Original Paper

Development of a Recommendation Engine to University Student Mental Health Support Aligned With Stepped Care: Longitudinal Cohort Study

Pedro Velmovitsky[1,2], MSc, PhD; Charles Keown-Stoneman[3,4], MSc, PhD; Kaylen J Pfisterer[1,5], MSc, PhD; Julia Hews-Girard[6,7,8], MN, PhD; Joseph Saliba[1,2], BHSc; Shumit Saha[1,2,9], MSc, PhD; Scott Patten[6], MD, PhD; Nathan King[10,11], MSc, PhD; Anne Duffy[10,11,12], MSc, MD; Quynh Pham[1,2,13,14], MSc, PhD

[1]Centre for Digital Therapeutics, Toronto General Hospital Research Institute, University Health Network, Toronto, ON, Canada
[2]Institute of Health Policy, Management, and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[4]Applied Health Research Centre, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Canada
[5]Systems Design Engineering, University of Waterloo, Waterloo, Canada
[6]The Mathison Centre for Mental Health Research & Education, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[7]Faculty of Nursing, University of Calgary, Calgary, Canada
[8]Alberta Children's Hospital Research Institute, Alberta Children's Hospital, Calgary, Canada
[9]Department of Biomedical Data Science, School of Applied Computational Sciences, Meharry Medical College, Nashville, United States
[10]Department of Psychiatry, Queen's University, Kingston, ON, Canada
[11]Department of Public Health Sciences, Queen's University, Kingston, ON, Canada
[12]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[13]Telfer School of Management, University of Ottawa, Ottawa, Canada
[14]School of Public Health Sciences, Faculty of Health, University of Waterloo, Waterloo, Canada

Corresponding Author:
Pedro Velmovitsky, MSc, PhD
Centre for Digital Therapeutics
Toronto General Hospital Research Institute
University Health Network
Toronto General Hospital/R Fraser Elliot Building, 4th Floor
190 Elizabeth St
Toronto, ON, M5G 2C4
Canada
Phone: 1 (416) 340 4800 ext 4765
Email: pedro.velmovitsky@uhn.ca

Abstract

Background: Mental health challenges are prevalent among Canadian higher education students, with significant rates of depression and anxiety often going untreated due to reduced early detection, stigmatizing beliefs, and practical barriers. The U-Flourish longitudinal electronic survey study launched in 2018 engages new cohorts of incoming undergraduate students and repeatedly collects data about mental health and well-being and access to support.

Objective: U-Flourish survey data provide a unique opportunity to train evidence-based prediction risk models and a personalized recommendation engine to signpost students to indicated mental health support based on their own data.

Methods: Two approaches were integrated in developing the risk prediction models and recommendation engine: (1) clinically defined rules by experts in the field to detect current and predict the risk of future anxiety and depression and to signpost students to appropriate care using a stepped care approach and based on clinical factors (ie, self-harm and suicidal thoughts, symptom levels, and lifetime history); and (2) machine learning models, trained with additional data including family history, early adversity, and stress indicators, to predict future risks of clinically significant depression (9-item Patient Health Questionnaire) and anxiety (7-item Generalized Anxiety Disorder questionnaire). Models were created using the XGBoost algorithm and a 70:30 ratio for training and testing with 10-fold cross-validation.

Results: In total, 27.5% of students at entry to university from 2018 to 2023 were identified as having potentially clinically significant levels of anxiety and depression and signposted to university mental health services based on the clinically defined rules. Optimizing thresholds to reduce false negatives, the machine learning models predicted anxiety and depression over the year in students screening negative at baseline with accuracy comparable with reported clinical screening as evidenced by sensitivity ≥90% for all models trained. Models had high negative predictive value (≥89%), balanced against low specificity. Individuals identified at risk for anxiety or depression were signposted primarily to self-guided resources supporting proactive prevention. Model findings also demonstrated that abbreviated screens (2-item Patient Health Questionnaire [PHQ-2] and 2-item Generalized Anxiety Disorder Questionnaire [GAD-2]), with potential to reduce respondent burden and improve adherence, can be used without compromising sensitivity. Indeed, PHQ-2 displayed a 90% sensitivity and GAD-2 displayed a 92% sensitivity. Shapley additive explanations analyses revealed other predictive factors including childhood trauma, family history of mental illness, and functional impairment associated with reported depression and anxiety symptoms.

Conclusions: The risk prediction models and recommendation engine's dual approach rationalize support allocation and promote targeted early intervention and prevention, potentially improving capacity to address the increasing burden on university mental health services. Future directions include further refinement based on a larger harmonized and enriched dataset, independent validation, and implementation studies to estimate the complex factors that influence uptake, reach to services, and acceptability across more diverse student users.

(J Med Internet Res 2025;27:e72669) doi: 10.2196/72669

KEYWORDS

mental health; university students; machine learning; depression; anxiety; stress; university, stepped care, prevention, early intervention, early detection

Introduction

Approximately one third of Canadian postsecondary students screen positive for clinically significant symptoms of depression or anxiety [1]. Young adults transitioning to university are particularly vulnerable. They are faced with increased responsibility for managing their daily lives including the demands of higher education, greater autonomy, and making new relationships while becoming more independent from previous support systems [1]. Furthermore, the transition to university coincides with the peak period of risk for the emergence of mental disorders—with anxiety and depression being the most common [1]. Untreated, mental health disorders are associated with a number of negative outcomes likely to affect economic stability (eg, reduced employment opportunities), education access and quality (eg, poor academic performance and increased risk of school dropout), and quality of life (eg, persisting and increasing mental disorders) [2-4]. However, less than 10% of students across Canadian university campuses receive treatment, often due to stigma associated with mental health care, not understanding how or when to reach out for support, and other practical barriers such as perceived lack of time [5-7]. Despite the low proportion of symptomatic students seeking care, student demand for mental health services is outpacing service capacity. For example, between 2009 and 2015, counseling service utilization rates at universities were 8 times greater than enrollment growth [8]. This also reflects that university counseling is the most frequently visited campus mental health service for students with a spectrum of needs and severity of concerns [9,10].

