also be associated with risk of extended work absence or application for federal disability benefits following an injury or illness. Contreary et al. (2018) found that STDI claimants who exhausted their benefits were more likely to be male, and Neuhauser et al. (2018) showed that the proportion of male claimants increases as benefit claim duration increases. Risk of exiting the workforce following disability also appears to be higher for men than it is for women, and for people who are Black compared to those who are not (Ben-Shalom et al. 2021).

Several interrelated socioeconomic factors such as education level, income level, and employment industry might also be associated with prolonged work absence or application for federal disability benefits following an injury or illness. The risk of exiting the workforce due to disability drops as education level rises (Ben-Shalom et al. 2021), and SSDI beneficiaries are less likely to have completed education beyond the high school level compared to the general population (Livermore et al. 2010). Relatedly, STDI

[1] Workers' compensation is available to virtually all employees in the United States, but the level of benefits and quality of care vary widely across states and employers.

[2] Five states have mandatory short-term disability insurance benefits: California, Hawaii, New Jersey, New York, and Rhode Island. In all other states, most workers do not have any short-term (let alone long-term disability) insurance coverage (Bureau of Labor Statistics 2020).

claimants in Rhode Island who did not return to work before exhausting their benefits had higher pre-injury poverty rates compared with those who did (Bourbonniere and Mann 2018). Finally, Contreary et al. (2018) found that claimants who exhausted their STDI benefits were more likely to be employed in labor-intensive industries (for example, agriculture, mining, construction, transportation, utilities) than claimants who did not exhaust their benefits.

Prior research highlights individual factors linked to a higher risk of prolonged work absence. However, a predictive model that integrates multiple baseline characteristics and circumstances can better capture the complex interactions among these factors. Such a model can also generate a simple binary flag or risk score, enabling program implementers to assess the need level of potential participants more effectively.

2. Identifying potential program beneficiaries through predictive analytics

Over the past several decades, researchers and program implementers have increasingly used predictive analytics models to target early intervention services more effectively to those who need them. For example, researchers have used health data from the United Kingdom, New Zealand, and Australia to develop models that identify patients at risk of hospital readmission or future emergency department visits (Billings et al. 2012; Panattoni et al. 2011).

One of the most well-known and successful uses of predictive analytics to identify people in need of public program services is the Allegheny Family Screening Tool (Hurley 2018). In 2016, Allegheny County, Pennsylvania, became the first U.S. jurisdiction to adopt a predictive analytics algorithm to help child protective services (CPS) workers assess the risk associated with child abuse or neglect allegations reported to the county's CPS hotline. To supplement the often-limited information available in hotline reports, the Allegheny Family Screening Tool draws on a wide array of data available in the county's integrated data system, including prior use of CPS and mental health, drug and alcohol use, and homeless services (Chouldechova et al. 2018). The tool generates a risk score indicating the likelihood that the child or children involved will be removed from the home within two years. Importantly, the tool's purpose is to support, not replace, CPS workers' decisions about whether to "screen in" (investigate) or "screen out" each case.

Despite early implementation challenges, evidence suggests the tool improved screening decisions. Sixteen months after the tool was rolled out, CPS workers were recommending fewer low-risk cases and more high-risk cases for further investigation, signaling greater reliance on the tool and a potentially more efficient use of CPS resources (Hurley 2018).

Predictive analytics have also shaped workforce development programs. In DOL's Reemployment Services and Eligibility Assessment program, some states built statistical profiling models to identify unemployment insurance claimants most likely to exhaust benefits and therefore most in need of reemployment services (Trutko et al. 2022). These models relied on claimants' baseline characteristics and employment histories to guide service prioritization. States, however, varied widely in their use of predictive approaches, and many still relied on simpler administrative rules or self-assessments. This mix of strategies shows both the potential of predictive modeling to strengthen targeting in workforce programs and the fact that systematic adoption remains at an early stage.

3. Concerns about the accuracy and fairness of predictive models

One key concern with using predictive analytics models is the risk of perpetuating biases or inaccuracies, especially in public policy and healthcare contexts where vulnerable populations might be adversely affected. Predictive models are only as accurate as the data used to train them. If training data contain biases, those biases can be embedded in the model's outputs. For example, research on a widely used commercial model designed to identify patients for "high-risk care management" programs revealed substantial racial bias in its predictions (Obermeyer et al. 2019). The model predicted that White patients required more health services than Black patients with similar medical conditions. Researchers attributed this bias to the model's reliance on healthcare cost and service use data, which reflect historic barriers that have limited access to care for Black patients, resulting in lower recorded costs and service use compared with White patients.

To mitigate these risks, model developers must carefully assess and address potential biases in the data they use. Strategies include incorporating additional data elements, ensuring data cleanliness, and thoughtfully managing missing data (Whicher et al. 2022). Developing accurate and trustworthy predictive analytics models also means building transparent and interpretable models in collaboration with the people who will use them and critically considering the training data used to build them. Unlike some commercial tools that have drawn criticism for their opacity or potential bias,[3] the Allegheny Family Screening Tool is notable for its transparency and stakeholder engagement. The tool, which researchers developed but the county owns, has been described in academic publications and was developed with feedback from local officials, CPS experts, and the local community (Hurley 2018). This collaborative and transparent approach has contributed to more accurate and fair screening outcomes. Sixteen months after its rollout, researchers found that CPS workers were treating allegations against Black and White families more consistently (Hurley 2018).

Data on RETAIN enrollees and their outcomes are not publicly available, but this appendix provides transparency into the data elements and methods we used to develop the predictive model we present.

A. Data and Methods

1. Data

This analysis relied on state program data, data from Mathematica's surveys of enrollees, and SSA administrative data. RETAIN states collected information on enrollees at initial intake, before random assignment. The RETAIN enrollment data include baseline information on demographic characteristics, qualifying injury or illness, recent employment, and health insurance coverage. These data provide independent variables for the predictive models we discuss later.

Mathematica conducted two-follow up surveys of RETAIN enrollees, one two months after enrollment (field period January 2022 to October 2024) and another 12 months after enrollment (field period January 2023 to August 2025). The surveys included questions on services and training used, employment,

[3] See, for example, the controversy around the widely used Correctional Offender Management Profiling for Alternative Sanctions model, which aims to predict a criminal defendant's risk of committing another crime (Yong 2018).

earnings, and health. We used the survey data to construct two outcomes for the predictive models: (1) not working two months after enrollment and (2) not working 12 months after enrollment.

Finally, we obtained information on RETAIN enrollees' SSI and SSDI applications from SSA's Structured Data Repository. We used these data to construct an outcome representing application to SSDI or SSI in the 12 months after enrollment.

2. Analytical samples

For prediction modeling, the analytical sample pooled control group enrollees from all five RETAIN programs. Pooling enrollees from all five programs into one prediction model increases the number of observations and variety of features we exposed to the model in the training data, enabling it to learn similarities between enrollees across RETAIN programs. Pooling across states also ensured the model's predictions were not specific to one state's program or population but were applicable nationwide. We expected the different RETAIN programs to share common factors that predict not returning to work or applying to SSA benefits.[4] Exhibit 1 summarizes means and sample sizes for the outcomes we sought to predict for all control group enrollees; the sample size varies by outcome based on the data source.[5]

For the predictive model, we restricted the sample to the control group because we sought to predict outcomes in the absence of RETAIN services. For the analysis of impacts by predicted need, we used baseline and outcome information for both the control and treatment groups in each state.

