Functional improvement is a better predictor of steady work than medical improvement for individuals with mental health conditions

Joshua C. Chang[1*], Julia Porcino[1*], Elizabeth Marfeo[1,2], Larry Tang[1,3], Howard Goldman[1], Elizabeth Rasch[1]

[1] Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, Maryland, United States of America
[2] Department of Occupational Therapy & Community Health, Tufts University, Boston, Massachusetts, United States of America
[3] Department of Statistics and Data Science, National Center for Forensic Science, University of Central Florida, Orlando, Florida, United States of America

*Corresponding authors

E-mails: josh.chang@nih.gov, julia.porcino@nih.gov

Abstract

The Supported Employment Demonstration (SED) offered vocational and mental health services to recently denied disability benefit applicants with mental health conditions, along with other multiple co-morbidities, to evaluate the impact of evidence-based interventions on fostering employment and downstream benefits such as self-sufficiency, improved quality of life, and improved mental health. Using the SED public use file, we analyzed work outcomes for the study participants in relation to functional improvement, as measured by the Work Disability Functional Assessment Battery, vs. medical improvement. Using both Bayesian logistic regression models and neural networks, we found that functional improvement is a better predictor of steady work than medical improvement.

Introduction

Employment and economic stability are critical social determinants of health and a key component of meeting sustainable development goals (1). Labor force participation is known to have a positive effect on health, while conversely, loss of employment can have a negative impact as observed during the COVID-19 pandemic (2–7). Participation in employment is particularly advantageous for individuals with mental health and physical limitations, as it fosters a sense of purpose, supports economic independence, and enhances emotional well-being. People with disabilities tend to have poorer health outcomes as well as lower rates of employment than their non-disabled peers (8). Developing policies and interventions to help people with disabilities gain or maintain employment is critical for meeting public health goals and

promoting health and well-being. Current models of disability incorporate a broad biopsychosocial perspective on factors that drive health and participation in social roles such as work. There are two key components to be able to apply these models to support programs for individuals with disabilities. First, it is important to know which factors inform and predict labor force participation. Second, there must be appropriate measures for these factors incorporated into research, policy, and support programs. In this work, we focus on the role of function and whether a self-report measure of whole-person function contributes to the prediction of employment participation and outcomes.

The Work Disability Functional Assessment Battery (WD-FAB)

The Work Disability Functional Assessment Battery (WD-FAB) was developed to address gaps in data collection around function (9,10).The WD-FAB is a self-reported assessment of functional abilities that relate to work that provides scores in 8 scales across two domains of mental and physical functioning: Basic Mobility, Upper Body Function, Fine Motor Function, Community Mobility, Resilience & Sociability, Self-Regulation, Communication & Cognition, and Mood & Emotions. The WD-FAB is based on item response theory (IRT) and administered using computer adaptive testing (CAT) technology to make the assessment efficient and tailored to the respondent. IRT-CAT-based assessments use advanced psychometric methods to enhance the precision and efficiency of measurement by dynamically selecting the most relevant questions for each individual based on their previous responses. The WD-FAB has undergone rigorous development and psychometric testing, and recent efforts are now focused on applications of the WD-FAB across use cases. Previous work has considered whether certain thresholds or profiles of functioning are indicative of a person's ability to work (11). From a longitudinal perspective, we want to understand how changes in WD-FAB scores relate to an individual's ability to return to work and maintain steady work.

The Supported Employment Demonstration (SED)

The SED, a randomized controlled trial, aimed to understand whether providing work, behavioral, and health supports for recently denied Social Security disability benefit applicants can help such individuals achieve self-sufficiency. The study recruited individuals aged 18 to 49 who had a mental health impairment, alongside other co-morbidities, who were assigned to one of three study arms (full service, basic service, and usual service). Over the three-year study period, the study tracked work, quality of life, and income outcomes along with functioning information (provided by the WD-FAB), health condition, and health care utilization information. Functioning information was collected via the WD-FAB on an annual basis. Our goal was to understand the predictive power of functional improvement versus medical improvement on work outcomes with a particular focus on steady work (12–19) .

Materials and Methods

Data

This study used the SED Public Use File (SED-PUF), a research dataset provided by the US Social Security Administration (SSA) that contains information collected from 2,944 individuals who participated in the SED. The accompanying codebook (20) provides the full list of variables and basic statistics for each variable.

Statistical analyses

In this section we provide a high-level overview of our performed analyses. Please refer to the Supplemental Methods for more details.

We developed Bayesian hierarchical logistic regression models (21) to predict the annual odds of steady work for each individual in the SED study, where steady work refers to employment of at least half-time, as defined in the SED datasets. These models used three classes of covariates: demographic (age, employment history, education, housing status, vehicle access), medical – consisting of outpatient/emergency/inpatient utilization as well as assessments like the Drug Alcohol Screen Test (DAST (22)), Alcohol Use Disorders Identification Test (AUDIT (23–25)), Colorado Symptoms Index (CSI (26)), and function (WD-FAB mental and physical scores).

In predicting whether an individual had steady work in a given study year, we used both the baseline value of a predictor and the change from baseline for each predictor for that study year. Within our models, we scaled baseline values of predictors by subtracting the mean and dividing by the standard deviation. We scaled differences by subtracting the mean difference for a variable and scaling by the standard deviation of differences. Since Community Mobility scale scores had extreme missingness, which itself is informational, we omitted the scale scores in our modeling and instead used indicators for the presence of these scale scores.

The SED-PUF contained a substantial number of missing values, particularly for function measurements. Notably, Community Mobility - Driving and Public Transportation (transit) - scales are not administered to respondents who do not use those transportation modes. Largely, for the seven remaining functioning scales, we found that participants were missing either zero, seven, fourteen, or twenty-one measurements. With the goal of retaining as many study participants as possible in the analysis, we performed missing value imputation within our model by jointly predicting the missing values and marginalizing over them during overall model inference.

Fitting submodels

One main objective was to quantify the relative impact of function in predicting the likelihood of work, controlling for demographic differences. To this end, we also fitted submodels where we used the following sets of predictors: demographic only, demographic + function (omitting medical), demographic + medical (omitting functional), and demographic + medical + function (full model).