The U-Flourish Study, launched at Queen's University in 2018, is a longitudinal successive cohort study of undergraduates designed to understand mental health trajectories and associated academic outcomes among university students. The study collects biannual data (at entry and completion of each academic year) from incoming and continuing undergraduates, including mental health symptoms, risk and lifestyle factors, and help seeking through validated measures. Examples of collected metrics include the 9-item Patient Health Questionnaire (PHQ-9) [11,12] and the 7-item Generalized Anxiety Disorder questionnaire (GAD-7) [13,14]. Identifiers are removed from the dataset. Early findings confirm high rates of anxiety and depressive symptoms among incoming students, which are associated with distress, academic impairment, and lifestyle challenges, including sleep difficulties and substance use. Data collection continued through the COVID-19 pandemic which, as anticipated, led to a worsening of mental health issues—students' sleep and mental health quality were significantly negatively affected, with increasing rates of insomnia, anxiety, and depressive symptoms as pandemic restrictions intensified [15-17]. Self-harm and suicidal ideation also worsened, particularly among female students [17]. On one hand, as confirmed by the U-Flourish study, mental health conditions are increasingly prevalent among students, who often fail to seek treatment; on the other hand, help-seeking students are already overburdening university resources. This is an unsustainable situation and has been recognized as a mental health crisis [18].

In this context, there is a gap that needs to be addressed, namely, early detection and mapping of at-risk students to individually indicated support. This can be done by leveraging student data to proactively assess their mental health status and provide actionable feedback while reducing the burden on health care services. In other words, there is a need for innovative and rationalized models of care that can meet individual student needs while efficiently using existing campus resources. Student health services are under pressure to facilitate access to students with unmet needs, while ensuring that scarce resources are used with maximum efficiency.

The primary aim of this work was to develop and evaluate risk prediction models and an engine that combined 2 different approaches to provide treatment recommendations and signpost (ie, recommend) students to indicated levels of support rationalized in accordance with a stepped care model and based on student mental health data using (1) clinically defined rules and (2) machine learning (ML)–based predictive models. Specifically, we sought to explore the use of predictive factors such as mental health symptoms, substance use, and lifetime mental health history in a novel recommendation engine, which uses clinically defined rules and predictive models to signpost students to appropriate stepped levels of care. Such a digital resource would provide personalized recommendations in nonstigmatizing language and be available upon demand, increasing efficiency in support delivery, and potentially improving student mental health outcomes. If successful, this system and the general approach described in this paper could be implemented and scaled at other universities and higher education institutions.

The secondary aims of this work were to (1) inform future implementation of the recommendation engine into service delivery through exploration of alternative metrics (ie, the abbreviated questionnaires 2-item Patient Health Questionnaire [PHQ-2] and 2-item Generalized Anxiety Disorder Questionnaire [GAD-2] to reduce respondent burden) and (2) determine the strongest risk predictors of clinically significant anxiety or depression over the academic year, which could be used to further personalize prevention recommendations to mitigate future risk.

While many digital mental health interventions have been proposed to help students navigate university resources and provide support (eg, scheduling of appointments and self-guided resources) [19-21], these services are typically a "one-size-fits-all" solution to students and, as such, do not follow a stepped care approach, essential to provide appropriate personalized support recommendations to students and rationalize and reduce the burden on health care services. On the same token, such solutions are usually directed toward specific universities rather than providing a general approach that can be adapted and customized for different institutions. Finally, digital solutions for students are typically not co-designed with clinicians and do not apply more advanced analytical methods, such as ML, to gain further insights into student mental health. This work will address these aspects by co-designing the solution with experts, applying a hybrid rule-based and ML-based approach, and providing a general stepped care framework that can be adapted by different institutions.

Methods

Study Design

This work involved a systematic and phased approach to the design, development, and evaluation of risk prediction models, and a mental health recommendation engine that leveraged the richness of student data available from the longitudinal prospective U-Flourish study.

Requirements were defined by the team as follows. (1) Users were undergraduate university students. (2) The recommendation engine was not meant to replace in-person assessment or clinical triaging but rather to provide useful feedback to students about their estimated current mental health and risk over the academic year and automated recommendations only. (3) Prediction models and recommendation engine inputs were to be student-reported, mental health and mental health-related psychosocial and lifestyle variables comprising a subset of the U-Flourish survey self-report data. (4) These data were to be evidence-based and expert-informed as having strong predictive validity with mental health outcomes. (5) Recommendation outputs were to signpost students to the least intensive care level appropriate for their reported mental health status, that is, they were to follow a stepped care approach.

Next, we first describe the U-Flourish study data.

U-Flourish Data

The U-Flourish Student Well-being Survey study [22] was launched at Queen's University in the Fall of 2018. After a robust student-led engagement campaign, all incoming first-year students were invited to complete a web-based well-being survey via their university email. Students who completed the baseline Fall survey in mid-September were then invited to complete a Spring follow-up survey at the end of the academic year (mid-March) before final examinations. Fall and Spring surveys were subsequently sent out to students who completed each prior survey for up to 5 years. This study used data from the first 5 cohorts of students who participated in the U-Flourish survey, spanning the 2018-2019 to 2022-2023 academic years, and who had completed a follow-up survey in the subsequent Spring term. U-Flourish includes responses from 10,823 students with up to 5 years of follow-up data; a total of 4843 (44.8%) students completed at least 1 follow-up survey. This dataset is used to inform this work.

The baseline U-Flourish survey collected information on demographics, distal and proximal risk factors for mental health problems, and measures of current mental health and well-being using validated measures [22]. Subsequent follow-up surveys collected repeated data on psychosocial risk factors, mental health and well-being, and experiences over the academic year, including with mental health services. The specific subset of variables used in this study, defined in conjunction with the study clinicians, includes recreational drug use and binge drinking (Table 1), symptoms of depression (PHQ-9) [23], anxiety (GAD-7) [24], personal and familial history of mental illness, adverse childhood experiences, suicidal thoughts and behaviors and self-harm (Columbia-Suicide Severity Rating Scale, C-SSRS [25]), and self-perceived stress (Perceived Stress Scale–4 [PSS-4]) (Table 2).

Each survey was completed in a 2-week window with email reminders sent out at regular intervals. The survey was run on the Qualtrics platform. Participation was encouraged through a student-led engagement campaign that targeted first-year students and included in-class presentations, web-based and in-person advertisements, and booths at student residences and welcoming events. Incentives offered for participation included a free coffee voucher and entry into a draw to win an iPad.