Exhibit 1. Outcomes of RETAIN enrollees in the control group (percentages)

Outcome	Percentage	Sample size
Not working 2 months after enrollment	37.4	5,080
Not working 12 months after enrollment	28.9	4,209
Applied for SSDI or SSI benefits by 12 months after enrollment	9.0	6,045

Sources: RETAIN enrollee survey data; SSA administrative data.

SSA = Social Security Administration; SSDI = Social Security Disability Insurance; SSI = Supplemental Security Income.

3. Predictive models

We assessed different types of predictive models, of varying complexity, before deciding which model to use to identify the need level of RETAIN enrollees. Specifically, we fit three types of predictive models to the pooled data set: (1) linear regression classifiers,[6] (2) random forest classifiers, and (3) neural network classifiers. We trained each model to predict each of the binary outcomes listed in Exhibit 1 for each participant using the baseline characteristics of RETAIN enrollees as predictors (for a total of three

[4] Although state policies and economic environments could moderate the relationship between specific risk factors and return to work or SSA applications, we sought to identify factors that are predictive across state environments. The state programs also varied in their approach to recruitment for RETAIN (Croake et al. 2023), which could affect the composition of control group enrollees in each program. However, each RETAIN program included enrollees across all categories of predictive factors (see Exhibit 7 for summary statistics on predictive factors).

[5] Sample sizes varied widely by program; the predictive models were driven more by control group enrollees in programs with larger enrollment numbers.

[6] We also fit a logistic regression. Because the predictive performance of the logistic and linear regressions were very similar, we describe findings based on the linear regression for further ease in interpretation.

independently trained models per model type).[7] Prior to estimating each model, we randomly split the data set into train and test samples. The train samples contain 75 percent of the full analytic samples, and the test samples contain the remaining 25 percent.

To estimate model performance on unseen data, we evaluated model performance metrics on the 25 percent test sample. For each outcome, we optimized any hyperparameters[8] of each model using a 10-fold cross-validation technique on the train sample only.[9] We then fit the model on the entire train sample using the optimal hyperparameters learned during the cross-validation process.

We provide more detail on each type of predictive model and the performance metrics we used to assess those models below.

Linear regression classifiers model the relationship between a dependent variable and the independent variables by fitting a linear equation to the observed data. Linear regression classifiers are parametric in nature: they are easy to interpret and work well when the relationship between the input variables and the outcomes is linear and additive. We did not regularize the linear regression classifiers, so we did not need to optimize hyperparameters for these models.

Random forest classifiers predict outcomes by combining the results of many decision trees. Each decision tree splits the data into groups based on different features, step by step, until it reaches a classification. To build the trees, the random forest algorithm randomly selects a subset of features at each split and then chooses the feature that best separates the data into outcome classes. The algorithm repeats this process down the branches of the tree until it reaches a stopping rule..

During training, the algorithm grows a set number of independent trees. To make a prediction, the random forest combines the "votes" from all decision trees and assigns the majority class to the observation. Random forests are nonparametric models that can capture complex relationships between input variables and outcomes, but they are harder to inspect and interpret than linear regression classifiers. Random forest models include several hyperparameters; to maximize out-of-sample performance, we tuned the maximum tree depth, the number of features considered at each split, and the maximum number of terminal nodes per tree.

Neural network classifiers extend linear classifiers by stacking multiple layers of linear and nonlinear transformations. This nonparametric method uses nonlinear functions of the input variables to predict outcomes. Because neural networks can model complex interactions and contain many parameters, they are generally less interpretable than random forests or linear classifiers. To improve out-of-sample

[7] See Exhibit 7 for the full list of baseline characteristics we include in each model.

[8] Hyperparameters, or user-specified settings that control how a model is trained, can affect various aspects of the training process, including the complexity of the model, the learning rate, and the regularization techniques used.

[9] Ten-fold cross-validation is a model evaluation technique where the data set is split into 10 equal parts; the model is trained on nine parts and tested on the remaining part, repeating this process 10 times so each part serves as the test set once. The results from each fold are averaged to provide a robust estimate of the model's performance on unseen data.

performance, we tuned several hyperparameters: the number of layers, the size of each layer, the activation function, and a regularization parameter used during training.[10]

Performance metrics. We assessed each model's performance in the test sample using the following metrics: accuracy, recall, F1 score,[11] and area under the receiver operating characteristic (ROC) curve (AUC). To assess the first three metrics, we computed an optimal predicted outcome threshold for each prediction model using Youden's method (Youden 1950). Youden's method selects a threshold, above which a predicted outcome is considered a "positive" outcome, by maximizing a metric balancing sensitivity (true positive rate) and specificity (true negative rate) defined as sensitivity + specificity -1.

The ROC curve is a parametric function relating the true positive rate of a classifier to its false positive rate at various classification thresholds (Fawcett 2006). The ROC AUC summarizes a classifier's ability to distinguish between classes by measuring the area under the ROC curve. This feature of the ROC AUC makes it useful for evaluating prediction models that might be used to direct scarce resources: if the observations are sorted by predicted outcome, models with a greater ROC AUC will do a better job sorting the true positive outcomes first in the list than models with a lesser ROC AUC.

Model selection. Ultimately, as we explain in the results section, we chose the linear regression classification as the main model for further analysis. We based this decision on four key considerations: (1) how accurately each model predicted outcomes, (2) how familiar relevant audiences might be with the model, (3) how well the model supports scaling, and (4) how easy it is to interpret. While only the first factor directly reflects model performance, the others influence how useful and accessible the results are for policy audiences.

4. Sample classification and subgroup analysis

We used the main model, the linear regression classifier, to classify enrollees in both the treatment and control groups into "high-need" and "not high-need" enrollees based on the prediction of their probability of not returning to work within two months after enrollment.[12] The rationale for focusing on this outcome is that if an enrollee in the control group returned to work within two months, access to RETAIN services would likely have had a minimal impact on their return to work. In contrast, if an enrollee in the control group was still not working after two months, access to RETAIN services could plausibly have influenced their RTW trajectory. We then conducted subgroup analyses, separately by state, to understand whether and how the impacts of RETAIN vary based on whether enrollees are in the high-

[10] We use the logistic or hyperbolic tangent activation function for all layers, and the adaptive moment estimation error backpropagation algorithm implemented in version 1.6.1 of the Python scikit-learn package (Pedregosa et al. 2011).

[11] The accuracy rate captures the share of enrollees for whom the model's prediction was correct. Recall is the true positive rate or the share of true positive instances that the model correctly identified. Precision is the share of all instances predicted as positive by the model that were indeed positive. The F1 score is the harmonic mean of precision and recall; it provides a single number to reflect the balance between precision and recall. If both precision and recall are high, the F1 score is high.

[12] For both the treatment and control groups, the predicted probability of not returning to work within two months after enrollment represents what we expect would happen in the absence of using RETAIN services, because the predictive model only uses information that is available at baseline, before any of the treatment group receives any services.

need group. We hypothesized that if RETAIN services have positive impacts, impacts in the high-need group will be larger because enrollees in this group are less likely to return to work within two months in the absence of RETAIN and therefore have the most room for improvement.

B. Model Selection

After fitting the three models, we compared their predictive performance and other features (audience familiarity, interpretability, and scalability) and chose the linear regression classification as the primary model for further analysis. As detailed below, the predictive performance of the linear regression classification was comparable to that of the random forest and neural network classifications. However, linear regression is more widely understood among policy stakeholders, offers interpretable coefficients, and is less computationally intensive, making it easier to scale and apply in practical settings.