Beyond linear regression

For the sake of completeness, we also replicated the same analysis using two nonlinear modeling techniques. First, we fit Bayesian artificial neural networks to the same data, settling on a shallow model with a hidden layer size of 12, after finding it relatively optimal compared to layer sizes between two and twenty. Second, we fit a piecewise linear Bayesian generalized linear model that is in-effect a multilevel regression model with both random slopes and intercepts based on an additive decomposition method (28,29) for representing nonlinearity in model coefficients. In these models, each local region of the data is associated with a generalized linear regression model.

Model evaluation

We adjudicated the models on predictive accuracy using Bayesian leave one out (LOO) cross validation (CV). Specifically, we used Pareto-smoothed importance sampling (30,31) (PSIS) in order to estimate the LOO predictive distribution and log-likelihood for each observation modeled (31). Using these quantities we estimated leave-one-out cross-validated receiver operator characteristic (ROC) and precision-recall (PRC) curves (32) and computed the area under these curves as a basis of comparison.
We used the Python bayesianquilts (28,29) wrapper for Tensorflow-probability (33) to perform all statistical analyses.

Results

Descriptive statistics

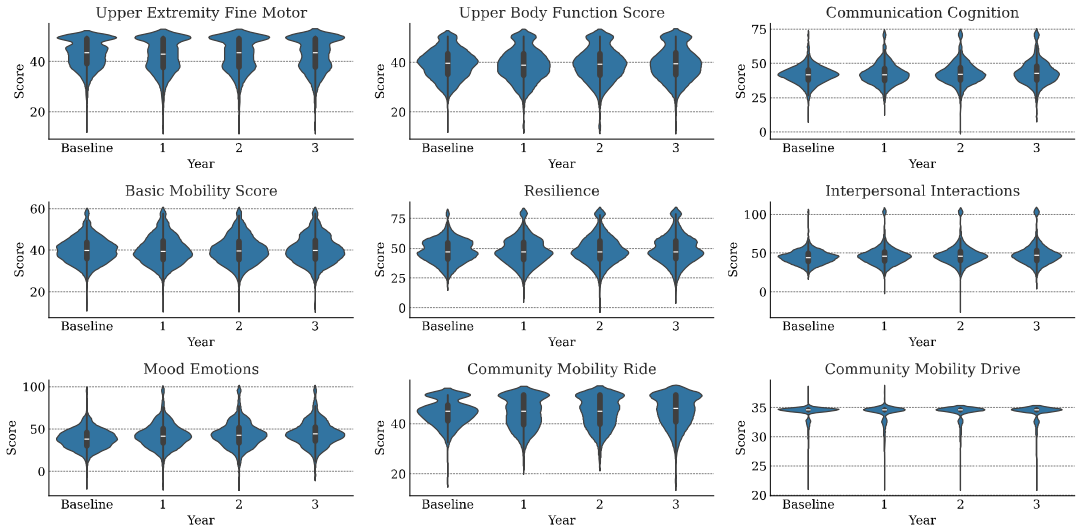

Figure 1. Distributions of WD-FAB scores for each study year by scale.

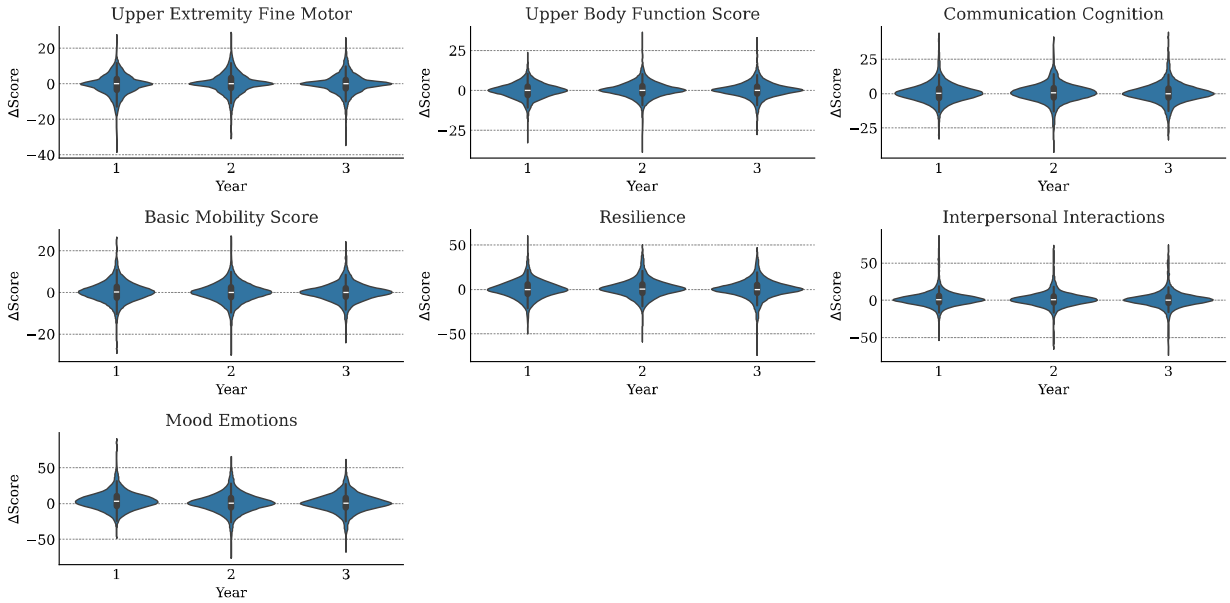

Figure 2. Distributions of WD-FAB changes in function (relative to baseline) for each study year. Community Mobility scale omitted.

Ignoring missingness, the annual means and standard deviations of functioning variables (along with their changes from baseline) are presented in **Error! Reference source not found.**. Corresponding statistics for medical variables are available in Table 2. The distributions of WD-FAB scores by year are presented in Fig 1**Error! Reference source not found.**. The distributions of change relative to baseline in these scores is presented in Fig 2**Error! Reference source not found.**. As seen in **Error! Reference source not found.**, approximately 60% of Driving and 80% of Public Transportation Community Mobility scores were missing. As mentioned in the Methods, this fact motivated us to only incorporate their presence rather than the scores themselves into our predictive models. Except for Community Mobility, most study participants were missing either zero, seven, or fourteen functioning measurements (Fig 3**Error! Reference source not found.**).

Table 1. Statistics mean (standard deviation) for outcome and functioning variables in the SED dataset, and their changes relative to baseline, ignoring missing values.