Student participants were provided a letter of information outlining the study aims and potential risks and benefits to participation, and they were provided consent to access and complete the survey.

It should be noted that nonsuicidal self-harm was collected using the C-SSRS [25], while item 9 of the PHQ-9 questionnaire [11,12] refers to thoughts of being better off dead or hurting oneself. Except otherwise stated, we will consider self-harm as the C-SSRS measure (self-harm using C-SSRS and item 9 of the PHQ-9 in different decision nodes).

Table 1. List of recreational drug use variables.

Variable description	Response options	Threshold used in rules
Nonprescribed sleeping pills, past month	0 = Never, 1 = Less than once a week, 2 = Once a week, 3 = 2-3 times a week, 4 = 4+ times a week, and 999 = prefer not to answer	≥2
Nonprescribed stimulants or wake-up pills, past month	0 = Never, 1 = Less than once a week, 2 = Once a week, 3 = 2-3 times a week, 4 = 4+ times a week, and 999 = prefer not to answer	≥2
Cannabis, past month	The first 2 terms defined options as: 0 = Never, 1 = Less than once a week, 2 = Once a week, 3 = 2-3 times a week, 4 = 4+ times a week, and 999 = prefer not to answer For other terms: 0 = Never, 1 = Less than once a month, 2 = 1-3 days a month, 3 = 1-2 days a week, 4 = 3-4 days a week, and 5 = Every day or nearly every day	≥2 (first 2 terms) ≥3 (other terms)
Pain killers or opiates, past month	0 = Never, 1 = Less than once a week, 2 = Once a week, 3 = 2-3 times a week, 4 = 4+ times a week, and 999 = Prefer not to answer	≥2
Psychedelics, past month	0 = Never, 1 = Less than once a week, 2 = Once a week, 3 = 2-3 times a week, 4 = 4+ times a week, and 999 = Prefer not to answer	≥2
Cocaine, past month	0 = Never, 1 = Less than once a month, 2 = 1-3 days a month, 3 = 1-2 days a week, 4 = 3-4 days a week, and 5 = Every day or nearly every day	≥3
Other street drugs (eg, opioids, LSD[a], MDMA[b])	0 = Never, 1 = Less than once a month, 2 = 1-3 days a month, 3 = 1-2 days a week, 4 = 3-4 days a week, and 5 = Every day or nearly every day	≥3
Prescription drug without a prescription to get high, buzzed, or numbed out	0 = Never, 1 = Less than once a month, 2 = 1-3 days a month, 3 = 1-2 days a week, 4 = 3-4 days a week, and 5 = Every day or nearly every day	≥3
Nonprescribed medication to enhance academic performance (eg, modafinil and stimulant medication)	0 = Never, 1 = Less than once a month, 2 = 1-3 days a month, 3 = 1-2 days a week, 4 = 3-4 days a week, and 5 = Every day or nearly every day	≥3
Binge drinking, 5+ alcoholic drinks on one occasion	0 = Never, 1 = Less than monthly, 2 = Monthly, 3 = Weekly, and 4 = Daily or almost daily	≥3

[a]LSD: lysergic acid diethylamide.

[b]MDMA: methylenedioxymethamphetamine.

Table 2. A priori set of features.

Variable names	Description
PHQ-9[a] Items: Over the past 2 weeks, how often have you been bothered by the following problems: 0 = Not at all, 1 = Several days, 2 = Over half the days, and 3 = Nearly every day?	
PHQ9_1_1	Little interest or pleasure in doing things
PHQ9_2_1	Feeling down, depressed, or hopeless
PHQ9_3_1	Trouble falling or staying asleep or sleeping too much
PHQ9_4_1	Feeling tired or having little energy
PHQ9_5_1	Poor appetite or overeating
PHQ9_6_1	Feeling bad about yourself, or that you are a failure, or have let yourself or your family down
PHQ9_7_1	Trouble concentrating on things, such as reading the newspaper or watching TV
PHQ9_8_1	Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual?
PHQ9_9_1	Thoughts that you would be better off dead or of hurting yourself in some way
PHQ9_DIFF_1	If checked off any problems (PHQ-9), how difficult have these made for you to do your work, take care of things at home, or get along with other people? 0 = Not at all difficult, 1 = Somewhat difficult, 2 = Very difficult, and 3 = Extremely difficult
GAD-7[b] Items: Over the past 2 weeks, how often have you been bothered by the following problems: 0= Not at all, 1= Several days, 2= Over half the days, and 3 = Nearly every day?	
GAD7_1_1	Feeling nervous, anxious, or on edge
GAD7_2_1	Not being able to stop or control worrying
GAD7_3_1	Worrying too much about different things
GAD7_4_1	Trouble relaxing
GAD7_5_1	Being so restless that it is hard to sit still
GAD7_6_1	Becoming easily annoyed or irritable
GAD7_7_1	Feeling afraid as if something awful might happen
GAD7_DIFF_1	If checked off any problems (GAD-7), how difficult have these made for you to do your work, take care of things at home, or get along with other people? 0 = Not at all difficult, 1 = Somewhat difficult, 2 = Very difficult, and 3 = Extremely difficult
Disorders: Have you ever been diagnosed with any of the following mental health conditions or learning difficulties? 1 = Yes; 2 = No	
ANXIETY_1	Anxiety disorder (eg, PTSD[c], OCD[d], panic disorder, social anxiety disorder, and generalized anxiety disorder)
MOOD_1	Mood disorder (eg, depression, dysthymia, and bipolar disorder)
PSYCHOTIC_1	Psychotic disorder (eg, schizophrenia and drug-induced psychosis)
EATING_1	Eating disorder (eg, bulimia nervosa, anorexia nervosa, and binge eating disorder)
NEURO_1	Neurodevelopmental disorder (eg, autism spectrum)
SUBSTANCE_1	Substance use disorder (eg, cannabis and alcohol)
EMERGE_1	In your lifetime, have you ever visited a hospital emergency department or been admitted to a hospital for help with a mental health condition?
Family history: Have any of your first-degree blood relatives (eg, biological parents or siblings) ever been diagnosed with any of the following mental health conditions or learning difficulties? 1 = Yes; 2 = No	
FANXIETY_1	Anxiety disorder (eg, PTSD, OCD, panic disorder, social anxiety disorder, and generalized anxiety disorder)
FMOOD_1	Mood disorder (eg, depression, dysthymia, and bipolar disorder)
FPSYCHOTIC_1	Psychotic disorder (eg, schizophrenia and drug-induced psychosis)
FSUBSTANCE_1	Substance use disorder (eg, cannabis and alcohol)
Abuse: In terms of difficult childhood experiences… 1 = Yes; 2 = No	