1. Predictive performance

A model with the highest predictive performance is best able to correctly predict the outcomes of interest. To evaluate each model, we calculated several performance metrics: accuracy, recall, the F1 score, and the ROC AUC. All three models demonstrated broadly similar performance in the test data set (Exhibit 2). For the outcome of not working two months after enrollment, the AUC was 0.77 for all three models. For not working 12 months after enrollment, AUCs were slightly lower, spanning 0.72 (linear regression and neural network) to 0.73 (random forest). For predicting whether an enrollee applied for SSDI or SSI within 12 months, AUCs ranged from 0.71 (neural network) to 0.74 (linear regression and Random Forest). Exhibit 3 presents ROC curves of the test data sets for all models and three outcomes examined.

Following the guidance that AUC values of 0.70–0.80 indicate acceptable discrimination and values above 0.80 indicate excellent discrimination (Mandrekar 2010), our models achieve acceptable predictive power, though none reach the excellent benchmark of 0.80. This moderate performance likely stems from our reliance solely on information collected at enrollee intake; richer data—such as detailed injury or illness information and job characteristics—could boost predictive performance but are not typically available to program implementers.

Exhibit 2. Performance of predictive models

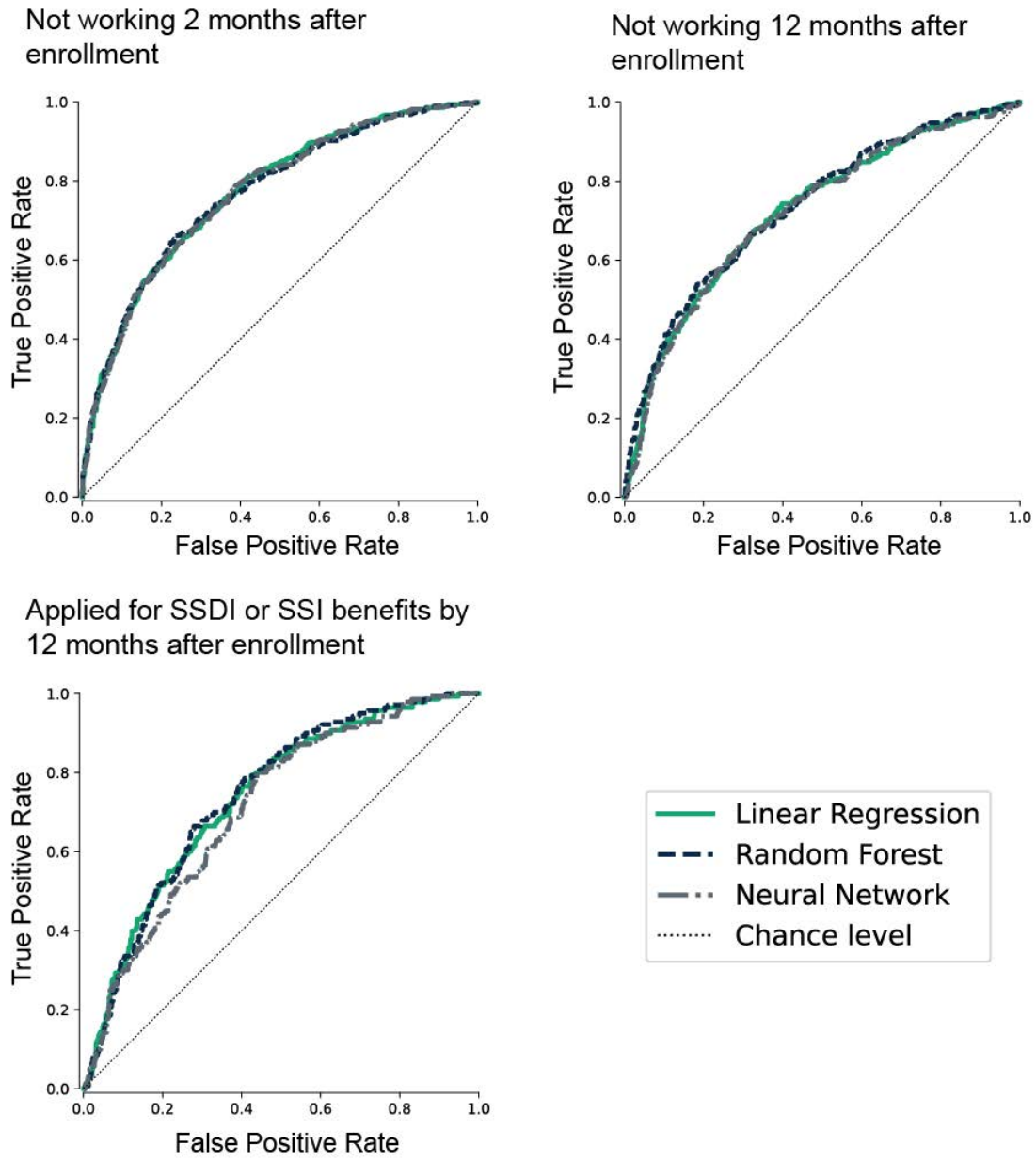
Outcome	AUC	Accuracy	Recall	F1
Not working 2 months after enrollment				
Linear regression	0.77	0.70	0.67	0.63
Random forest	0.77	0.71	0.70	0.64
Neural network	0.77	0.70	0.68	0.63
Not working 12 months after enrollment				
Linear regression	0.72	0.70	0.60	0.52
Random forest	0.73	0.69	0.62	0.52
Neural network	0.72	0.68	0.64	0.52
Applied for SSDI or SSI by 12 months after enrollment				
Linear regression	0.74	0.66	0.68	0.27
Random forest	0.74	0.72	0.62	0.29
Neural network	0.71	0.68	0.56	0.25

Source: RETAIN enrollment and survey data; SSA administrative data.

Note: The accuracy rate captures the share of enrollees for whom the model's prediction was correct. Recall is the true positive rate or the share of true positive instances that the model correctly identified. Precision is the share of all instances predicted as positive by the model that were indeed positive. The F1 score is the harmonic mean of precision and recall; it provides a single number to reflect the balance between precision and recall. If both precision and recall are high, the F1 score is high.

AUC = area under the receiver operating characteristic curve; SSA = Social Security Administration; SSDI = Social Security Disability Insurance; SSI = Supplemental Security Income.

Exhibit 3. ROC curves of predictive models, by outcome



Source: RETAIN enrollment and survey data; SSA administrative data.

2. Audience familiarity, interpretability, and scalability

In addition to predictive performance, we considered three other factors when selecting a model for further analysis: audience familiarity, interpretability, and scalability. Most consumers of policy-oriented quantitative research are familiar with linear regression, which is widely taught, well documented, and relatively easy to explain to nontechnical audiences. While machine learning methods are becoming more common in applied research, they remain less familiar to many policy stakeholders, due in part to their algorithmic complexity and the perception that they are less transparent than traditional parametric models like linear regression. Others have noted that policy audiences often prefer models that are easier to interpret and explain (Peet et al. 2022).

Interpretability is another advantage of linear regression. It produces coefficients that can be read as the change in the outcome for a one-unit change in the predictor, holding other variables constant—making it straightforward to understand and communicate relationships between variables. Most machine learning models, by contrast, do not yield such directly interpretable coefficients. Instead, they often provide measures of variable importance, which, while informative, describe relative contributions to predictions rather than precise effect sizes.

Finally, scalability matters for practical deployment. Complex machine learning models such as random forests and neural networks can require greater computing resources and longer training times than linear regression. If a predictive model were to be deployed for eligibility determination or caseload targeting in a future RETAIN-like program, linear regression would offer the advantage of being less computationally demanding and easier to implement with available technical capacity.