Variable	**MDC90**	**Overall**	**Baseline**	**Year 1**	**Year 2**	**Year 3**
Outcome						
Steady Work	N/A	0.173		0.155	0.192	0.173
Physical and mental function						
Fine Motor Function **Change (from baseline)**	8.2	42.7 (6.2) -0.1 (5.5)	43.0 (5.6) -	42.4 (6.4) -0.5 (5.6)	42.7 (6.5) 0.4 (5.5)	42.7 (6.5) -0.0 (5.2)
Upper Body Function **Change**	4.6	39.3 (6.4) 0.1 (4.8)	39.4 (5.8) -	39.0 (6.5) -0.2 (4.8)	39.3 (6.7) 0.3 (5.0)	39.4 (6.6) 0.2 (4.7)
Communication &	7.1	42.6 (7.8)	41.8 (6.6)	42.4 (7.8)	43.0 (8.4)	43.5 (8.7)

Cognition Change		0.6 (6.9)	-	0.7 (6.7)	0.7 (7.0)	0.5 (7.1)
Basic Mobility Change	4.7	40.2 (6.5) 0.2 (4.7)	40.2 (6.0) -	40.2 (6.7) 0.3 (4.8)	40.2 (6.6) 0.1 (4.9)	40.3 (6.7) 0.1 (4.5)
Resilience Change	8.2	48.3 (11.1) 0.6 (10.2)	47.6 (9.6) -	47.6 (11.3) 0.2 (10.1)	48.9 (12.0 1.2 (10.1)	49.4 (12.1) 0.5 (10.3)
Interpersonal Interactions Change	8.6	46.5 (12.1) 1.5 (11.6)	44.4 (8.8) -	46.5 (12.3) 2.1 (10.6)	47.6 (13.4) 1.3 (12.2)	48.6 (14.3) 1.0 (12.1)
Mood & Emotions Change	10.6	42.0 (14.1) 2.2 (13.1)	38.6 (12.0) -	42.5 (14.3) 4.0 (12.9)	43.6 (15.1) 1.3 (13.5)	44.6 (14.7) 1.0 (12.7)
Has Community Mobility (Ride) score	N/A	0.21	0.31	0.22	0.17	0.13
Has Community Mobility (Drive) score	N/A	0.40	0.51	0.36	0.38	0.36
Has Wheelchair Score	N/A	0.03	0.04	0.03	0.03	0.03

Table 2. Statistics (mean and standard deviation) for selected medical variables: overall, at Baseline, Years 1-3, and changes relative to baseline.

Variable	Overall	Baseline	Year 1	Year 2	Year 3
Colorado Symptom Index (CSI) **Change**	22.1 (12.4) -2.1 (10.3)	25.2 (11.2) -	22.2 (12.7 -3.0 (10.7	20.4 (12.7) -2.0 (10.2)	19.2 (12.5) -1.0 (9.7)
Body Mass Index (BMI) **Change**	31.5 (8.7) 0.1 (3.4)	31.1 (8.9) -	31.5 (8.6) -0.2 (3.7)	31.6 (8.5) 0.1 (3.2)	32.2 (8.8) 0.3 (3.3)
Inpatient Hospital Admissions **Change**	0.3 (0.8) -0.1 (0.9)	0.5 (0.9) -	0.3 (0.8) -0.2 (1.0)	0.2 (0.8) -0.1 (0.9)	0.2 (0.6) -0.1 (0.8)
Drug Abuse Screening Test (DAST) **Change**	0.8 (1.8) -0.2 (1.7)	1.1 (2.0) -	0.7 (1.6) -0.5 (1.9)	0.6 (1.5) -0.0 (1.6)	0.6 (1.6) 0.0 (1.5)
AUDIT **Change**	2.9 (5.1) -0.3 (4.5)	3.5 (5.7) -	2.7 (5.8) -0.7 (4.9)	2.5 (4.6) -0.1 (4.5)	2.5 (4.6) -0.1 (4.1)
Total Emergency room visits **Change**	0.9 (1.6) -0.2 (1.8)	1.2 (1.9) -	0.9 (1.6) -0.3 (1.0)	0.7 (1.5) -0.1 (0.9)	0.6 (1.3) -0.1 (0.8)
Emergency room drug visits **Change**	0.01 (0.2) -0.01 (0.2)	0.03 (0.3) -	0.01 (0.1) -0.02 (0.3)	0.005 (0.07) -0.01 (0.1)	0.003 (0.06) -0.002 (0.1)
Emergency room physical visits **Change**	1.0 (1.6) -0.2 (1.5)	1.0 (1.6) -	0.7 (1.4) -0.2 (1.7)	0.6 (1.5) -0.2 (1.5)	0.4 (1.1) -0.1 (1.4)
Emergency room mental visits **Change**	0.2 (0.8) -0.1 (0.7)	0.2 (0.8) -	0.1 (0.5) -0.1 (0.9)	0.1 (0.4) -0.0 (0.6)	0.0 (0.3) -0.0 (0.5)
Total inpatient nights **Change**	1.9 (6.6) -0.1 (0.9)	1.9 (6.6) -	1.0 (4.2) -0.9 (7.4)	1.0 (4.7) -0.0 (5.5)	0.9 (3.8) -0.3 (5.3)
Total ER visits **Change**	0.85 (1.6) -0.23 (1.8)	1.2 (1.9) -	0.88 (1.6) -0.29 (2.0)	0.71 (1.5) -0.21 (1.7)	0.59 (1.3) -0.18 (1.6)
Total ER physical visits	0.67 (1.4) -0.19 (1.5)	0.98 (1.6) -	0.73 (1.4) -0.25 (1.7)	0.56 (1.5) -0.17 (1.5)	0.41 (1.1) -0.15 (1.4)
Total ER mental visits	0.11 (0.55) -0.05 (0.66)	0.21 (0.82) -	0.12 (0.49) -0.09 (0.86)	0.07 (0.44) -0.05 (0.58)	0.05 (0.33) -0.02 (0.48)
Total ER drug visits	0.01 (0.15) -0.01 (0.19)	0.03 (0.26) -	0.01 (0.12) -0.02 (0.28)	0.005 (0.07) -0.01 (0.13)	0.003 (0.06) -0.002 (0.09)
Total ER alcohol visits	0.01 (0.14) -0.01 (0.17)	0.020 (0.204) -	0.012 (0.140) -0.008 (0.21)	0.008 (0.126) -0.004 (0.15)	0.004 (0.071) -0.004 (0.13)