Variable names	Description
REJECTED_1	When you were a child or teenager was someone in your household very harsh, critical, or rejecting toward you?
PHYSABUSE_1	When you were a child or teenager were you ever hit repeatedly with an implement (such as a belt or stick) or punched, kicked, or burnt by someone in the household?
BULLYING_1	When you were a child or a teenager were you physically or verbally bullied or teased very badly by peers?
SABUSE_1	When you were a child or teenager did you ever have any unwanted sexual experiences?
PDEATH_1	Parental death before 10 years old (either parent)
Suicidality: Thinking about your past, have you ever: 1 = Yes; 2 = No	
WISHDEAD_1	Wished you were dead or wished you could go to sleep and never wake up?
SUICIDEA_1	Had thoughts about ending your life?
SUICATMP_1	Made any suicide attempts?
SELFHARM_1	Hurt yourself on purpose without trying to end your life?
PSS-4[e] : Thinking about stress, please indicate in the last month, how often have you… 0 = Never, 1 = Almost never, 2 = Sometimes, 3 = Fairly often, and 4 = Very often	
STRESS_1_1	Felt that you were unable to control the important things in your life?
STRESS_2_1	Felt confident about your ability to handle your personal problems?
STRESS_3_1	Felt that things were going your way?
STRESS_4_1	Felt difficulties were piling up so high that you could not overcome them?

[a]PHQ-9: 9-item Patient Health Questionnaire.

[b]GAD-7: 7-item Generalized Anxiety Disorder questionnaire.

[c]PTSD: posttraumatic stress disorder.

[d]OCD: obsessive-compulsive disorder.

[e]PSS-4: Perceived Stress Scale–4.

Missing Data Imputation

Missing baseline variables were imputed using multivariate imputations by chained equations [26,27]. Ten multiple-imputed datasets were produced using a combination of predictive mean matching and polynomial regression models. For tools with individual item–level data (eg, individual questions from the GAD-7 and PHQ-9), the individual items were imputed using predictive mean matching, while the total scores were derived using passive imputation [26]. All features required for the ML models were included in the imputed variables, as well as variables collected as part of the U-Flourish study that may inform the imputation process that were not included in the ML model (eg, academic program of the student). A single-imputed dataset was then constructed for use in the ML models, using the mode for each imputed baseline variable. All imputation preparation and completion were performed using R (version 4.3.2: R Core Team) for Windows 64bit [28].

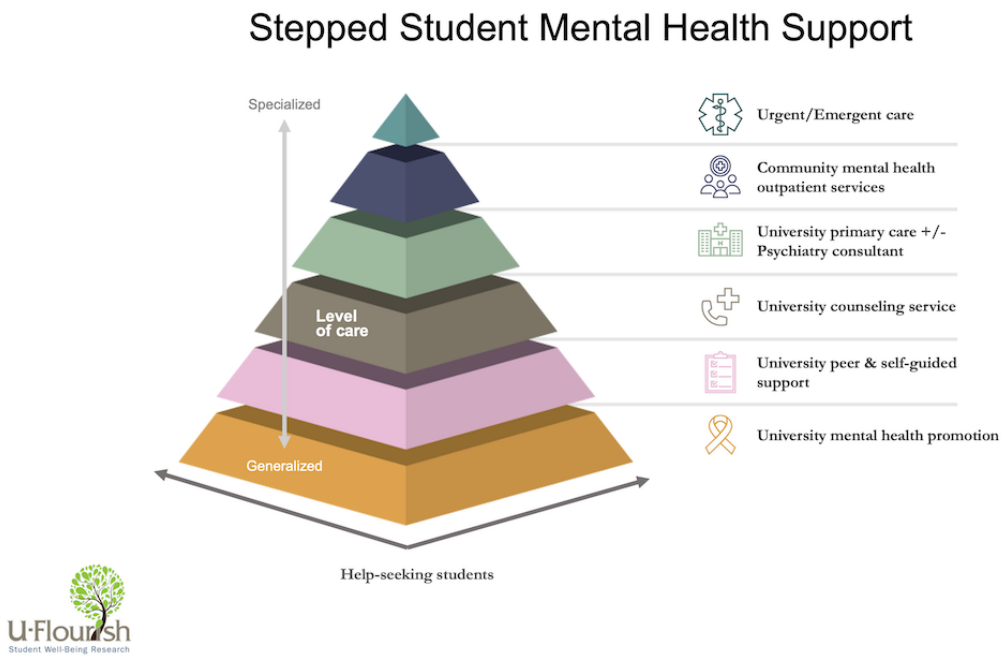
Defining Stepped Care Levels

Iterative co-design sessions were held biweekly between an interdisciplinary group of researchers and clinicians in health informatics, psychiatry, public health, nursing, and statistics. These sessions were initially used to define a consensus map of stepped care levels for the recommendation engine to address (Figure 1).

Of note, the stepped care levels were designed to be a general template that can be customized for any higher education institution and their respective mental health prevention and early intervention services. These university supports can then be classified within the general stepped care levels, which are based on the nature and intensity of the intervention.

Lower levels were defined as broader and less resource-intensive, emphasizing mental health promotion, psychoeducation, and self-guided support for managing common mental health problems (ie, insomnia, stress, and time management). Higher levels were defined as more resource-intensive and would typically involve in-person assessment and guided (supervised by a mental health professional) intervention.

Figure 1. General stepped care framework.