As noted above, predictive performance was similar across the three model types we evaluated. Given the comparable accuracy, combined with its greater audience familiarity, interpretability, and scalability, we selected the linear regression classifier for further analysis.

C. Results

1. Characteristics that predict non-return to work and applications for SSDI or SSI

After selecting the linear regression model as the primary model for further analysis, we used it to examine which baseline characteristics were most predictive of the three outcomes: not working two months after enrollment, not working 12 months after enrollment, and applying for SSDI or SSI benefits in the 12 months after enrollment.

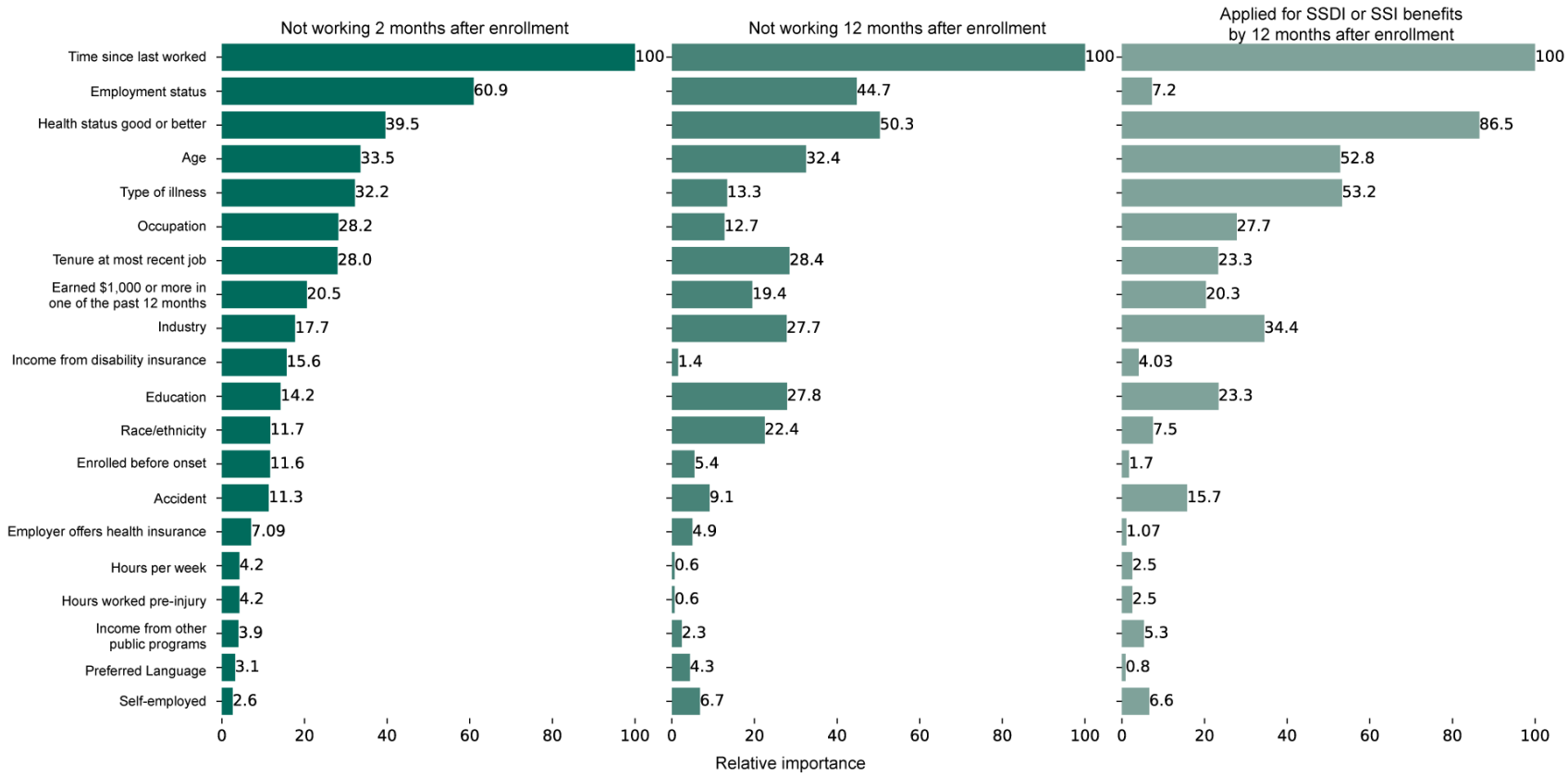
In the next section, we assess the relative importance of independent variables in predicting outcomes. We report a statistic called the aggregate Shapley value, which represents the average marginal contribution of an independent variable or category of independent variables to a prediction, computed across all combinations of input variables. These values are model-agnostic and can capture interaction effects and nonlinear relationships. For categorical variables, we summed Shapley values across categories to generate an overall score.[13]

[13] Shapley values are additive. At an observation level, the sum of Shapley values across a combination of independent variables represents that combination of variables' contribution to explaining the difference between the average outcome across the sample and that individual observation's predicted outcome.

a. Understanding predictors of non-return to work and SSA applications via Shapley values

Exhibit 4 presents the aggregated Shapley values of the 20 most predictive variables for each outcome, scaled so that the most predictive variable for a given outcome has a value of 100, and ordered by their relative importance for predicting *not working two months after enrollment*. Unsurprisingly, there is substantial overlap across outcomes: factors that strongly predict short-term work status often also predict long-term work status and the likelihood of applying for SSDI or SSI benefits. The time since last worked stands out as the single most important predictor across all outcomes, with age and self-reported health status also consistently ranking among the top five.

Exhibit 4. Relative importance of the top 20 predictors in linear regression models predicting three outcomes



Note: Predictors are sorted by relative importance for predicting the outcome of not working two months after enrollment. Relative importance is computed as the mean absolute Shapley value for each predictor over all observations, scaled such that the value for the strongest predictor for each outcome is 100.

Some predictors differ in importance by outcome. Employment status at enrollment is highly predictive of both short- and long-term work outcomes but plays a much smaller role in predicting SSA applications. The type of illness or injury is a strong predictor for benefit applications and short-term non-return to work, but its influence fades at the 12-month mark. Similarly, industry category is a more important predictor for long-term work outcomes and SSA claims than it is for short-term work outcomes.

Several baseline characteristics—such as sex, veteran status, preferred language, health insurance type, workers' compensation status, and hours worked—did not appear among the top predictors for any outcome. Although these factors may still shape individual experiences, their predictive value in this analysis was limited.

b. Factors associated with non-return to work two months after enrollment

One advantage of using a linear regression model is that it produces coefficients that are straightforward to interpret. Below, we present findings from the fit linear regression model, focusing on *not returning to work within two months of enrollment*, as measured by the early follow-up survey.