Admitted after alcohol ER visit **Change**	0.007 (0.1) -0.003 (0.13)	0.012 (0.141) -	0.004 (0.071) -0.007 (0.15)	0.007 (0.110) 0.003 (0.118)	0.003 (0.052) -0.005 (0.117)
Admitted after drug ER visit **Change**	0.007 (0.11) -0.004 (0.14)	0.014 (0.50) -	0.006 (0.09) -0.008 (0.19)	0.004 (0.07) -0.002 (0.11)	0.002 (0.04) -0.003 (0.08)
Admitted after mental ER visit **Change**	0.06 (0.36) -0.03 (0.45)	0.12 (0.50) -	0.068 (0.36) -0.05 (0.55)	0.038 (0.28) -0.03 (0.41)	0.025 (0.24) -0.01 (0.35)
Admitted after physical health ER **Change**	0.16 (0.57) -0.04 (0.65)	0.23 (0.66) -	0.16 (0.56) -0.07 (0.70)	0.14 (0.56) -0.02 (0.66)	0.11 (0.50) -0.03 (0.60)
Admitted after other ER visit **Change**	0.02 (0.17) -0.003 (0.24)	0.02 (0.16) -	0.02 (0.19) 0.004 (0.24)	0.02 (0.19) -0.007 (0.26)	0.01 (0.13) -0.005 (0.22)
ER visits for other problems **Change**	0.07 (0.56) -0.01 (0.71)	0.08 (0.94) -	0.09 (0.38) 0.004 (1.0)	0.05 (0.39) -0.03 (0.52)	0.05 (0.28) -0.006 (0.47)
Hospital stays for drug problems **Change**	0.003 (0.06) -0.001 (0.08)	0.005 (0.08) -	0.003 (0.07) -0.002 (0.10)	0.002 (0.04) -0.001 (0.08)	0.002 (0.05) 0.00 (0.06)
Hospital stays for mental health **Change**	0.024 (0.18) -0.011 (0.24)	0.04 (0.25) -	0.03 (0.19) -0.02 (0.29)	0.02 (0.14) -0.01 (0.23)	0.01 (0.12) -0.004 (0.18)
Hospital stays for physical health	0.058 (0.29) -0.02 (0.39)	0.09 (0.38) -	0.06 (0.28) -0.03 (0.46)	0.05 (0.27) -0.009 (0.37)	0.04 (0.21) -0.013 (0.33)
Hospital stays for other problems **Change**	0.01 (0.11) -0.002 (0.16)	0.02 (0.12) -	0.01 (0.12) -0.002 (0.17)	0.01 (0.11) -0.003 (0.16)	0.01 (0.10) -0.001 (0.15)
Routine outpatient mental visits	3.7 (7.9) 0.36 (8.3)	2.2 (4.6) -	5.5 (9.7) 3.4 (8.9)	4.0 (8.5) -1.5 (8.2)	3.3 (7.8) -0.7 (6.6)
Self-help group visits	0.90 (4.5) 0.00 (4.6)	0.66 (2.9) -	1.3 (5.5) 0.67 (5.0)	0.93 (4.8) -0.40 (4.8)	0.66 (4.5) -0.27 (4.1)
Public clinic visits	0.31 (1.5) 0.02 (2.1)	0.21 (1.1) -	0.48 (2.1) 0.28 (2.1)	0.29 (1.2) -0.19 (2.3)	0.26 (1.3) -0.03 (1.8)
Private outpatient physician visits	1.3 (2.7) 0.16 (3.0)	0.75 (1.8) -	1.8 (3.2) 1.3 (3.1)	1.4 (2.7) -0.37 (3.1)	1.2 (2.7) -0.19 (2.7)
Outpatient psychiatric visits	0.96 (2.1) 0.11 (2.6)	0.55 (1.5) -	1.4 (2.7) 0.88 (2.8)	1.0 (2.1) -0.40 (2.8)	0.87 (2.1) -0.15 (2.1)
Outpatient other mental health visits	1.6 (4.1) 0.26 (4.2)	0.82 (2.1) -	2.3 (4.8) 1.5 (4.4)	1.8 (4.6) -0.52 (4.2)	1.6 (4.2) -0.22 (3.7)
Outpatient other professional visits	0.24 (1.2) -0.02 (1.7)	0.18 (1.2) -	0.43 (1.7) 0.25 (2.1)	0.21 (0.92) -0.22 (1.8)	0.14 (0.71) -0.075 (0.89)
Other outpatient visits	0.24 (1.2) 0.04 (1.7)	0.13 (0.92) -	0.30 (1.4) 0.17 (1.4)	0.24 (1.6) -0.05 (1.9)	0.24 (1.4) -0.02 (1.9)

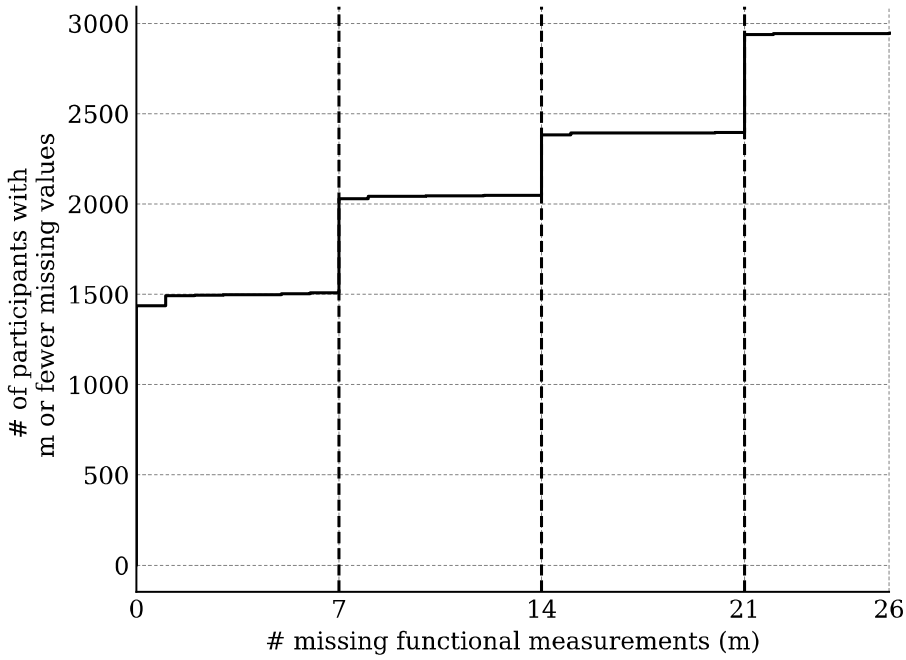

Figure 3. Number of participants with at most m missing function measurements (not including Community Mobility scales).