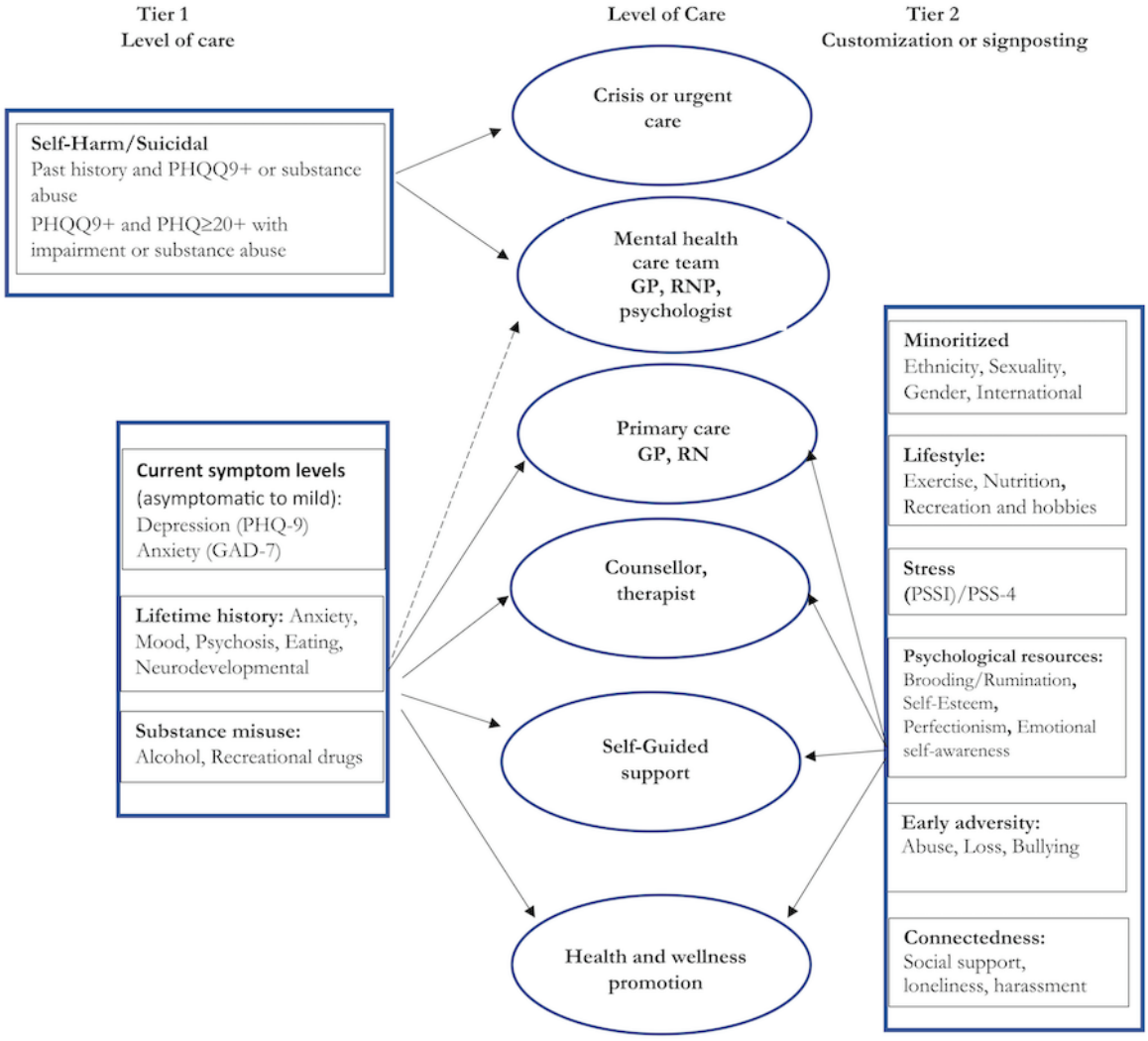


Prediction Models and Recommendation Engine Design and Development

After defining the stepped care levels, our biweekly co-design discussions focused on the design, development, and evaluation of the recommendation engine. First, we defined a priori features that were important contributors to mental health outcomes and that could be used to inform signposting to treatment recommendations. We also identified and described features that could be used for personalizing the treatment recommendations in the future (Figure 2). Then, as outlined in the following sections, two strategies were used to develop the engine: (1) we created a set of clinically defined rules, that is, a rules-based algorithm, to assess current mental health status and related level of treatment signposting; and (2) we developed ML-based models to predict future risk of worsening mental health for targeting mental health promotion and prevention.

Of note, tier 2 (right-hand side of Figure 2) is meant to provide more personalized filters based on individual preferences and characteristics (eg, international student, members of minority communities, and early adversity). While this level of customization is outside the scope of this work, we decided to include it as part of Figure 2 for future considerations.

Figure 2. Mapping of a priori clinically relevant U-Flourish features. List of features that are important contributors to mental health outcomes and that could be used to inform high-level signposting to treatment recommendations (tier 1), and features that could be used to further personalize the recommendations (tier 2). GAD-7: 7-item Generalized Anxiety Disorder questionnaire; GP: general practitioner; PHQ-9: 9-item Patient Health Questionnaire; PSS-4: Perceived Stress Scale–4; PSSI: Secondary Student Stressors Index; RN: registered nurse; RNP; registered practical nurse.



Rules-Based Algorithm for Categorizing Current Level of Treatment Needed

Overview of Levels

We sought consensus among the working group members—defined as agreement by all group members as to the best course of action—as to clinical decision points for clinically predictive variables at 2 workflow levels. The first level examines urgency, that is, establishing whether a student requires a timely in-person assessment. The second level provides less-intensive support recommendations based on depression and anxiety symptoms (level 2a), lifetime history (level 2b), and substance use (level 2c). Table 1 describes the variables used in the rules that consider drug use, while Table 2 describes the remaining variables used both in the rules and in the ML models. It should be noted that the ML models were developed in parallel with the rules.