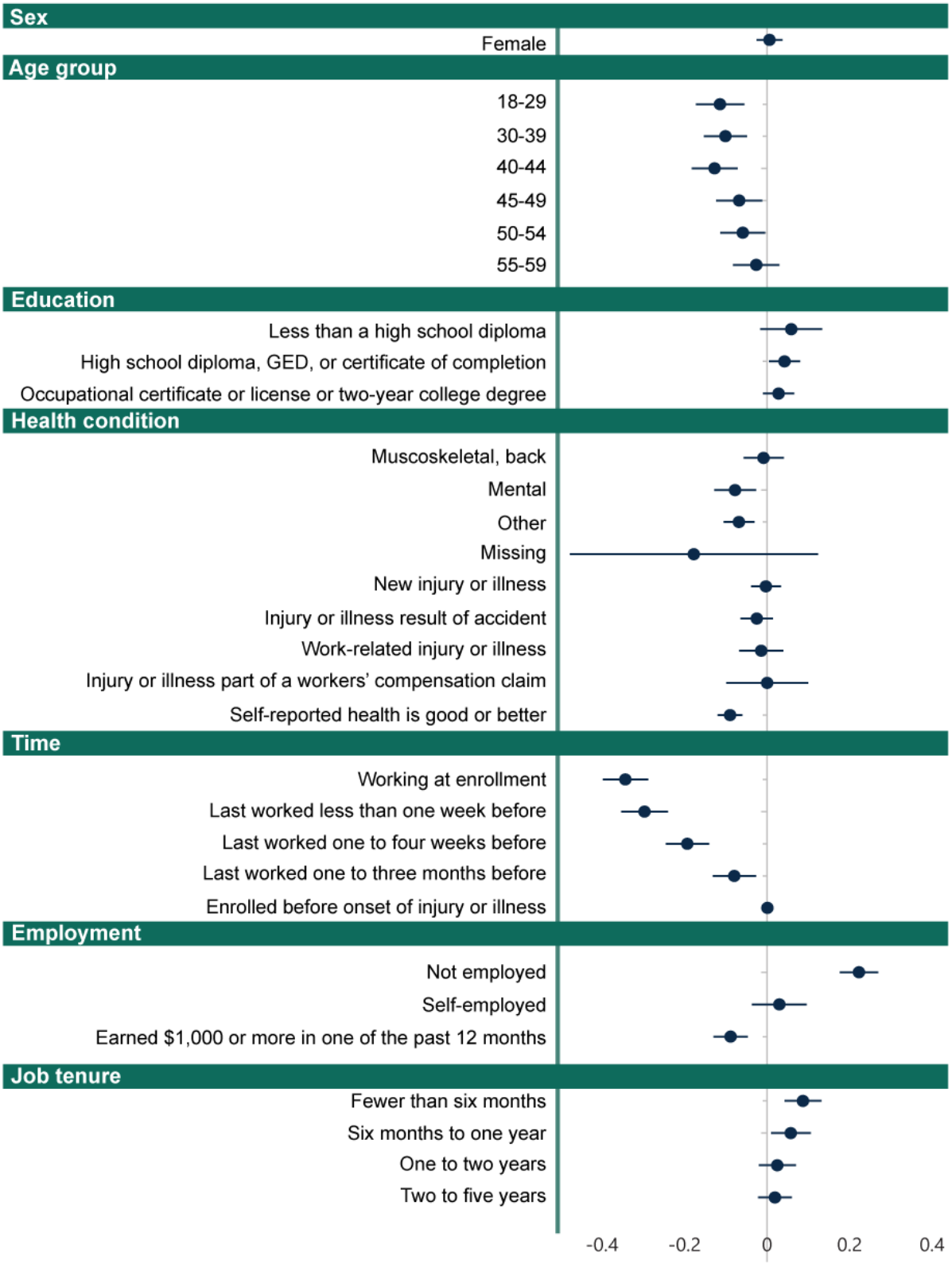
The fit regression coefficients (Exhibit 5) align closely with the Shapley value estimates—coefficients that are statistically significantly different from zero are for predictors in categories with Shapley values indicating high relative importance (Exhibit 4). Enrollees who were not employed at enrollment, had been out of work longer before enrollment, and had shorter job tenures were more likely to still be out of work at follow-up. These findings suggest that stronger pre-injury attachment to an employer may support return-to-work, possibly by enabling access to leave, accommodations, or job protection.

Diagnosis type was also strongly associated with outcomes. Compared with enrollees with musculoskeletal non-back injuries, those with mental health conditions or conditions categorized as "other" were less likely to be out of work at the two-month mark. This pattern may reflect the possibility that certain musculoskeletal injuries—particularly acute cases—can be more physically limiting or require longer recovery times than some mental health or miscellaneous conditions in the "other" category.

Socioeconomic and health-related factors also played a meaningful role. Higher educational attainment and having earned at least $1,000 in one of the prior 12 months were associated with a lower likelihood of being out of work at two months. Similarly, participants who rated their health more favorably were less likely to be out of work.

While Shapley values and regression coefficients help identify the characteristics most associated with return-to-work or benefit application outcomes, they represent associations that may be influenced by unobserved factors. Some characteristics flagged as predictive may be correlated with other, unmeasured attributes that themselves affect outcomes. For example, while pre-injury industry of employment appears to predict benefit applications, industry may be a proxy for unobserved traits such as skill level or socioeconomic status—meaning that the estimated effect for industry partly captures the influence of these unmeasured factors.

Exhibit 5. Select linear regression coefficients (outcome: not working two months after enrollment)



2. Predicting need for RETAIN services

We defined "high need" for RETAIN services as the likelihood of not returning to work within two months after enrollment (specifically, by the early follow-up survey). To identify people with these needs, we used the estimated coefficients from the linear regression model to calculate the predicted probability of not returning to work within two months for all enrollees, across both treatment and control groups, assuming no RETAIN services were provided. We then applied the threshold that maximized Youden's index to classify participants as either high need or not. Exhibit 6 presents the share of enrollees predicted to be high need within each of the five RETAIN programs. This share was lowest in Vermont RETAIN (36 percent) and highest in Kentucky (50 percent).

Exhibit 6. Percentage of enrollees predicted to be high need, by program

RETAIN program	High need percentage
All programs combined	**42.6**
RETAINWOR*KS*	38.1
RETAIN Kentucky	49.5
Minnesota RETAIN	45.7
Ohio RETAIN	37.7
Vermont RETAIN	36.0

a. Characteristics of enrollees predicted to be high need and not high need

We compared the baseline characteristics of enrollees predicted to be high need with those predicted not to be high need (Exhibit 7). Several notable differences emerged. In terms of demographic characteristics, high-need enrollees were less likely to be women (56 percent compared with 63 percent), less likely to be White (69 percent versus 81 percent), and more likely to be Black (19 percent versus 10 percent). They also had lower educational attainment: 59 percent had a high school diploma or less compared with 36 percent of those not predicted to be high need.

Health-related characteristics also differed between the groups. High-need enrollees were less likely to have a non-back musculoskeletal condition and more likely to have a mental health or back-related musculoskeletal condition. They were also less likely to report their health as very good or excellent (53 percent compared with 73 percent).

Labor force characteristics showed especially pronounced differences. High-need enrollees had much lower attachment to the labor force, with 46 percent not employed at enrollment compared with less than 1 percent of others. Only 7 percent had worked within a week of enrollment compared with 75 percent of those not predicted to be high need, and their most recent job tenure was typically shorter, with 34 percent having been employed for six months or less compared with 14 percent of others. The two groups worked a similar number of hours per week before their injury, though high-need enrollees worked slightly fewer hours on average (37 compared with 39). The distribution of occupations also differed: high-need enrollees were less likely to have worked in management, professional, or related occupations (17 percent versus 42 percent) and more likely to have been employed in production, transportation, or material-moving occupations (23 percent versus 12 percent) and service occupations (43 percent versus 32 percent).

Economic well-being indicators also favored the non–high-need group. About two-thirds of high-need enrollees earned at least $1,000 in one of the past 12 months, compared with 91 percent of others. High-need enrollees also relied more heavily on income from public programs beyond those queried in the one-year follow-up survey, with 13 percent reporting such income compared with 3 percent among those not predicted to be high need.