We analyzed the impact of our missing value imputation marginalization scheme by repeating the analysis while retaining all study participants with at most zero, seven, and fourteen missing functioning measurements (excepting Community Mobility). We found that the differences between the models were minimal (Supplemental Materials). For this reason, we report only on the most-inclusive model (accepting a tolerance of 14 missing scale scores).

Predictors of steady work

Logistic regression analyses

We standardized all predictor variables used in our models so that their effect sizes are directly comparable. In Fig 4, we display the odds ratios for the top 32 predictors for steady work, where the mean and 95% credible intervals are annotated. The top ten predictors were: baseline working status (demographic), change in Communication & Cognition score (function), having a Community Mobility score (function), change in Upper Body Function score (function), “Other” race (demographic), having a bachelor’s degree (demographic), having worked in the past 2

years at baseline (demographic), being in a treatment arm of the study (demographic), baseline Upper Body Function score (function), change in BMI (medical), and mental health-related ER visits (medical).

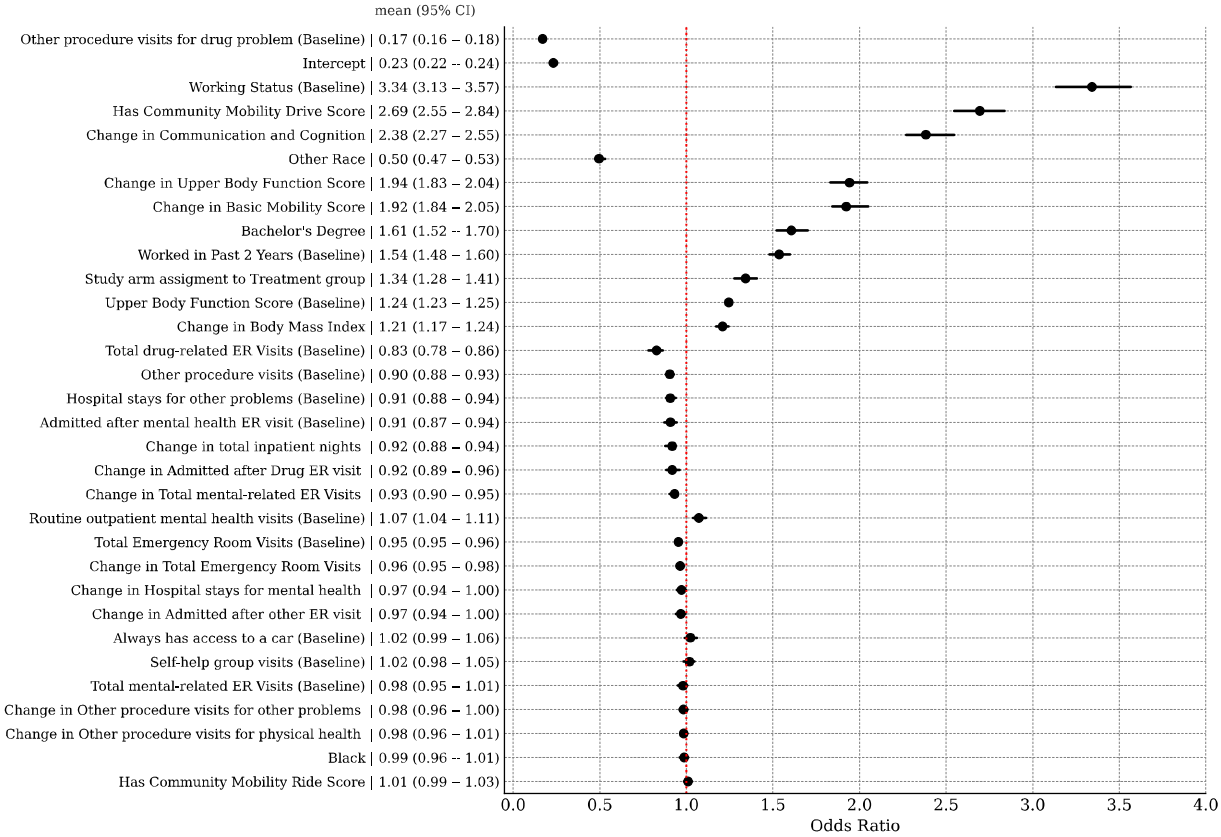

***Figure 4**. Odds ratios for the top 32 predictors of steady work. Units are standardized. Mean and 95% credible intervals presented.*

Fig 5**Error! Reference source not found.** shows the top ten predictors of steady work when restricted to demographic variables, medical + demographic variables, and WD-FAB + demographic variables. The baseline working status was the top predictor in all three models. When looking at demographic variables only, the top predictors are related to work history and education. Additionally, being in a treatment arm of the study is predictive of achieving steady work.

