A Learning Algorithm for Predicting Mental Health Symptoms and Substance Use

Anthony T. Fojo, MD, MHS [Assistant Professor],
School of Medicine, Johns Hopkins University, Baltimore, MD

Catherine R. Lesko, PhD [Assistant Professor],
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Kelly S. Benke, PhD [Assistant Scientist],
Johns Hopkins Bloomberg School of Public Health, Department of Mental Health, Baltimore, MD

Geetanjali Chander, MD, MPH [Associate Professor],
School of Medicine, Johns Hopkins University, Baltimore, MD

Bryan Lau, PhD [Associate Professor],
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Richard D. Moore, MD [Professor],
School of Medicine, Johns Hopkins University, Baltimore, MD

Peter P. Zandi [Professor],
Johns Hopkins Bloomberg School of Public Health, Department of Mental Health, Baltimore, MD

Scott L. Zeger [Professor]
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD

Keywords

Mental Health; Patient Reported Outcomes; Substance Use; Learning Health System

INTRODUCTION:

A learning health system refers to infrastructure that iteratively uses data to generate knowledge that can be applied to clinical care (Friedman et al., 2015). With increasing quantities of data available in health care, learning health systems offer the promise of substantial improvements in health (2011; Friedman et al., 2015). In 2012, the Institute of Medicine called for the development of a continuously learning healthcare system (Institute of Medicine, 2013).

Psychiatric diseases and substance misuse are highly prevalent both in the US and globally, and impose substantial morbidity and mortality (G. B. D. Alcohol Drug Use, 2018; Kessler et al., 2012). These disorders frequently co-occur; concurrent disorders can increase the

Corresponding Author: Anthony Todd Fojo, <u>Phone:</u> 650-906-8938, tfojo13@gmail.com, <u>Address:</u> 2024 East Monument Street, Suite 2-600, Baltimore, MD 21287.

clinical severity and be more treatment-resistant (Arostegui et al., 2012; Pettinati et al., 2013). The ability to predict future mental health and substance use burden could allow for targeted interventions in patients at highest risk for poor outcomes.

Patient-reported measures of mental health and substance use are increasingly collected by health systems (Krägeloh et al., 2015). When mental health symptoms and substance use are measured serially over time, we can conceive of them as following a trajectory in a multidimensional space, where each dimension corresponds to one psychiatric disease or substance use disorder; this approach can be leveraged to predict future symptoms and substance use (Fojo et al., 2017).

We sought to develop a statistically rigorous approach to use serial, patient-reported measurements of mental health and substance use to predict future mental health comorbidity and substance use. With an eye to learning health systems, we specifically designed an approach that fits a model in the population in which it will be used to make predictions. We validated our approach using data from two cohorts of patients in outpatient care.

MATERIALS AND METHODS:

Study Populations:

We evaluated the performance of our model in two cohorts in clinical care, with the aim of demonstrating generalizability across two very different populations:

1. **The National Network of Depression Centers (NNDC):** The National Network of Depression Centers is a consortium of centers in the United States that care for patients with mood disorders; we included patient data contributed by sixteen sites from January 2011 through December 2014 (Greden, 2011). At clinic visits, patients completed self-reported assessments which include (1) the Patient Health Questionnaire 9-item scale (PHQ-9), a measure of depression (Kroenke et al., 2001), (2) the Generalized Anxiety Disorder 7-item scale (GAD-7) (Spitzer et al., 2006), and (3) the Altman Self-Rating Mania Scale (ASRM) (Altman et al., 1997). The consortium also collected information on patients' demographics, employment, and education.

2. **The Johns Hopkins HIV Clinical Cohort (JHHCC):** The Johns Hopkins HIV Clinical Cohort (JHHCC) comprises all HIV-infected person 18 years or older who enroll in HIV care at Johns Hopkins outpatient HIV clinic and consent to share their electronic health record and prescription data (Moore, 1998). A subset of patients completed a computer-assisted self-interview approximately every six months, which includes: (1) the 8-item Patient Health (PHQ-8), a subset of the PHQ-9, (2) the GAD-7, (3) the Alcohol Use Disorders Identification Test consumption questions (AUDIT-C) (Reinert and Allen, 2007), and (4) the National Institute on Drug Abuse modified Alcohol, Smoking And Substance Involvement Screening Test (NIDA-ASSIST) (National Institute on Drug Abuse (NIDA), 2010). We included data from June 2013 through December 2017 on all patients who completed at least one self-interview. The PHQ-8, AUDIT-C, and

NIDA-ASSIST are collected at the first clinic visit greater than 6 months since the prior collection; the GAD-7 is recorded at the first visit one year or more after the previous time it was recorded.

Both cohorts collect patient reported outcomes independently of the clinical care visit; the collection of outcomes does not depend on patient history or what is discussed in visits.

Outcomes:

For the NNDC, we sought to predict scores on the PHQ-9 (which range from 0 to 27), GAD-7 (range 0–21), and ASRM (range 0–20). For the JHHCC, we sought to predict scores on the PHQ-8 (range 0–24), GAD-7, and AUDIT-C (range 0–12), as well as binary indicators of reported heroin and cocaine use within the past 3 months. Missing single questions from multi-question scales (PHQ, GAD-7, and ASRM) were handled according to standard scoring protocols: if one item was missing, the scale was scored as though that item were zero. If two or more items were missing, the entire scale was treated as missing. The second and third items from the AUDIT-C could be missing only if the first item indicated no alcohol use.

Because outcomes in both cohorts are systematically collected independent of the clinical visit, most missing data are the result of either how often outcomes are collected (GAD-7 in the JHHCC is collected yearly instead of every 6 months) or constraints regarding starting the visit on time. Patients can refuse to report outcomes, but this is rare. Consequently, we made the assumption that observations were missing at random.