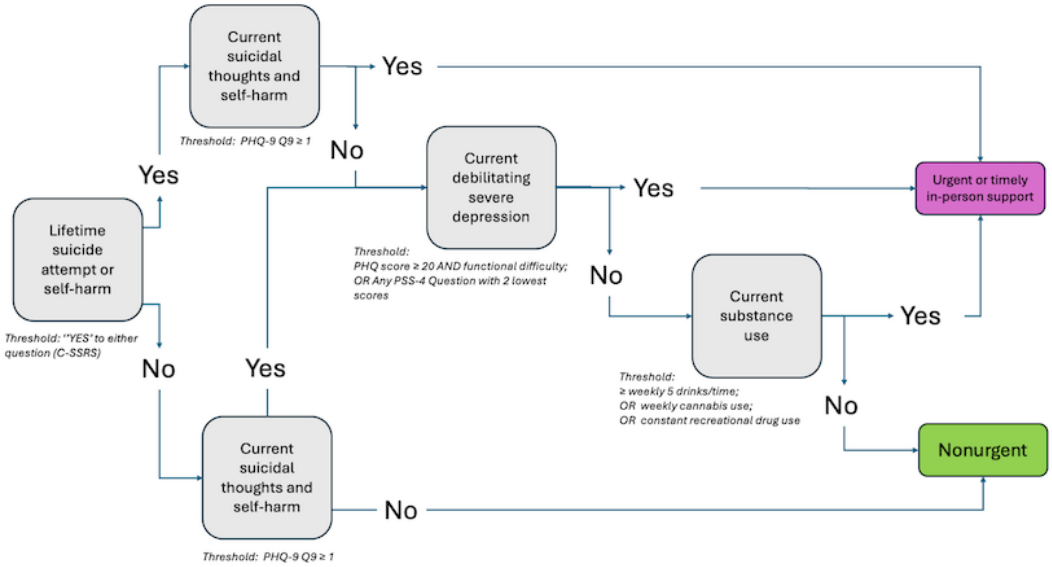
Level 1: Urgency of Assessment and Support Workflow

Signposting to either more urgent and timely, in-person assessment and support, or less immediate support (ie, less intensive levels of assessment and support) was based on four self-report criteria: (1) whether a student had a lifetime history of a suicide attempt or self-harm, (2) whether the student had current suicidal thoughts and self-harm, (3) whether the student was experiencing current debilitating and severe depression, and (4) whether the student was currently using substances (alcohol, cannabis, or recreational drugs).

Figure 3 shows the flow of decision points starting with a screen for lifetime suicide attempt or self-harm (C-SSRS), followed by a screen for current suicidal thoughts or self-harm (PHQ-9 item 9). If lifetime attempt or self-harm is screened positive and current thoughts or self-harm are negative (score of 0 for PHQ-9 item 9), or if the opposite is true (ie, lifetime attempt or self-harm is negative and current thoughts or self-harm is positive), the rule considers presence of severe depression and substance misuse. Aside from these decision points, in the

presence of any additional thresholds met the student is signposted to requiring a more timely and in-person assessment and intervention.

Figure 3. In-person urgent or timely level rules. Summary of the rules-based algorithm signposting students as either urgent (requiring more immediate or timely in-person assessment) or nonurgent (requiring less immediate follow-up, represented by rules in Figures 4-6). The algorithm shows the flow of decision points starting with a screen for lifetime suicide attempt or self-harm. If positive, it screens for current suicidal thoughts or self-harm. If negative, it screens for severe debilitating depression and current substance use. Aside from the first decision point (positive for lifetime suicide attempt), in the presence of any additional thresholds met, the student is signposted to requiring urgent intervention. C-SSRS: Columbia-Suicide Severity Rating Scale; PHQ-9: 9-item Patient Health Questionnaire; PSS-4: Perceived Stress Scale–4.



Level 2: Treatment Recommendation–Level Workflow

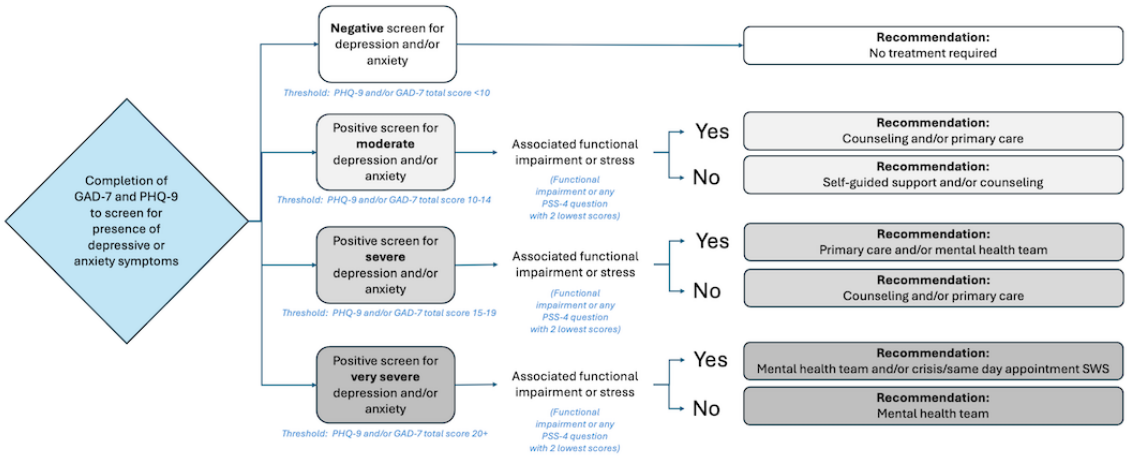
Overview of Level 2

All students who were classified as not requiring urgent or timely in-person assessment enter the treatment recommendation workflow which considers each of depression and anxiety symptoms (level 2a), lifetime history (level 2b), and substance use (level 2c). This process allows for robust and nuanced signposting of students to each of the stepped care levels outlined in Figure 1. These 3 sets of rules (ie, considering first by symptom, lifetime history, or substance use) are evaluated in parallel (ie, an individual could be screened for different sets of rules at the same time); in this case, the signposting should be the highest level of treatment recommended by the rules.

Level 2a: Presence of Depressive and Anxiety Symptoms

Figure 4 shows the decision nodes of level 2a. Recommendations were based on the combination of severity of depressive and anxiety symptoms defined from the PHQ-9 and the GAD-7, paired with associated functional impairment (asked after the PHQ-9 and GAD-7 questionnaires, as seen in Table 2) or stress defined from the PSS-4 (having any PSS-4 question achieving the lowest scores, that is, score of 3 or higher for questions 1 and 4, and score of 1 or lower for questions 2 and 3). Cutoffs were defined by the experts on the team.