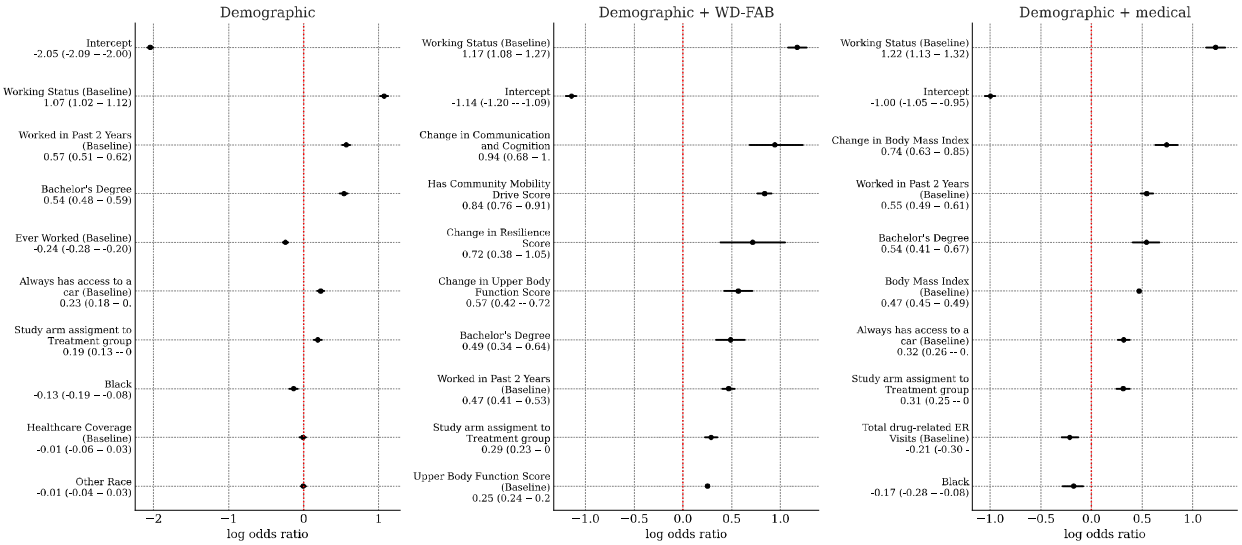

Figure 5. *Top predictors for steady work when restricted to demographic, demographic + function, or demographic + medical variables.*

When adding WD-FAB measures, improvements in Communication and Cognition, Resilience, and Upper Body Function were the most influential predictors of steady work. Additionally, having a Community Mobility Drive score (implying that a person can operate a vehicle) is also a positive predictor of steady work. Ignoring the WD-FAB and adding medical predictors, both the BMI and the change in BMI are positively associated with steady work whereas drug related ER visits at baseline is negatively associated.

Fig 6**Error! Reference source not found.** presents cross-validation-based model classification metrics, specifically the ROC and PR curves for each logistic regression model. Overall, these metrics provide an estimate of how well a given model can predict new outcomes based on new data.

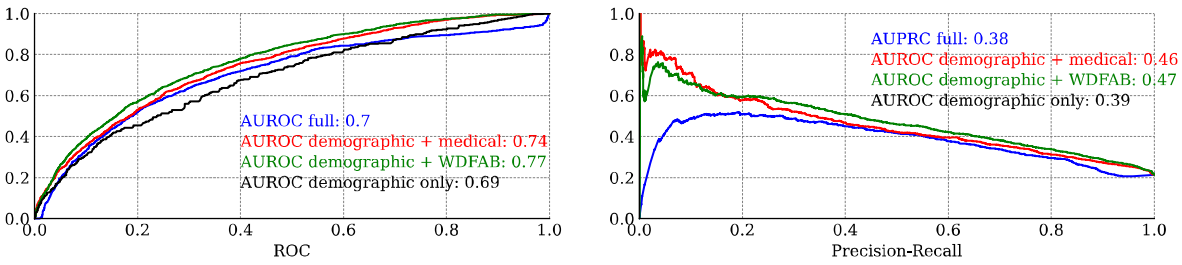

Figure 6. Leave One Out (LOO) cross validated classification metrics for logistic regression models. (a) Receiver operator characteristic and corresponding area under the curve. (b) Precision-recall curve.

Nonlinear models

In Fig 7**Error! Reference source not found.**, we present classification metrics for each of the non-linear model types that we fitted: Piecewise generalized linear regression, and Bayesian neural network. The classification performance of these two types was remarkably similar. Both types of models performed best when using the demographic + WD-FAB predictors, with that submodel performing better than the model fitted using all predictors.

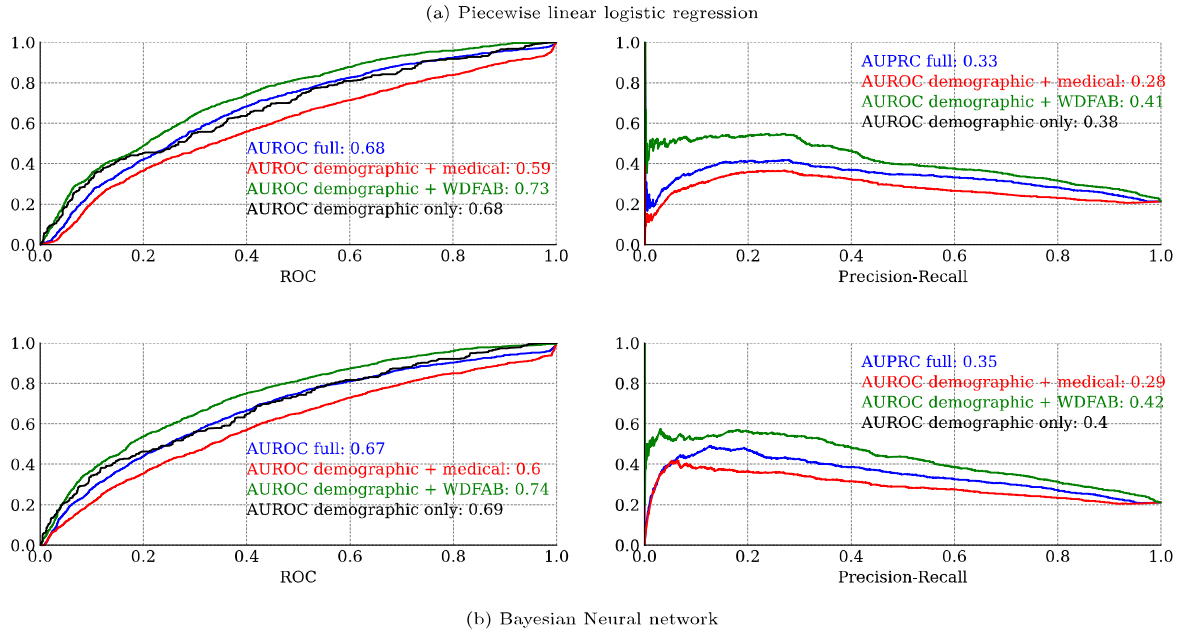

Figure 7. Leave One Out (LOO) cross validation classification metrics (ROC and PRC curves) for nonlinear models (a) Piecewise generalized linear (b) Bayesian neural network AUROC refers to the area under the receiver operator curve, PRC refers to the area under the precision recall curve, larger values are better.