The Predictive Model:

Our objective was to design a flexible approach to operate with any mix of continuous symptom scales and binary indicators of substance use. We developed a Bayesian, hierarchical model in which we represented each continuous outcome at a particular time with a Tobit model, assuming that the symptom scale we observed was actually a truncated latent normal variable (Arostegui et al., 2012) and binary outcomes with a Probit model, assuming a latent normal variable which was greater than zero if substance use occurred and less than zero if it did not (the mathematical details are presented in the Supplement). Observations of symptoms and substance use could be made at any arbitrary times. We used a continuous autoregressive process to specify that observations were more closely related if they were made closer in time. Observations were also allowed to be missing and assumed to be missing at random (conditional on covariates).

We represented the set of all latent variables as following a multivariate normal distribution with four components: (1) fixed effects (common to all subjects), (2) random effects (specific to each individual), (3) the continuous autoregressive process that defined how outcomes covary through time (after removing fixed and random effects), and (4) Gaussian measurement error (Diggle, 2013). We allowed the random effects and measurement errors to covary across different outcomes. Fundamentally, our model encapsulates the idea that patients move through a "symptom/substance use" space through time, and that changes in one symptom or substance use pattern within individuals correlate to changes in other symptoms or substance use patterns.

Model Fitting:

We fit the model using Hamiltonian Markov-Chain Monte Carlo simulation, a Bayesian method in which values for model parameters are simulated thousands of times to approximate their probability distributions. We used the Rstan package (Carpenter et al., 2017) in R version 3.6.2 to run two chains with 2,000 warm-up iterations and 2,000 sampling iterations for each population; we inspected trace plots to assess convergence.

Covariates:

Our approach allows for a flexible specification of covariates that can be customized to specific settings. In both the NNDC and JHHCC validation sets, we used sex, and race/ethnicity as reported at the first visit, as well as time-varying age as a restricted quadratic spline with knots at the 20th, 40th, 60th, and 80th percentiles.(Howe et al., 2011) In the NNDC validation, we additionally used indicators for education taken at the first visit and time-varying current employment. In the JHHCC validation, we also included HIV acquisition risk factor as reported at the first visit (indicators for men who have sex with men and history of injection drug use), as well as time-varying binary indicators of a recent (within the past two years) diagnosis in the medical chart of depression, anxiety, alcohol use disorder, and opioid and cocaine use, and an indicator for concurrent prescription of any mental health medication (any selective serotonin reuptake inhibitor, serotonin–norepinephrine reuptake inhibitor, tricyclic antidepressant, serotonin modulator and stimulator, monoamine oxidase inhibitor, bupropion).

Evaluation of Predictive Performance:

We used an internal, temporal validation approach with five-fold cross validation to evaluate our model's performance. We divided each of the two validation sets into five subsets of equal size. We iteratively fit the model on 80% of each dataset and used the fitted model to make predictions on the remaining 20%. For each subset, all observations for the 80% in the training set were used to fit the model, even if participants had only one visit.

In making predictions for each validation subset, we considered all participants in the 20% not used to fit the model who had more than one clinic visit with observations. We passed observations from all but their last clinic visit to the fitted model to predict symptom scores and substance use at the last visit. This approach simulates the way the algorithm might be deployed in a clinic setting – using all of a patient's history up to a point to predict where they may be at their next clinic visit. Consequently, if a particular outcome (for example, the GAD-7) was not recorded for a given participant at their last visit, that outcome was not validated for that subject, even if it was recorded at prior visits. In this situation, the GAD-7 at prior visits for a participant would be used to inform predictions for that participant's other outcomes at the last visit. We compared our predicted outcomes to the actual outcomes observed at the last visit. While participants with only one clinic visit were used in fitting the models, to help estimate fixed and random effects, they were not used in the validation process. Our primary outcome for predictive performance was the area under the receiver-operating characteristic (ROC) curves of binary outcomes and of continuous outcomes dichotomized at their common thresholds (10 on the PHQ, 7 on the GAD, 6 on the ASRM, and 3 on the AUDIT-C for women or 4 on the AUDIT-C for men) (Altman et al.,

1997; Kroenke et al., 2001; Reinert and Allen, 2007; Spitzer et al., 2006). We compared our algorithm's predictions to the simple predictors of (a) the symptoms or substance use recorded at the last visit, and (b) the average of all prior symptoms or substance use indicators. We additionally calculated root mean squared error (RMSE – how far the estimates are from the truth), mean bias, estimated variance, and prediction interval coverage (how often the 95% prediction intervals included the truth) for continuous outcomes.

We conducted secondary analyses to evaluate predictive performance depending on the number of prior observations, as well as to assess whether predictive performance declined with increasing time from the last observation to the prediction. We used DeLong's algorithm to compare different ROC curves (DeLong et al., 1988). All calculations and comparisons of ROC curves were made with the pROC package (Robin et al., 2011) in R version 3.6.2.

RESULTS:

Table 1 details the demographic characteristics of the 2,444 participants in the two validation sets (1,234 in the NNDC and 1,210 in the JHHCC). Each of these 2,444 participants was used in fitting models in exactly 4 of the 5 subsets for the validation procedure. The NNDC is majority female (66.9%) and predominantly white (83.2%), whereas the JHHCC is majority male (63.1%) and predominantly Black (84.0%).

708 participants in the NNDC and 1,205 participants in the JHHCC had more than one visit with symptoms recorded and were included in the validation set. In the NNDC, 648 of the 708 participants (91.5%) had a PHQ recorded at their last visit, 636 (89.8%) had a GAD-7 recorded, and 602 (85%) had an ASRM recorded. In the JHHCC, 1,166 of the 1,205 participants (96.8%) had a PHQ recorded at their last visit, 1,197 (99.3%) had an AUDIT-C recorded, 1,193 (99%) had heroin use vs. non-use recorded, and 1,196 (99.3%) had cocaine use vs non-use recorded. A GAD-7, which is collected yearly (as opposed to every 6 months for other outcomes in the JHHCC) was recorded for 592 participants (49.1%) at their last visit.