Exhibit 7. Baseline characteristics of RETAIN enrollees, by predicted level of need

Variable	All (A)	Not high need (B)	High need (C)	Difference (B–C)	*p*-value
Demographic characteristics					
Sex					
Female	60.1	63.0	56.2	6.9	0.00***
Age					
18–29	16.5	15.9	17.3	-1.4	0.04**
30–39	23.4	23.7	22.9	0.8	0.30
40–44	13.3	14.5	11.5	3.0	0.00***
45–49	12.6	12.8	12.4	0.3	0.58
50–54	13.1	13.3	12.8	0.5	0.42
55–59	11.5	11.0	12.2	-1.2	0.04**
60 and older	9.6	8.7	10.8	-2.1	0.00***
Race and ethnicity					
Hispanic	5.2	4.5	6.1	-1.6	0.00***
White, non-Hispanic	75.7	80.6	69.1	11.5	0.00***
Black, non-Hispanic	13.6	9.8	18.8	-9.0	0.00***
Asian, non-Hispanic	0.9	1.0	0.6	0.4	0.01**
More than one race	3.2	2.7	3.8	-1.1	0.00***
Other, non-Hispanic	0.9	0.8	1.0	-0.2	0.25
Missing	0.6	0.6	0.6	0.0	0.79
Preferred language					
English	98.9	99.5	98.2	1.3	0.00***
Spanish	0.5	0.2	0.9	-0.8	0.00***
Other	0.6	0.3	0.9	-0.6	0.00***
Education					
Less than a high school diploma	4.7	2.3	7.9	-5.6	0.00***
High school diploma, GED, or certificate of completion	41.1	33.4	51.5	-18.1	0.00***
Occupational certificate, license, or two-year college degree	27.0	27.8	26.0	1.8	0.02**
Four-year college or postgraduate degree	27.2	36.5	14.7	21.9	0.00***
Injury or illness characteristics					
Type of injury or illness					
Musculoskeletal, back	10.2	9.0	11.9	-2.9	0.00***
Musculoskeletal, non-back	47.4	50.1	43.7	6.3	0.00***

Variable	All (A)	Not high need (B)	High need (C)	Difference (B–C)	*p*-value
Mental health condition	15.4	12.1	19.9	-7.7	0.00***
Other	26.6	28.4	24.2	4.2	0.00***
Missing	0.4	0.4	0.3	0.1	0.25
Other characteristics of injury or illness					
New injury or illness	37.8	38.3	37.0	1.3	0.13
Injury as result of an accident	40.6	42.7	37.6	5.1	0.00***
Work-related injury or illness	10.1	10.8	9.2	1.6	0.00***
Injury or illness as part of a workers' compensation claim	3.3	3.9	2.5	1.4	0.00***
Time between injury or illness and enrollment					
Total days	41.7	37.2	47.8	-10.61	0.00***
Enrolled before onset of injury or illness	1.9	1.7	2.1	-0.4	0.09*
Missing	0.1	0.2	0.1	0.1	0.06*
Work history and other sources of income					
Employment status at enrollment					
Not employed	19.9	0.3	46.3	-46.0	0.00***
Self-employed	5.1	5.5	4.6	0.9	0.02**
Employed	75.0	94.2	49.0	45.2	0.00***
Time since last worked at enrollment					
Working at enrollment	28.7	48.2	2.4	45.8	0.00***
Last worked less than one week before	17.4	26.5	5.1	21.5	0.00***
Last worked one to four weeks before	24.5	22.7	26.9	-4.2	0.00***
Last worked one to three months before	16.2	2.4	34.7	-32.3	0.00***
Last worked more than three months before	13.3	0.2	30.9	-30.7	0.00***
Tenure at most recent job					
Less than six months	22.6	14.0	34.2	-20.2	0.00***
Six months to one year	13.5	11.5	16.1	-4.5	0.00***
One to two years	14.2	15.6	12.4	3.2	0.00***
Two to five years	17.9	20.1	14.9	5.2	0.00***
More than five years	31.8	38.8	22.4	16.3	0.00***
Occupational classification of pre-injury or pre-illness job					
Management, professional, or related	31.3	42.2	16.5	25.7	0.00***
Service	36.6	31.8	43.0	-11.2	0.00***
Sales and office	8.6	9.3	7.6	1.7	0.00***
Natural resources, construction, or maintenance	7.0	4.9	9.8	-5.0	0.00***
Production, transportation, or material moving	16.5	11.7	22.9	-11.2	0.00***
Missing	0.1	0.2	0.1	0.1	0.06*
Receipt of income other than earnings					
Veterans benefits	1.2	1.4	0.9	0.6	0.00***

Variable	All (A)	Not high need (B)	High need (C)	Difference (B–C)	*p*-value
Workers' compensation	0.9	1.1	0.5	0.6	0.00***
Employer-provided or other private disability insurance	11.6	9.7	14.2	-4.6	0.00***
Other public programs	7.0	2.9	12.6	-9.8	0.00***
Additional baseline characteristics					
Hours per week usually worked before injury or illness	38.3	38.9	37.4	1.44	0.00***
Earned $1,000 or more in one of the past 12 months	81.1	90.2	68.8	21.4	0.00***
Applied for or received SSDI or SSI in the past three years	2.0	1.1	3.1	-1.9	0.00***
Covered by health insurance	95.0	95.6	94.2	1.4	0.00***
Veteran status	4.6	5.0	4.1	0.9	0.02**
Health status good or better	64.7	74.3	51.9	22.4	0.00***
Industry					
Agriculture or mining	1.6	1.4	1.8	-0.4	0.08*
Construction or utilities	7.0	5.7	8.9	-3.2	0.00***
Manufacturing	11.6	10.6	13.1	-2.5	0.00***
Retail trade and wholesale trade	10.0	9.3	10.8	-1.4	0.01***
Transportation and warehousing	5.2	3.7	7.1	-3.4	0.00***
Information	2.1	2.5	1.5	1.0	0.00***
Finance and insurance or real estate and leasing	3.0	4.0	1.7	2.2	0.00***
Professional, management or administrative services	6.7	8.3	4.5	3.8	0.00***
Education	8.2	10.7	4.8	5.9	0.00***
Health care	20.6	23.4	16.9	6.4	0.00***
Accommodation and food services or arts and entertainment	12.7	8.9	17.8	-8.9	0.00***
Other services	7.7	7.1	8.4	-1.2	0.01***
Public administration	3.6	4.3	2.6	1.7	0.00***
missing	0.1	0.2	0.1	0.1	0.06*
Total number of enrollees	**12,638**	**7,253**	**5,385**		

Source: RETAIN enrollment data.

Note: The *p*-value is based on a two-tailed *t*-test.

*/**/*** Difference is significantly different from zero (p-value is less than .10/.05/.01) using a two-tailed t-test.

SSDI = Social Security Disability Insurance; SSI = Supplemental Security Income.

b. Impacts of RETAIN on enrollees predicted to be high need and not

Exhibit 8 presents the estimated impacts of RETAIN for enrollees predicted to be high need and those predicted not to be, by program. For RETAINWOR*KS*, the only program that generated positive impacts across the three primary outcomes, we find favorable results for both subgroups on employment, earnings, and applications for SSDI or SSI. Although the differences in impacts across the two subgroups are not statistically significant, they suggest that RETAINWOR*KS* was more effective in improving outcomes for the high-need group.