DiscussionError! Reference source not found.

Functional information as a stronger predictor of steady work than medical and health care utilization data

In this manuscript, we leverage data from the SED to evaluate the relative predictive power of medical versus functional improvement in forecasting an individual's ability to maintain steady work on an annual basis, specifically among individuals with mental health conditions. While SSA has increasingly acknowledged the importance of functioning information in disability

determinations, evaluation criteria still heavily rely on impairment data and healthcare utilization metrics as primary indicators of impairment severity. This approach often overlooks the direct impact of functional limitations on work capacity. Our findings strongly support that functional improvement, as measured by the WD-FAB, provides a more accurate prediction of work status for these individuals than traditional medical impairment and healthcare utilization measures.

As evident in Table 1, the average trend for change in function is slightly negative. However, large variability in these changes exists, indicating that a significant contingent of individuals shows improvement. The observed standard deviations of WD-FAB scale changes are comparable to their empirical test-retest minimal detectable change (MDC90) thresholds (10), indicating that approximately one-third of participants experience functional improvements beyond the MDC90 threshold. However, even modest sub-threshold improvements portend increased odds of steady work. See thresholds presented in Table 1 for further details.

A detailed examination of the full predictive model (Fig 4) highlights that the top predictors of steady work are predominantly functional measures and their changes over time. Improvements in key functioning domains - such as Communication & Cognition, Upper Body Function, and Basic Mobility - emerge as strong predictors of steady work in each study year. Furthermore, when medical predictors are removed from the model (Fig 5), improvements in Resilience also become a significant factor in predicting sustained employment. Notably, while the participants were recruited into the SED based on their mental health conditions, a significant proportion also had co-occurring physical limitations. By considering a multidimensional profile of function encompassing both mental and physical domains, we obtain a more comprehensive and accurate measurement of overall ability and work potential.

Among the top ten predictors of steady work in the full model, BMI is the only medical variable, whereas functional measures dominate. However, some health care utilization variables, such as baseline mental health and substance-related ER visits, changes in total ER visits, and increases in inpatient nights, are negatively associated with steady work. Additionally, changes in overall ER visits and DAST scores appear as negative predictors of work ability when functional measures are removed from the model. These findings highlight how multidimensional functional assessments, such as the WD-FAB, can significantly outperform traditional indicators of disability - such as medical diagnoses and healthcare utilization - in predicting steady work outcomes.

Beyond the prominence of functional measures as top predictors of steady work, our analysis also demonstrates that models incorporating functional variables (i.e. the WD-FAB) alone have superior predictive accuracy compared to models relying solely on medical data. As illustrated in Fig 6, models that include demographic factors alongside WD-FAB scores achieve predictive accuracy better than the full model, as measured by both receiver operating characteristic (ROC) and precision-recall curves.

Limitations and extensions

Our analysis is based on medical variables recorded in the SED Public Use File (SED-PUF). There may be additional medical variables, especially in the domain of condition-specific impairments that were not recorded in this dataset that are predictive of work outcomes. Additionally, we incorporated the medical variables directly whereas the function predictors are a low-dimensional representation of overall physical and mental function. It is possible that low-

dimensional representations of medical utilization (34) may be more predictive of work outcome than the original variables measured in this study.

Conclusion

Functional improvement as measured by the WD-FAB is highly predictive of steady work, and more reliably predicts this outcome compared to medical impairment and healthcare utilization measures alone. These findings underscore the necessity of shifting disability assessment and work-capacity frameworks toward a more whole-person approach, moving beyond the reliance on medical diagnoses and healthcare utilization. By integrating multidimensional functional assessments such as the WD-FAB into processes and programs to help individuals with disabilities obtain and maintain employment, policymakers and practitioners can more accurately identify work potential and develop targeted interventions to support sustained employment for individuals with mental health conditions.

CRediT authorship contributions statement

Joshua C. Chang: Writing – review & editing, Writing - original draft, Formal analysis, interpretation, & visualization. **Julia Porcino**: Conceptualization, Data Curation, Investigation, Project administration, Supervision, Writing – review & editing. **Elizabeth Marfeo**: Writing – review & editing. **Larry Tang**: Writing – review & editing. **Harold Goldman**: Writing – review & editing. **Elizabeth K. Rasch**: Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgements

We would like to acknowledge the contributions of Dr. Christine McDonough, who helped define the scope and interpretation of the study but passed away before the completion of this work. This research was supported, in part, by the Intramural Research Program of the National Institutes of Health and the U.S. Social Security Administration.

References

1. Armenti K, Sweeney MH, Lingwall C, Yang L. Work: A Social Determinant of Health Worth Capturing. International Journal of Environmental Research and Public Health. 2023 Jan;20(2):1199.

2. Fiori F, Rinesi F, Spizzichino D, Di Giorgio G. Employment insecurity and mental health during the economic recession: An analysis of the young adult labour force in Italy. Social Science & Medicine. 2016 Mar 1;153:90–8.

3. Milner A, LaMontagne AD, Aitken Z, Bentley R, Kavanagh AM. Employment status and mental health among persons with and without a disability: evidence from an Australian cohort study. J Epidemiol Community Health. 2014 Nov 1;68(11):1064–71.

4. Paul KI, Moser K. Unemployment impairs mental health: Meta-analyses. Journal of Vocational Behavior. 2009 Jun 1;74(3):264–82.

5. Reuschke D, Houston D, Sissons P. Impacts of Long COVID on workers: A longitudinal study of employment exit, work hours and mental health in the UK. PLoS One. 2024;19(6):e0306122.

6. Chan XW, Shang S, Brough P, Wilkinson A, Lu C. Work, life and COVID□19: a rapid review and practical recommendations for the post□pandemic workplace. Asia Pacific Journal of Human Resources. 2022 Sep 28;10.1111/1744-7941.12355.

7. Griffiths D, Sheehan L, van Vreden C, Petrie D, Whiteford P, Sim MR, et al. Changes in work and health of Australians during the COVID-19 pandemic: a longitudinal cohort study. BMC Public Health. 2022 Mar 12;22(1):487.

8. Krahn GL, Walker DK, Correa-De-Araujo R. Persons With Disabilities as an Unrecognized Health Disparity Population. Am J Public Health. 2015 Apr;105(Suppl 2):S198–206.