Table 2 gives the distribution of the patient reported outcomes across the six mental health and substance use domains, for all 19,088 clinic visits as well as for the subset of 1,213 visits used in the validation procedure. Across all visits, psychiatric symptoms were greater in the NNDC, with a mean PHQ-9 score of 10.9 and GAD-7 score of 8.5 compared to 3.4 and 2.2 for JHHCC. Substance use was prevalent in the JHHCC with participants endorsing cocaine use at 11.4% of visits, heroin use at 6.4% of visits, and heavy alcohol use at 16.0% of visits. Visits in the NNDC were more closely spaced, with a median of 33 days in between visits (interquartile range 14 to 78), compared to 91 days (interquartile range 46 to 133) between visits with self-reported symptoms in the JHHCC. The distribution of outcomes did not differ substantially from the full set of all visits to the subset of visits used in validation.

We iteratively blinded ourselves to all observations of any mental health symptom scale or self-reported substance use at each participant's last clinic visit and predicted those values

conditional on observations at all prior visit; participants could contribute observations (and predictions) for multiple symptom scales and substance use indicators at their last clinic visit. To help illustrate this predication process, figure 1 depicts the prior trajectory and future prediction for one participant from the NNDC and one participant from the JHHCC. Figure 2 and table 3 detail the predictive performance of our algorithm on the 5,383 predictions at those 1,913 final visits when scales were dichotomized at common cut points. Figure S3 in the Supplement shows calibration plots for the continuous outcomes.

Predicted probabilities of a PHQ score 10 yielded an AUC of 0.85 in the NNDC (across 648 predictions) and 0.84 in the JHHCC (1,166 predictions). Continuous predictions of the PHQ score yielded a RMSE of 5.3, mean bias of 1.1, and estimated variance of 27.20 in the NNDC and RMSE 3.9, mean bias −1.1, and estimated variance 10.1 in the JHHCC (the score ranges from 0–27 for the 9-item version used in the NNDC and 0–24 for the 8-item version used in the JHHCC). Prediction intervals included the true value 96.3% (NNDC) and 96.0% (JHHCC) of the time, and were on average 19.1 (NNDC) and 11.5 (JHHCC) points wide.

Predicted probabilities of a GAD-7 score 7 yielded an AUC of 0.85 in the NNDC (across 636 predictions) and 0.84 in the JHHCC (592 predictions). Continuous predictions of GAD-7 score (which ranges from 0–21) yielded RMSE of 4.5, mean bias of 0.8, and estimated variance of 23.2 in the NNDC, and RMSE 4.0, mean bias −1.5, and estimated variance 3.9 in the JHHCC, with prediction interval coverage of 96.1% (NNDC) and 97.3% (JHHCC) and average width of 15.8 (NNDC) and 11.6 (JHHCC) points.

In the NNDC, the AUC was 0.77 for a predicted probability that the ASRM was 6 (602 predictions). The RMSE for continuous predictions was 2.8 (the scale ranges from 0–20), with mean bias −0.7 and estimated variance 3.1. Prediction interval coverage was 97.3% with an average width of 8.9 points. In the JHHCC, the predicted probability of an AUDIT-C 3 for women or 4 for men was 0.90 (1,197 predictions), with an RMSE for continuous predictions of the AUDIT-C (which ranges from 0–12) of 1.4, mean bias of −0.4, and estimated variance of 2.5, as well as a prediction interval coverage of 97.4% and width of 4.6 points. Lastly, in the JHHCC, the AUC was 0.92 for predicted heroin use (1,193 predictions) and 0.90 for predicted cocaine use (1,196 predictions).

Table 3 also compares our algorithm's predictions to the simple predictors of (a) symptoms or substance use at the preceding visit and (b) the average of symptoms or substance use indicators across all prior visits. These simple predictors performed well for the AUDIT-C (AUCs of 0.86 and 0.90 respectively) and cocaine use (AUCs of 0.81 and 0.87), but were lower by 0.1 or more than our algorithm's predictions for other psychiatric symptoms and substance use indicators.

Predictive performance improved marginally with more prior observations to inform predictions and declined marginally with a longer interval over which to predict (Figure 3). The ROC for predictions made based off of zero or one prior observations were 0.83 for the NNDC and 0.86 for the JHHCC compared to 0.87 and 0.89 for predictions based off of 2 or more prior observations (p value 0.048 for NNDC and 0.044 for JHHCC). In the NNDC the

ROC for predictions made 60 days in the future was 0.89 compared to 0.82 for predictions 60–180 days (p=0.003) and 0.81 for predictions >180 days in the future (p=0.005). In the JHHCC there were no observations 60 days apart, but there was no significant difference in the ROC for predictions 180 days into the future (0.89) compared to >180 days into the future (0.87, p=0.18).

DISCUSSION:

We developed a statistically rigorous approach to predict patients' future mental health comorbidity and substance use from prior self-reported mental health symptom scores and substance use indicators. With an eye towards learning health systems, our approach allows a model to be fitted in the population in which it will be used. In two populations of patients in outpatient care, our predictions achieved strong discrimination between participants with moderate-to-severe mental health symptoms vs mild-to-no symptoms, and between participants with subsequent substance use vs none. Our approach is robust to missing data and irregular time intervals between observations.

Psychiatric symptoms and reported of substance use are inherently imprecise and variable (McMahon, 2014). The strength of our approach is to quantify this imprecision, and reduce variability by leveraging repeated measurements of an individual's trajectory and the covariance across different psychiatric and substance use domains. Our algorithm has the virtue of being flexible with respect to variable time spans of measurement and robust to missing data, both of which are the rule in clinical data. Furthermore, this approach is not limited to the specific domains we evaluated here (depression, anxiety, mania, heroin, cocaine, and alcohol use). It can be adapted to any set of numeric symptom scales or binary indicators that are serially measured over time. Predictive performance was consistent across two very different clinical populations, which suggests that this approach has broad generalizability.