Among high-need enrollees, employment in the fourth quarter after enrollment increased by 11.0 percentage points, compared with an increase of 6.2 percentage points among those not predicted to be high need. Earnings gains were similar across the groups—$3,083 for high-need participants and $3,856 for those not predicted to be high need. RETAINWOR*KS* also reduced applications for SSDI or SSI by 7.7 percentage points among high-need participants, from a control group rate of 24 percent; it reduced applications by 3.9 percentage points in the not–high-need group, whose control group rate was 9.9 percent.

Our findings for other RETAIN programs were mixed. Minnesota RETAIN reduced disability applications in the not-high-need group by 1.6 percentage points and Vermont RETAIN reduced earnings in the high-need group by $4,268. RETAIN Kentucky and Ohio RETAIN did not impact employment, earnings, or disability benefit applications for enrollees in either the high-need or not-high-need groups.

These findings underscore the complexity of identifying and serving people who might benefit most from RETAIN. Not–high-need participants generally have stronger labor force attachment (as shown in Exhibit 7), which may make them more responsive to RETAIN's supports in maintaining or regaining employment. High-need participants, by contrast, tend to face greater barriers to work.

Exhibit 8. One-year impacts on enrollee outcomes, by predicted level of need

Program	Outcome measure	Not high need					High need					*p*-value for subgroup difference
		Control mean	Impact	*p*-value	Treatment N	Control N	Control mean	Impact	*p*-value	Treatment N	Control N	
RETAINWOR*KS*	Employed in the fourth quarter after enrollment	71.0	6.2**	0.04	307	281	66.3	11.0**	0.02	196	167	0.40
	Earnings during Quarters 1-4 after enrollment	29,953	3,856***	0.00	307	281	28,681	3,083*	0.07	196	167	0.72
	Applied for SSDI or SSI	9.9	-3.9*	0.07	307	281	24.2	-7.7*	0.08	196	167	0.43
RETAIN Kentucky	Employed in the fourth quarter after enrollment	69.5	-1.5	0.47	800	742	56.3	2.9	0.25	791	705	0.18
	Earnings during Quarters 1-4 after enrollment	26,397	-306	0.75	800	742	20,876	669	0.45	791	705	0.45
	Applied for SSDI or SSI	4.9	0.6	0.62	800	742	17.7	0.5	0.80	791	705	0.96
Minnesota RETAIN	Employed in the fourth quarter after enrollment	72.7	1.6	0.36	843	864	61.8	-3.3	0.17	724	708	0.10
	Earnings during Quarters 1-4 after enrollment	40,931	-112	0.92	843	864	31,800	1,003	0.38	724	708	0.49
	Applied for SSDI or SSI	4.9	-1.6*	0.08	843	864	14.9	0.3	0.86	724	708	0.35
Ohio RETAIN	Employed in the fourth quarter after enrollment	75.9	0.3	0.83	1,394	1,397	75.7	-1.6	0.45	847	850	0.45
	Earnings during Quarters 1-4 after enrollment	41,322	-976	0.16	1,394	1,397	35,762	962	0.25	847	850	0.07†
	Applied for SSDI or SSI	3.3	0.3	0.63	1,394	1,397	10.0	0.2	0.88	847	850	0.98
Vermont RETAIN	Employed in the fourth quarter after enrollment	67.2	-2.2	0.56	285	205	61.3	-6.6	0.27	151	126	0.54
	Earnings during Quarters 1-4 after enrollment	27,936	-1,013	0.58	285	205	28,393	-4,268**	0.04	151	126	0.23
	Applied for SSDI or SSI	8.0	-0.8	0.71	285	205	15.7	-1.5	0.73	151	126	0.88

Source: RETAIN enrollment data; one-year follow-up survey; SSA data; state Unemployment Insurance wage records.

Note: Outcome measures reflect enrollees' outcomes in the year after enrollment. For each subgroup, this table shows the regression-adjusted means for the control group (the estimate of the counterfactual) and the regression-adjusted estimates of each program's impacts. To calculate the adjusted mean for the treatment group, add the impact estimate and the adjusted mean for the control group. The *p*-value for all outcomes is based on a two-tailed *t*-test.

*/**/*** Impact estimate is significantly different from zero at the .10/.05/.01 level.

†/††/††† Impact estimates for subgroups are significantly different from each other (*p*-value is less than .10/.05/.01) using an adjusted Wald test.

N = sample size; SSA = Social Security Administration; SSDI = Social Security Disability Insurance; SSI = Supplemental Security Income.

D. Discussion

This analysis tested whether a predictive model could identify RETAIN enrollees who were less likely to return to work quickly without intervention and whether these "high-need" participants experienced different program impacts. Using baseline data from the five RETAIN programs, we found that a simple linear regression model predicted outcomes about as well as more complex methods, with acceptable but not exceptional performance. This level of performance does not support using the model to determine eligibility, but it does offer insights into who is most at risk of delayed return to work—information that can inform program planning and resource allocation.

1. Predictive insights and program impacts

The predictive models highlighted a consistent set of baseline factors, especially time since last worked, employment status, and health status, that were strongly associated with return-to-work outcomes. Enrollees who had been out of work for less time, were employed at enrollment, and were healthier were more likely to return to work without assistance within 2 months of enrollment. In contrast, those with weaker labor force attachment, unemployed at enrollment, or less healthy were less likely to do so and thus be classified as high need.

Comparisons of impact estimates for high- and not–high-need enrollees show that both groups can benefit from RETAIN services, though the nature and size of impacts might differ by outcome. In KANSASWOR*KS*, the only program with broadly positive impacts, both groups saw improvements in employment and earnings gains along with a reduction in disability applications. These findings underscore that an effective SAW/RTW program can be beneficial even for participants not flagged as high need. Although not statistically significantly different from each other, the impact estimates for employment and disability applications were larger in magnitude for the high-need group compared to the not-high-need group. This divergence suggests that predictive classification can highlight where programs might have the greatest impact.

Taken together, these findings indicate that predicting high need can help identify participants who are more likely to apply for disability benefits and offer greater potential for reducing such applications and subsequent government expenditures. This distinction matters for program design. We did not examine whether different service types or intensities produced the observed subgroup differences. However, the findings suggest that programs could use predictive scores as one input when deciding how to tailor services. In practice, this might mean allocating more intensive support to participants facing greater challenges while still offering core services to others.

Importantly, the pattern of subgroup impacts when looking across the five programs suggests that targeting is only valuable if the underlying services are effective. Precise identification of high-need participants is unlikely to improve outcomes if programs cannot deliver interventions that address their needs.

2. Practical approach to predictive modeling

In this analysis, more complex machine learning models offered no meaningful advantage over a straightforward linear regression approach. The regression model's predictive performance was comparable to that of random forests and neural networks across all outcomes, reinforcing that in

contexts with structured intake data, simpler models can perform just as well as more sophisticated alternatives.

This finding is consistent with other applied research showing that regression methods can match or exceed the performance of machine learning when relationships are largely linear. A systematic review of clinical prediction studies, for example, found no overall performance advantage of complex machine learning methods over well-specified regression models when using structured data (Christodoulou et al. 2019). In practice, the absence of a performance edge for machine learning suggests that the additional complexity might not be justified in policy settings where clarity and interpretability are valued.

Because linear regression is transparent, familiar to policy audiences, and relatively easy to apply at scale, it represents a practical choice for program planning and management. Stakeholders are more likely to trust results that can be readily explained, and straightforward models both reduce the risk that analytic findings will be considered too opaque or technical and are easier to update as new information becomes available. The ability to interpret coefficients also helps connect predictions to tangible program design decisions, such as which participant characteristics warrant closer attention.