9. Jette AM, Ni P, Rasch E, Marfeo E, McDonough C, Brandt D, et al. The Work Disability Functional Assessment Battery (WD-FAB). Physical Medicine and Rehabilitation Clinics. 2019 Aug 1;30(3):561–72.

10. Meterko M, Marfeo EE, McDonough CM, Jette AM, Ni P, Bogusz K, et al. Work Disability Functional Assessment Battery: Feasibility and Psychometric Properties. Archives of Physical Medicine and Rehabilitation. 2015 Jun 1;96(6):1028–35.

11. Henly M, McDonough CM, Porcino J, Peterik K, Rasch EK, Marfeo EE, et al. Linking job duties, functioning, and employment status using the Work-Disability Functional Assessment Battery (WD-FAB): An expert coding and quantitative analysis. WORK. 2023 Jan 13;74(1):75–87.

12. Clymer C, Roberts B, Strawn J. States of Change: Policies and Programs to Promote Low-Wage Workers' Steady Employment and Advancement. 2001.

13. Goldberg JF, Harrow M. Consistency of remission and outcome in bipolar and unipolar mood disorders: a 10-year prospective follow-up. Journal of Affective Disorders. 2004 Aug 1;81(2):123–31.

14. Salyers MP, Becker DR, Drake RE, Torrey WC, Wyzik PF. A ten-year follow-up of a supported employment program. Psychiatr Serv. 2004 Mar;55(3):302–8.

15. Becker D, Whitley R, Bailey EL, Drake RE. Long-term employment trajectories among participants with severe mental illness in supported employment. Psychiatr Serv. 2007 Jul;58(7):922–8.

16. Bush PW, Drake RE, Xie H, McHugo GJ, Haslett WR. The long-term impact of employment on mental health service use and costs for persons with severe mental illness. Psychiatr Serv. 2009 Aug;60(8):1024–31.

17. Kukla M, Bond GR, Xie H. A prospective investigation of work and nonvocational outcomes in adults with severe mental illness. J Nerv Ment Dis. 2012 Mar;200(3):214–22.

18. McHugo GJ, Drake RE, Xie H, Bond GR. A 10-year study of steady employment and non-vocational outcomes among people with serious mental illness and co-occurring substance use disorders. Schizophr Res. 2012 Jul;138(2–3):233–9.

19. Davis LL, Kyriakides TC, Suris AM, Ottomanelli LA, Mueller L, Parker PE, et al. Effect of Evidence-Based Supported Employment vs Transitional Work on Achieving Steady Work Among Veterans With Posttraumatic Stress Disorder: A Randomized Clinical Trial. JAMA Psychiatry. 2018 Apr 1;75(4):316–24.

20. SSA. Contents of sed_puf_final [Internet]. 2023 [cited 2025 Mar 6]. Available from: https://www.ssa.gov/disabilityresearch/documents/sed/Contents%20of%20sed_puf_final_upd_20230510.pdf

21. Gelman A, Hill J, Vehtari A. Regression and Other Stories. Cambridge University Press; 2021. 551 p.

22. Skinner HA. The drug abuse screening test. Addict Behav. 1982;7(4):363–71.

23. Babor TF, Robaina K. The Alcohol Use Disorders Identification Test (AUDIT): A review of graded severity algorithms and national adaptations. International Journal of Alcohol and Drug Research. 2016 Jul 19;5(2):17–24.

24. Bohn MJ, Babor TF, Kranzler HR. The Alcohol Use Disorders Identification Test (AUDIT): validation of a screening instrument for use in medical settings. J Stud Alcohol. 1995 Jul;56(4):423–32.

25. Reinert DF, Allen JP. The Alcohol Use Disorders Identification Test (AUDIT): a review of recent research. Alcohol Clin Exp Res. 2002 Feb;26(2):272–9.

26. Boothroyd RA, Chen HJ. The Psychometric Properties of the Colorado Symptom Index. Adm Policy Ment Health. 2008 Sep 1;35(5):370–8.

27. Piironen J, Paasiniemi M, Vehtari A. Projective Inference in High-dimensional Problems: Prediction and Feature Selection. arXiv:181002406 [cs, stat] [Internet]. 2018 Oct 4 [cited 2020 May 13]; Available from: http://arxiv.org/abs/1810.02406

28. Xia H, Chang JC, Nowak S, Mahajan S, Mahajan R, Chang TL, et al. Interpretable (not just posthoc-explainable) heterogeneous survivors bias-corrected treatment effects for assignment of postdischarge interventions to prevent readmissions. In: Proceedings of the 8th Machine Learning for Healthcare Conference [Internet]. PMLR; 2023 [cited 2024 Apr 2]. p. 884–905. Available from: https://proceedings.mlr.press/v219/xia23a.html

29. Chang JC, Chang TL, Chow CC, Mahajan R, Mahajan S, Maisog J, et al. Interpretable (not just posthoc-explainable) medical claims modeling for discharge placement to prevent avoidable all-cause readmissions or death [Internet]. arXiv; 2023 [cited 2024 Apr 2]. Available from: http://arxiv.org/abs/2208.12814

30. Vehtari A, Simpson D, Gelman A, Yao Y, Gabry J. Pareto Smoothed Importance Sampling. Journal of Machine Learning Research. 2024;25(72):1–58.

31. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017 Sep 1;27(5):1413–32.

32. Chang JC, Li X, Xu S, Yao HR, Porcino J, Chow C. Gradient-flow adaptive importance sampling for Bayesian leave one out cross-validation for sigmoidal classification models [Internet]. arXiv; 2024 [cited 2024 Mar 28]. Available from: http://arxiv.org/abs/2402.08151

33. Dillon JV, Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, et al. TensorFlow Distributions [Internet]. arXiv; 2017 [cited 2024 Apr 2]. Available from: http://arxiv.org/abs/1711.10604

34. Chang JC, Fletcher P, Han J, Chang TL, Vattikuti S, Desmet B, et al. Sparse encoding for more-interpretable feature-selecting representations in probabilistic matrix factorization. arXiv:201204171 [cs, q-bio, stat] [Internet]. 2020 Dec 7 [cited 2020 Dec 9]; Available from: http://arxiv.org/abs/2012.04171

Supporting information

S1 Text. Supplemental methods and results.