Our algorithm had better predictive performance that the simple predictors of either symptoms/substance use at the preceding visit or the average of all prior symptoms/ substance use indicators for depression (PHQ), anxiety (GAD-7), mania (ASRM), and heroin use, but it provided minimal improvement for predicting alcohol misuse (AUDIT-C) or cocaine use in the Johns Hopkins HIV Clinical Cohort (JHHCC), for which the simple predictors performed highly. This likely reflects the fact that alcohol misuse and cocaine use are persistent and stable across the participants in the JHHCC, and are not strongly impacted by comorbid conditions. Conversely, our algorithm produced the largest gain for anxiety symptoms in the JHHCC, which may be due in part to the fact that the GAD-7 is only collected yearly, and other symptoms observed at intervening 6-month intervals can improve the predictions of anxiety. With the exception of manic symptoms (ASRM), predictions of binary high vs. low symptoms or substance use vs. non-use could achieve a sensitivity in the mid-70s to high 80s with cut points selected to yield a specificity of 75%; sensitivity fell off as specificity was increased to 90%.

This approach could form the basis for tools that are customized and fitted within specific health systems, and then deployed at the point of care. Our algorithm's discrimination of

high vs low symptoms was strong, but prediction intervals for specific symptom levels were wide. The algorithm thus does a good job of distinguishing "high burden" from "low burden" – particularly in the JHHCC where many individuals have no symptoms and some have a substantial symptom burden – but still infers a wide range around the high or low estimate. Distinguishing between high and low burden provides actionable intelligence for clinicians. A request by a clinician for predictions during a visit or an automated alert could help clinicians to identify patients at high risk for a substantial psychiatric symptoms or ongoing substance use and allow for tailored interventions to be targeted to those patients in real time.

Relatively few studies have evaluated prediction algorithms for mental health symptoms or substance use (Becker et al., 2018). Some work has focused on predicting health care utilization or readmissions relating to psychiatric disease (Donisi et al., 2016; van Orden et al., 2016). A number of studies have sought to predict binary outcomes, such as risk of suicide (Kessler et al., 2017), remission of depression (Sun et al., 2014), relapse of alcohol abuse (Farren and McElroy, 2010; Farren et al., 2013). However, predicting "hard" clinical outcomes overlooks the substantial comorbidity that severe symptoms can incur. The algorithm we have outlined in this manuscript is unique in focusing primarily on symptoms, considering multiple psychiatric and substance use domains simultaneously, and leveraging repeated measurements across time to make predictions.

Our approach has a number of limitations. First, we base predictions on measurements that are months apart. Changes in mental health and substance use often happen on a much faster time scale, and intermittent clinic measurements may miss changes that will have a substantial impact on a patient's future mental health. Second, we represent all participants as having the same covariance between different mental health and substance use metrics; in other words, we assume that changes in, for example, alcohol use are equally correlated with changes in depressive symptoms across all subjects. Future work could allow this covariance structure to vary across subgroups or individuals, and potentially incorporate genetic markers of psychiatric disease. Third, we rely on patient reported outcomes, which are subject to misreporting; our model accounts for this by postulating a measurement error. However, patient reported outcomes can be collected in a manner that imposes little additional burden on a health system and are thus potentially more scalable. Future work could explicitly incorporate imperfect self-report of substance use (Lesko et al., 2018). Fourth, we assume data are missing at random. In the two cohorts we use in this work, where symptoms are systematically collected according to pre-specified schedules, this is likely a reasonable assumption. Generalizing to other settings where patient reported outcomes are not systematically collected would require an extension of the algorithm explicitly linking symptom levels to the probability of observing those symptoms.

In conclusion, we have presented a statistically rigorous algorithm to predict patients' future mental health comorbidity and substance use based off of serial patient-reported mental health symptoms and substance use. Our predictions performed well in two validation sets, discriminating between moderate-to-severe vs. mild-to-no mental health symptoms and substance use vs. no use, and our approach can be adapted to a variety of symptoms or binary indicators and can be customized and fitted for specific health systems. A rigorous,

mathematically grounded approach to prediction of mental health and substance use can realize the potential of a learning health system to transform ever-increasing quantities of data into tangible guidance for patient care.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

REFERENCES

2011 in: Grossmann C, Powers B, McGinnis JM (Eds.), Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington (DC).

Altman EG, Hedeker D, Peterson JL, Davis JM, 1997. The Altman Self-Rating Mania Scale. Biological Psychiatry 42(10), 948–955. [PubMed: 9359982]

Arostegui I, Nunez-Anton V, Quintana JM, 2012. Statistical approaches to analyse patient-reported outcomes as response variables: an application to health-related quality of life. Stat Methods Med Res 21(2), 189–214. [PubMed: 20858689]

Becker D, van Breda W, Funk B, Hoogendoorn M, Ruwaard J, Riper H, 2018. Predictive modeling in e-mental health: A common language framework. Internet Interventions 12, 57–67. [PubMed: 30135769]

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A, 2017. Stan: A Probabilistic Programming Language. Journal of Statistical Software 76(1).

DeLong ER, DeLong DM, Clarke-Pearson DL, 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3), 837–845. [PubMed: 3203132]

Diggle P, 2013. Analysis of longitudinal data, Second Paperback Edition. ed. Oxford University Press, Oxford;.

Donisi V, Tedeschi F, Wahlbeck K, Haaramo P, Amaddeo F, 2016. Pre-discharge factors predicting readmissions of psychiatric patients: a systematic review of the literature. BMC Psychiatry 16(1).