At the same time, the models' moderate performance in this study highlights the limits of relying only on intake data. Incorporating richer information—such as more specific details on injury type, job demands, or local labor market conditions—might improve predictions. If such information is not readily available, collecting it could increase administrative burden and discourage program participation. This trade-off between predictive performance and information needs suggests that predictive models might be best suited for informing program planning and tailoring service delivery after intake, rather than as strict eligibility screens. Their value lies in supporting better resource allocation, not in determining who should or should not be served.

3. Implications for funders and implementers

For potential funders of future SAW/RTW programs, our findings suggest that predictive classification can be a useful planning tool, but its value depends critically on the effectiveness of the underlying services. When services generate consistent gains, targeting can help ensure those benefits are available to participants least likely to succeed without support. By contrast, when services show little or no impact, refining how participants are classified does little to improve outcomes and might distract from more fundamental questions about service design. Hence, investment in predictive analytics should follow, rather than precede, evidence that the program model itself is working.

For implementers, predictive models are best used to guide service intensity rather than to restrict eligibility. Programs often face pressure to ration resources, but the findings from RETAINWORKS suggest that excluding participants with lower predicted need can be shortsighted if they still realize meaningful gains from support. A more constructive approach could be to stratify caseloads, offering more intensive supports—such as more frequent communications with RTW coordinators, hands-on coordination of workplace accommodations, and extended service eligibility periods—to those with the highest predicted support needs, while still maintaining a baseline level of assistance for others. This strategy would acknowledge variation in participants' circumstances without cutting off access.

High-need participants might benefit from service adaptations that respond directly to the challenges identified in the predictive models. For example, those with weaker labor force attachment or shorter job tenure might need more sustained employer engagement, while those reporting poorer health may require stronger medical coordination. However, further research is needed to understand whether a different version of RETAIN—for example with more intensive supports—could produce similar or better results for similar costs. Testing such targeted approaches could help identify how predictive tools can be most effectively integrated with service delivery and maximize net benefits for both workers and the government.

Finally, the reliability of predictive models in applied settings warrants careful attention. Our model drew on pooled data from all five RETAIN states, which improves generalizability within the demonstration. Even so, predictive accuracy can vary across cohorts, program environments, or participants with different characteristics, and our analysis did not formally test for such differences. This limitation means that performance for some groups might be stronger or weaker than the pooled results suggest. Funders and implementers considering broader use of predictive tools should therefore plan to validate models regularly, both across time and across different participant groups. Doing so will help ensure that predictive classification supports resource allocation without introducing systematic blind spots or misclassifications.

This page has been left blank for double-sided copying.

References

Bureau of Labor Statistics. "Short-Term and Long-Term Disability Insurance for Civilian Workers in 2020." *TED: The Economics Daily*, December 31, 2020. https://www.bls.gov/opub/ted/2020/short-term-and-long-term-disability-insurance-for-civilian-workers-in-2020.htm.

Ben-Shalom, Yonatan, Ignacio Martinez, and Mariel McKenzie Finucane. "Who Is at Risk of Workforce Exit Due to Disability? State Differences in 2003–2016." *Journal of Survey Statistics and Methodology,* vol. 9, no. 2, 2021, pp. 209–230.

Billings, John, Ian Blunt, Adam Steventon, Theo Georghiou, Geraint Lewis, and Martin Bardsley. "Development of a Predictive Model to Identify Inpatients at Risk of Re-Admission within 30 Days of Discharge (PARR-30)." *BMJ Open*, vol. 2, no. 4, 2012, article e001667.

Bourbonniere, Annette M., and David R. Mann. "Benefit Duration and Return to Work Outcomes in Short Term Disability Insurance Programs: Evidence from Temporary Disability Insurance Program." *Journal of Occupational Rehabilitation,* 28, no. 4, 2018, pp. 597–609.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions." In *Conference on Fairness, Accountability and Transparency*, pp. 134–148. PMLR, 2018.

Christodoulou, Evangelia, Jie Mao, Gary S. Collins, Ewout W. Steyerberg, Johanna Verbakel, and Ben Van Calster. "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models." *Journal of Clinical Epidemiology,* vol. 110, 2019, pp. 12–22.

Contreary, Kara, and Todd Honeycutt. "Who's at Risk of Entering Social Security Disability Insurance? A Comparison of Application and Allowance Rates for Groups of At-Risk Individuals." *Journal of Vocational Rehabilitation,* vol. 53, no. 3, 2020, pp. 341–352.

Contreary, Kara, Yonatan Ben-Shalom, and Brian Gifford. "Using Predictive Analytics for Early Identification of Short-Term Disability Claimants Who Exhaust Their Benefits." *Journal of Occupational Rehabilitation,* vol. 28, no. 4, 2018, pp. 584–596.

Croake, Sarah, Moriah Bauman, Yonatan Ben-Shalom, Jill Berk, and Meagan Ager. "The RETAIN Demonstration: State Programs' Approaches to Recruiting Potential Enrollees." Disability Policy Issue Brief. Mathematica, 2023.

Fawcett, Tom. "An Introduction to ROC Analysis." *Pattern Recognition Letters,* vol. 27, no. 8, 2006, pp. 861–874.

Hurley, Dan. "Can an Algorithm Tell When Kids Are in Danger?" *The New York Times Magazine*, January 2, 2018.

Mandrekar, Jayawant N. "Receiver Operating Characteristic Curve in Diagnostic Test Assessment." *Journal of Thoracic Oncology,* vol. 5, no. 9, 2010, pp. 1315–1316.

Neuhauser, Frank, Yonatan Ben-Shalom, and David Stapleton. "Early Identification of Potential SSDI Entrants in California: The Predictive Value of State Disability Insurance and Workers' Compensation Claims." *Journal of Occupational Rehabilitation,* vol. 28, no. 4, 2018, pp. 574–583.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, vol. 366, no. 6464, 2019, pp. 447–453.

Patnaik, Ankita, Isabel Musse, Jillian Berk, Karen Katz, Monica Farid, and Yonatan Ben-Shalom. "The RETAIN Demonstration: Impacts Two Months after Enrollment." Mathematica, 2025.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research,* vol. 12, 2011, pp. 2825–2830.

Peet, Evan D., Brian G. Vegetabile, Matthew Cefalu, Joseph D. Pane, and Cheryl L. Damberg. "Machine Learning in Public Policy." Santa Monica, CA: RAND Corporation, 2022. https://www.rand.org/pubs/perspectives/PEA828-1.html.

Trutko, John, Alex Trutko, Andrew Clarkwest, Phomdaen Souvanna, Jacob A. Klerman, Amanda Briggs, Shayne Spaulding, Ian Hecker, Ayesha Islam, Batia Katz, Molly Scott, and Demetra Nightingale. "RESEA Program Strategies: State and Local Implementation." Abt Associates, 2022.

Whicher, Danielle, Emma Pendl-Robinson, Kyla Jones, and Allon Kalisher. "Avoiding Racial Bias in Child Welfare Agencies' Use of Predictive Risk Modeling." Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, 2022.

Youden, W.J. "Index for Rating Diagnostic Tests." *Cancer,* vol. 3, no. 1, 1950, pp. 32–35.

This page has been left blank for double-sided copying.

Mathematica Inc.

Our employee-owners work nationwide and around the world.

Find us at **mathematica.org** and **edi-global.com**.

Mathematica, Progress Together, and the "spotlight M" logo are registered trademarks of Mathematica Inc.