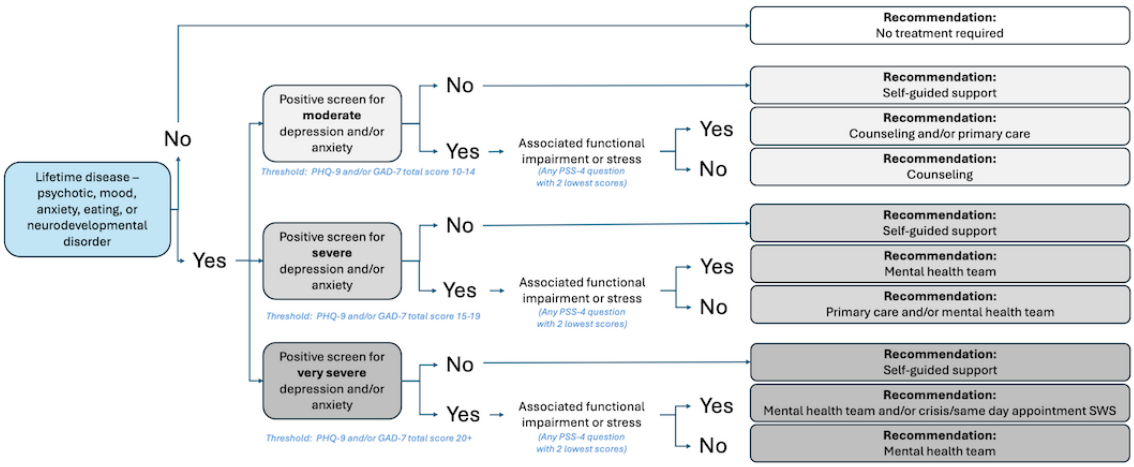
Figure 4. Rules-based algorithm—symptoms of depression and anxiety rules. Summary of the rules-based algorithm signposting students based on symptoms of depression and anxiety. Depending on the level of depressive or anxious symptoms (moderate, severe, or very severe) and presence of associated functional impairment or stress, individuals are signposted varied levels of stepped care from lower intensity (eg, self-guided support and counseling) to higher intensity (eg, mental health team and crisis/same day appointment Student Wellness Services). GAD-7: 7-item Generalized Anxiety Disorder questionnaire; PHQ-9: 9-item Patient Health Questionnaire; PSS-4: Perceived Stress Scale–4; SWS: Student Wellness Services.



Level 2b: Presence of Depressive and Anxiety Symptoms

Figure 5 shows the decision nodes of level 2b. To account for additional nuances of those with a lifetime history of diagnoses, an additional layer of self-guided support was added compared with level 2a. In addition to a lifetime diagnosis history (psychotic, mood, anxiety, eating, or neurodevelopmental disorders), recommendations were based on the combination of severity of screening positive for depression and anxiety defined from the PHQ-9 and the GAD-7, paired with associated functional impairment or stress (similarly to level 2a).

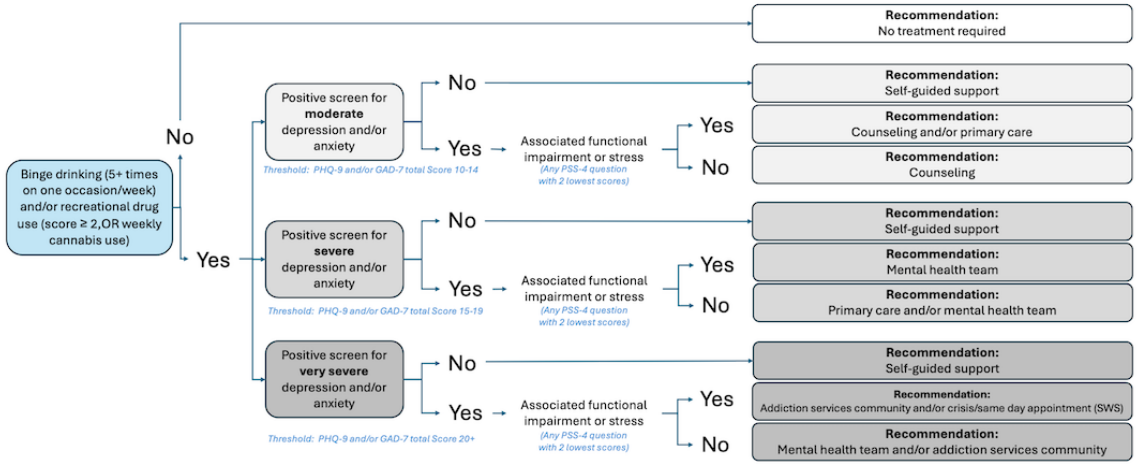
Figure 5. Rules-based algorithm—lifetime history rules. Summary of the rules-based algorithm signposting students based on symptoms of depression and anxiety and presence of lifetime conditions. Depending on the presence of these lifetime conditions, the level of depressive or anxious symptoms (moderate, severe, or very severe), and presence of associated functional impairment or stress, individuals are signposted to varied levels of stepped care from lower intensity (eg, self-guided support) to higher intensity (eg, mental health team and crisis/same day appointment Student Wellness Services). GAD-7: 7-item Generalized Anxiety Disorder questionnaire; PHQ-9: 9-item Patient Health Questionnaire; PSS-4: Perceived Stress Scale–4; SWS: Student Wellness Services.



Level 2c: Presence of Depressive and Anxiety Symptoms in Those Who Use Substances

Figure 6 shows the decision nodes of level 2c. As in level 2b, to account for additional nuances of those with a lifetime history of diagnoses, the same additional layer of self-guided support was included for those who use substances (ie, 5+ drinks on at least 1 occasion per week) or use recreational drugs (score ≥2 for any drug, or weekly cannabis use).

Figure 6. Rules-based algorithm—substance abuse rules. Summary of the rules-based algorithm signposting students based on symptoms of depression and anxiety and substance abuse. Depending on the presence of substance abuse, the level of depressive or anxious symptoms (moderate, severe, or very severe), and presence of associated functional impairment or stress, individuals are signposted to varied levels of stepped care from lower intensity (eg, self-guided support) to higher intensity (eg, addiction services in community and crisis/same day appointment Student Wellness Services). GAD-7: 7-item Generalized Anxiety Disorder questionnaire; PHQ-9: 9-item Patient Health Questionnaire; PSS-4: Perceived Stress Scale–4; SWS: Student Wellness Services.