Farren CK, McElroy S, 2010. Predictive Factors for Relapse after an Integrated Inpatient Treatment Programme for Unipolar Depressed and Bipolar Alcoholics. Alcohol and Alcoholism 45(6), 527–533. [PubMed: 20855410]

Farren CK, Snee L, Daly P, McElroy S, 2013. Prognostic Factors of 2-year Outcomes of Patients with Comorbid Bipolar Disorder or Depression with Alcohol Dependence: Importance of Early Abstinence. Alcohol and Alcoholism 48(1), 93–98. [PubMed: 23059424]

Fojo AT, Musliner KL, Zandi PP, Zeger SL, 2017. A precision medicine approach for psychiatric disease based on repeated symptom scores. J Psychiatr Res 95, 147–155. [PubMed: 28863391]

Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, Gunter C, Musen M, Platt R, Stead W, Sullivan K, Van Houweling D, 2015. Toward a science of learning systems: a research agenda

for the high-functioning Learning Health System. J Am Med Inform Assoc 22(1), 43–50. [PubMed: 25342177]

Alcohol Drug Use GBD, 2018. The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Psychiatry 5(12), 987–1012. [PubMed: 30392731]

Greden JF, 2011. The National Network of Depression Centers: progress through partnership. Depress Anxiety 28(8), 615–621. [PubMed: 21796738]

Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ Jr., 2011. Splines for trend analysis and continuous confounder control. Epidemiology 22(6), 874–875. [PubMed: 21968779]

Institute of Medicine, 2013. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. The National Academies Press, Washington, DC.

Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova MV, Rosellini AJ, Sampson NA, Schneider AL, Bradley PA, Katz IR, Thompson C, Bossarte RM, 2017. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psychiatr Res 26(3).

Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen HU, 2012. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. Int J Methods Psychiatr Res 21(3), 169–184. [PubMed: 22865617]

Krägeloh CU, Czuba KJ, Billington DR, Kersten P, Siegert RJ, 2015. Using Feedback From Patient-Reported Outcome Measures in Mental Health Services: A Scoping Study and Typology. Psychiatric Services 66(3), 224–241. [PubMed: 25727110]

Kroenke K, Spitzer RL, Williams JB, 2001. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 16(9), 606–613. [PubMed: 11556941]

Lesko CR, Keil AP, Moore RD, Chander G, Fojo AT, Lau B, 2018. Measurement of Current Substance Use in a Cohort of HIV-Infected Persons in Continuity HIV Care, 2007–2015. Am J Epidemiol 187(9), 1970–1979. [PubMed: 29701832]

McMahon FJ, 2014. Prediction of treatment outcomes in psychiatry--where do we stand? Dialogues Clin Neurosci 16(4), 455–464. [PubMed: 25733951]

Moore RD, 1998. Understanding the clinical and economic outcomes of HIV therapy: the Johns Hopkins HIV clinical practice cohort. J Acquir Immune Defic Syndr Hum Retrovirol 17 Suppl 1, S38–41. [PubMed: 9586651]

National Institute on Drug Abuse (NIDA), 2010. Clinician's Screening Tool for Drug Use in General Medical Settings. https://www.drugabuse.gov/nmassist/.

Pettinati HM, O'Brien CP, Dundon WD, 2013. Current status of co-occurring mood and substance use disorders: a new therapeutic target. Am J Psychiatry 170(1), 23–30. [PubMed: 23223834]

Reinert DF, Allen JP, 2007. The alcohol use disorders identification test: an update of research findings. Alcohol Clin Exp Res 31(2), 185–199. [PubMed: 17250609]

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M, 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12(1).

Spitzer RL, Kroenke K, Williams JBW, Löwe B, 2006. A Brief Measure for Assessing Generalized Anxiety Disorder. Archives of Internal Medicine 166(10), 1092. [PubMed: 16717171]

Sun HS, Demic S, Cheng S, 2014. Modeling the Dynamics of Disease States in Depression. PLoS ONE 9(10), e110358. [PubMed: 25330102]

van Orden M, Leone S, Haffmans J, Spinhoven P, Hoencamp E, 2016. Prediction of Mental Health Services Use One Year After Regular Referral to Specialized Care Versus Referral to Stepped Collaborative Care. Community Mental Health Journal 53(3), 316–323. [PubMed: 27430954]
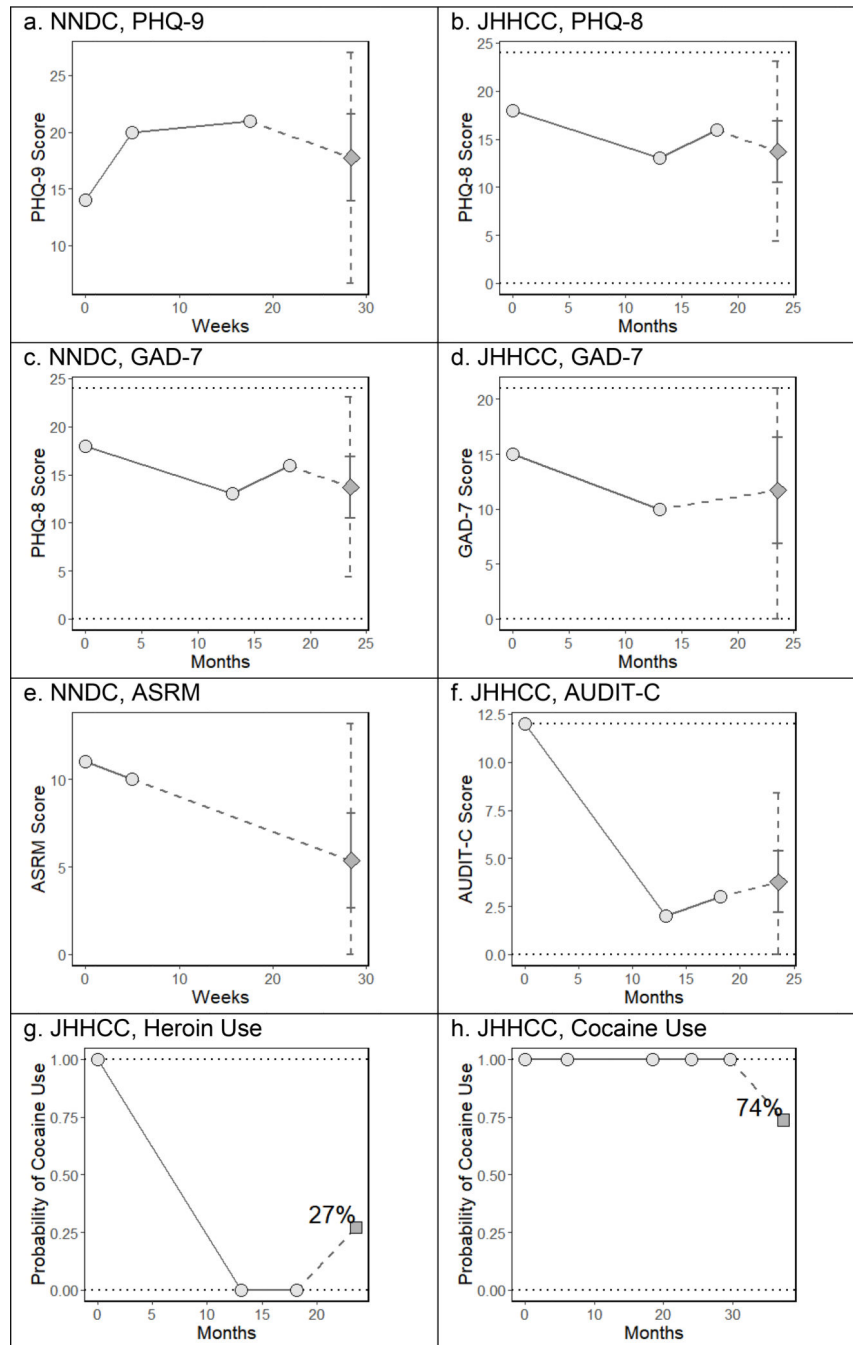
Figure 1: Sample Individual Predictions

Circles indicate prior observations. Diamonds indicate predicted future symptom scores and squares indicate predicted probability of future substance use. The solid error bars around predicted scores denote the 50% prediction interval, and the dashed error bars denote the 95% prediction interval.
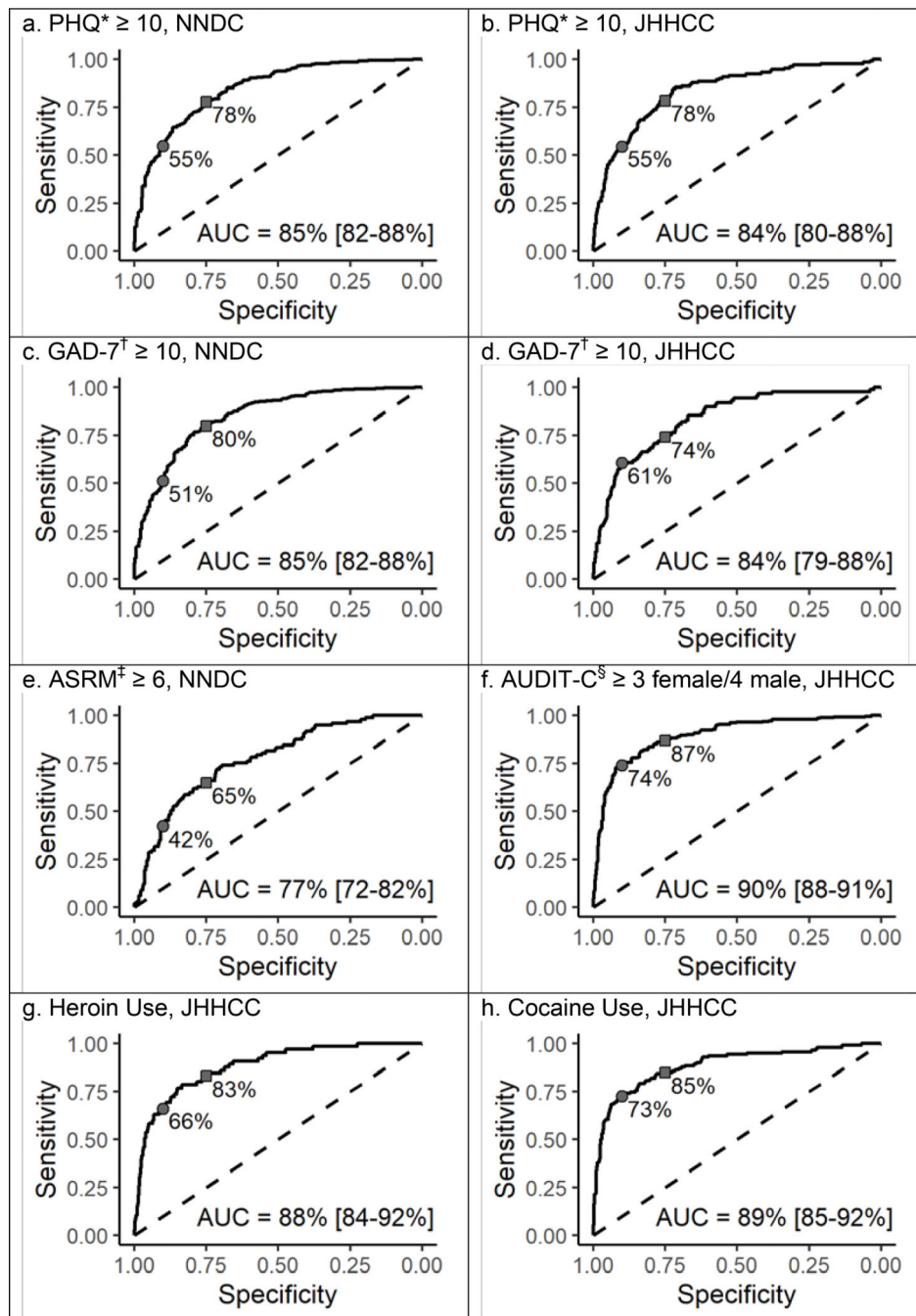
Figure 2: ROC Curves for Dichotomized and Binary Outcomes

*PHQ = Patient Health Questionnaire (9-item scale for NNDC ranges from 0–27, 8-item scale for JHHCC ranges from 0–24). †GAD-7 = Generalized Anxiety Disorder scale (ranges from 0–21). ‡ASRM = Altman Self-Rating Mania Scale (ranges from 0–20). §AUDIT-C = Alcohol Use Disorders Identification Test consumption questions (ranges 0–12). The sensitivity corresponding to a specificity of 75% (squares) and 90% (circles) are noted on each curve.

Figure 3:
Predictive Performance vs. Number of Prior Observations and Time Since Most Recent Prior Observation
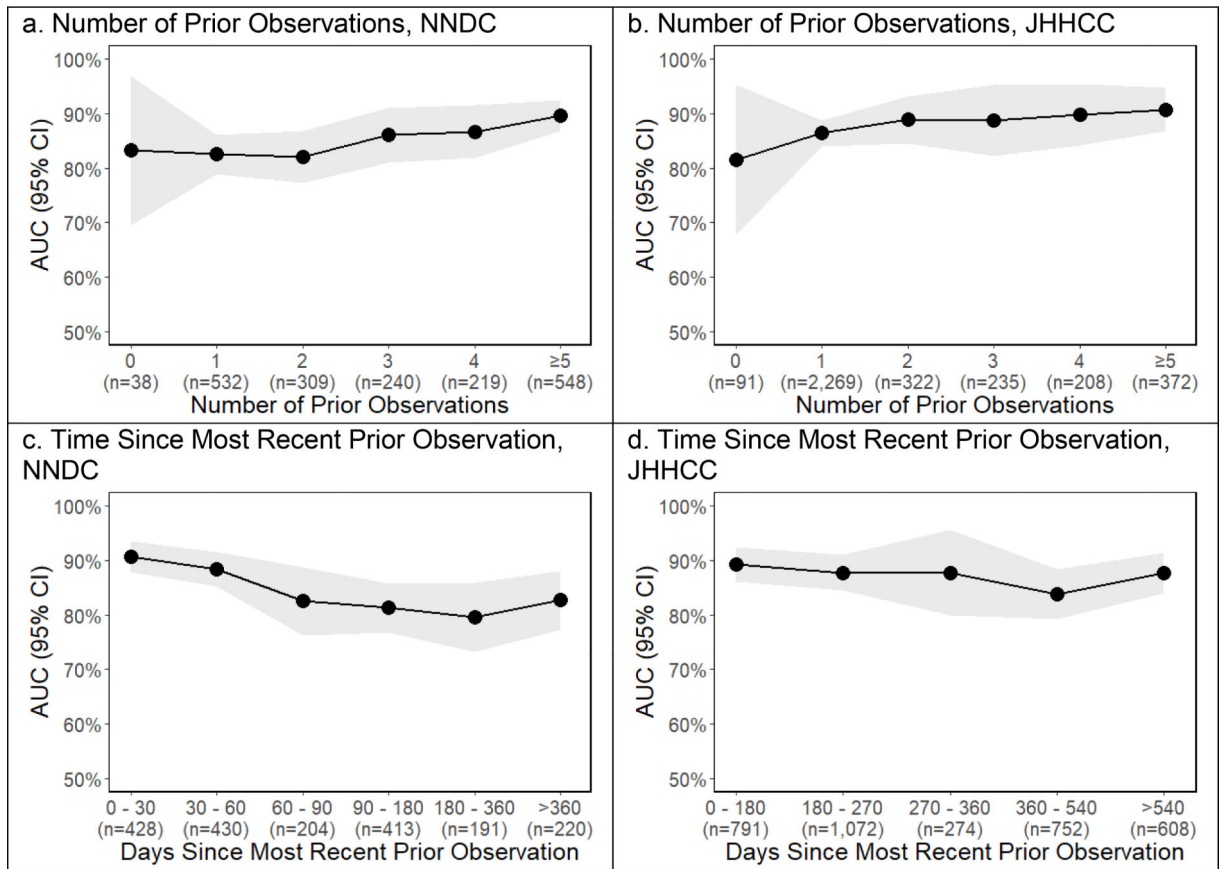
Table 1:

Demographic Characteristics of 2,444 Participants in the National Network of Depression Centers and Johns Hopkins HIV Clinical Cohort

	All Participants		Participants with >1 clinic visit	
	NNDC n=1,234	JHHCC n=1,210	NNDC n=708	JHHCC n=1,205
Age, median [IQR]	44 [31 to 57]	52 [45 to 58]	46.0 [32 to 58]	52 [45 to 58]
Female, % (n)	66.9% (826)	36.9% (446)	65.7% (465)	36.9% (445)
Race, % (n)				
Black	10.0% (123)	84.0% (1,015)	8.6% (61)	84.0% (1,012)
White	83.2% (1,027)	7.4% (89)	85.3% (604)	7.4% (89)
Hispanic	3.1% (38)	0.9% (11)	2.8% (20)	0.9% (11)
Other/Unknown	3.7% (46)	7.8% (94)	3.2% (23)	7.7% (93)
Number of Clinic Visits with Symptoms Recorded, median [IQR]	2 [1 to 4]	12 [8 to 16]	4 [2 to 6]	12 [8 to 16]
Education				
Grade School	0.5% (6)	N/A	0.7% (5)	N/A
Some High School	3.2% (40)	N/A	2.7% (19)	N/A
High School Grad or GED	13.5% (167)	N/A	12.4% (88)	N/A
Technical or Associate's Degree	7.4% (91)	N/A	6.5% (46)	N/A
Some College	25.7% (317)	N/A	24.6% (174)	N/A
Bachelor's Degree	26.4% (326)	N/A	27.4% (194)	N/A
Advanced/Professional Degree	23.3% (287)	N/A	25.7% (182)	N/A
Employment				
Unemployed	33.1% (408)	N/A	31.1% (220)	N/A
Homemaker	13.5% (166)	N/A	13.7% (97)	N/A
Part Time or Occasional	16.4% (202)	N/A	14.1% (100)	N/A
Full Time	37.1% (458)	N/A	41.1% (291)	N/A
HIV Acquisition Risk Factor, % (n)				
MSM [*]	N/A	20.6% (249)	N/A	20.7% (249)
IDU [†]	N/A	28.6% (346)	N/A	28.5% (344)
MSM+IDU	N/A	2.3% (28)	N/A	2.3% (28)
Heterosexual Contact	N/A	41.1% (497)	N/A	41.1% (495)
Unknown	N/A	7.4% (89)	N/A	7.4% (89)
Recent[‡] Diagnosis of Depression	N/A	27.4% (332)	N/A	27.5% (331)
Recent[‡] Diagnosis of Anxiety	N/A	6.0% (73)	N/A	6.1% (73)
Recent[‡] Diagnosis of Alcohol Use Disorder	N/A	9.8% (119)	N/A	9.9% (119)
Recent[‡] Diagnosis of Opioid Use	N/A	10.4% (126)	N/A	10.4% (125)
Recent[‡] Diagnosis of Cocaine Use	N/A	9.3% (113)	N/A	9.3% (112)

	All Participants		Participants with >1 clinic visit	
	NNDC n=1,234	JHHCC n=1,210	NNDC n=708	JHHCC n=1,205
Taking[‡] Mental Health Medication	N/A	28.6% (346)	N/A	28.5% (344)

All participants were used in fitting models. Only participants with >1 clinic visit were used in validating the algorithm. p-values for all characteristics for All Participants vs those with >1 clnic visit in both cohorts were >0.05.

[*]
MSM = men who have sex with men.

[†]
IDU = injection drug use.

[‡]
Recent = within past two years. Sex, race, and education level, were measured at the baseline clinic visit. Employment, recent clinical diagnoses, and mental health medications were time-varying; measurements at the first visit are presented here.

Table 2:

Distribution of Observed Symptoms and Substance Use in 19,088 Clinic Visits

	All Visits		Last Visit per Participant	
	NNDC (4,342 visits)	**JHHCC (5,409 visits)**	**NNDC (708 visits)**	**JHHCC (1,205 visits)**
PHQ [*], mean (SD)	10.9 (7.2)	3.4 (4.5)	10.5 (7.1)	3.4 (4.5)
	(n=4,008)	(n=5,138)	(n=648)	(n=1,166)
GAD-7 [†], mean (SD)	8.5 (6.3)	2.2 (4.2)	8.3 (6.2)	2.2 (4.2)
	(n=3,982)	(n=2,868)	(n=636)	(n=592)
ASRM [‡], mean (SD)	2.5 (3.0)	N/A	2.4 (3.0)	N/A
	(n=3,884)		(n=602)	
AUDIT-C [§], mean (SD)	N/A	1.3 (2.0)	N/A	1.3 (2.0)
		(n=5,363)		(n=1,197)
Heroin Use, % (n)	N/A	6.4% (272)	N/A	5.0% (271)
		(n=4,272)		(n=1,193)
Cocaine Use, % (n)	N/A	11.4% (520)	N/A	9.6% (519)
		(n=4,566)		(n=1,196)
Days Since Prior Visit with Symptoms Recorded, median [IQR]	33 [14 to 78]	91 [46 to 133]	33 [14 to 78]	91 [47 to 133]

All participants were used in fitting models. Only participants with >1 clinic visit were used in validating the algorithm.

[*] PHQ = Patient Health Questionnaire (9-item scale for NNDC ranges from 0–27, 8-item scale for JHHCC ranges from 0–24).

[†] GAD-7 = Generalized Anxiety Disorder scale (range 0–21).

[‡] ASRM = Altman Self-Rating Mania Scale (range 0–20).

[§] AUDIT-C = Alcohol Use Disorders Identification Test consumption questions (range 0–12).

Table 3:

Area Under the ROC Curve [95% CI] for Dichotomized and Binary Outcomes

Outcome	Cohort	Predictor		
		Probability from Predictive Model	Mean of All Prior Observations	Most Recent Prior Observation
PHQ [*] 10	NNDC (n=648)	0.85 [0.82 – 0.88]	0.75 [0.70 – 0.79]	0.75 [0.71 – 0.79]
	JHHCC (n=1,166)	0.84 [0.80 – 0.88]	0.75 [0.71 – 0.79]	0.72 [0.68 – 0.76]
GAD-7 [†] 10	NNDC (n=636)	0.85 [0.82 – 0.88]	0.63 [0.57 – 0.69]	0.63 [0.57 – 0.69]
	JHHCC (n=592)	0.84 [0.79 – 0.88]	0.75 [0.71 – 0.78]	0.69 [0.65 – 0.73]
ASRM [‡] 6	NNDC (n=602)	0.77 [0.72 – 0.82]	0.62 [0.55 – 0.69]	0.57 [0.50 – 0.65]
AUDIT-C [§] 3 female/4 male	JHHCC (n=1,197)	0.90 [0.88 – 0.92]	0.90 [0.87 – 0.92]	0.86 [0.83 – 0.90]
Heroin Use	JHHCC (n=1,193)	0.88 [0.84 – 0.92]	0.64 [0.58 – 0.70]	0.64 [0.58 – 0.70]
Cocaine Use	JHHCC (n=1,196)	0.89 [0.85 – 0.92]	0.87 [0.84 – 0.91]	0.81 [0.77 – 0.85]

[*] PHQ = Patient Health Questionnaire (9-item scale for NNDC ranges from 0–27, 8-item scale for JHHCC ranges from 0–24).

[†] GAD-7 = Generalized Anxiety Disorder scale (ranges from 0–21).

[‡] ASRM = Altman Self-Rating Mania Scale (ranges from 0–20).

[§] AUDIT-C = Alcohol Use Disorders Identification Test consumption questions (ranges 0–12).