ML-Based Models for Mental Health Worsening Risk Prediction

The clinically defined rules assess current mental health status. We also sought to predict future risk of screening positive for clinically significant anxiety and depressive symptoms, that is, for those who might be at risk in the future but screen negative at baseline—through the development of ML models. In particular, we trained the XGBoost model with a subset of U-Flourish features identified a priori by clinicians during the iterative sessions as model inputs (Table 2). As a reminder, these features were collected upon school entry, in the Fall term, that is, time 1 (Table 2), and included self-report indicators of depression (PHQ-9), anxiety (GAD-7), lifetime diagnosis or family history of diagnoses (anxiety, mood, psychotic, disordered eating, and neurodevelopmental conditions), personal and family history of substance use, childhood adversity (ie, abuse, neglect, or death of a parent), history of suicidality or self-harm, and current stress (Multimedia Appendix 1). This process was repeated for the PHQ-2 and GAD-2 questionnaires to evaluate whether these brief 2-item measures could be used to successfully predict depression or anxiety, without the remaining PHQ-9 and GAD-7 items, respectively. In other words, models using the PHQ-2 and the GAD-2 contained the first 2 items of each questionnaire in the model input to support more frequent collection of data points without increasing respondent burden.

Defining Thresholds for Output Labels

To label students as having clinically significant levels of anxiety or depression, a standard cutoff score of 10+ for each of the GAD-7 and PHQ-9 questionnaires was used [11,29-32]. Model outputs were created to predict anxiety and depression symptom scores measured by the GAD-7 and the PHQ-9 at the end of the academic year (ie, time 2: before Spring term examinations). In other words, if a student scored 10 or above on the PHQ-9 or GAD-7 at time 2, they were labeled as having a positive screen for depression or anxiety, respectively.

Classification Metrics

Binary classification was performed on the labeled data and the metrics of accuracy, precision, recall, and F_1-score were calculated to assess prediction performance. Accuracy was defined as the proportion of correct predictions compared with the total predictions. Precision was defined as the proportion of true-positive predictions (ie, if an individual is clinically anxious or depressed) and was used to indicate positive predictive value (PPV) and negative predictive value (NPV) as follows:

Precision:

- PPV: True positives / (True positives + False positives)
- NPV: True negatives / (True negatives + False negatives)

We used recall to indicate sensitivity and specificity. Sensitivity was defined as the proportion of true-positive predictions among all actual positives. Conversely, specificity was defined as the proportion of true-negative predictions among all actual negatives.

Recall:

- Sensitivity: True positives / (True positives + False negatives)
- Specificity: True negatives / (True negatives + False positives)

Finally, F_1-score is a typically used metric to evaluate ML models and represent the harmonic mean between precision and recall, calculated as follows:

F_1-score: 2 × (Precision × Recall) / (Precision + Recall)

Due to class imbalance (ie, having more examples of the negative class than the positive class), the F_1-macro metric was given priority, as it considers all classes as having equal weights when calculating the score and avoids artificially inflated results due to predictions from one class severely outperforming the other [33,34]. Libraries used include sci-kit learn and Python's XGBoost implementation.

Training and Testing

The datasets for ML training consisted of 2285 people who completed the U-Flourish survey in the Fall between 2018 and 2022, were not signposted to an in-person assessment (timely or urgent), were not screened in the level 2a rules, and had survey data for the following Spring of the academic year 2019-2023.

The dataset was split into train and test sets using a 70:30 ratio, stratified by clinical outcome (ie, anxiety and depression). For the depression outcome, the dataset's distribution had an imbalance: from1806 samples screened negative for depression to only 479 screening positive (1:3.77 ratio). Anxiety displayed a similar imbalance consisting of 1792 individuals not screening positive for anxiety compared with 493 screening positive (1:3.63 ratio). To alleviate this imbalance on model training, we included "scale_pos_weight" as part of XGBoost's training parameter, which scales the punishment of errors in predicting the minority class, causing the model to overcorrect. Values tested for this parameter were 1 (indicating the new parameter does not scale weights), followed by proportions based on the ratio of negative instances divided by positive instances, following best practices listed in the XGBoost's documentation for the general use of the parameter [35]. Finally, we evaluated and fine-tuned the model through 10-fold Cross-Validation through sci-kit learn's StratifiedKFold function.

Feature Importance

We used the Shapley additive explanations (SHAP) method [36] to investigate feature importance (ie, which features had the most influence on our model's predictions). SHAP gives each feature a score, called a "Shapley score," that shows how important that feature is for making accurate predictions [36]. To visualize the results, we used a SHAP summary plot that displays how each feature affects the model. Specifically, each point represents a sample from our data. The position of the point on the x-axis shows the feature's effect on the prediction (with positive values pushing toward a positive result and negative values toward a negative one). The color of each point represents the feature's value, with red indicating higher values and blue indicating lower ones. This setup aids in identifying not only which features are important but also how they interact with the model's output [36].

Comparative Analysis and Optimizations

To clinically assess the quality of the ML models, we conducted a comparative analysis between models using inputs from the PHQ-9, PHQ-2, GAD-7, and GAD-2, specifically evaluating their sensitivity (recall of the positive class), specificity (recall of the negative class), PPV, and NPV as the comparison metrics. As confirmed by clinicians during working meetings, there is no set "gold standard" target for these metrics in university students. We therefore defined our internal targets based on literature investigating clinical use of the PHQ and GAD questionnaires in student and general adult populations (Tables 3 and 4).

Table 3. Metrics based on student populations.

	Sensitivity (recall 1)	Specificity (recall 0)	PPV[a] (precision 1)	NPV[b] (precision 0)
PHQ-9[c] [11-13,29,37-40]	83%	83%	65%	95%
GAD-7[d] [13,14]	84%	65%	60%	92%
PHQ-2[e] [41]	N/A[f]	N/A	N/A	N/A
GAD-2[g] [42]	94%	86%	65%	98%

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GAD-7: 7-item Generalized Anxiety Disorder questionnaire.

[e]PHQ-2: 2-item Patient Health Questionnaire.

[f]N/A: not applicable.

[g]GAD-2: 2-item Generalized Anxiety Disorder Questionnaire.

Table 4. Metrics based on the general population.

	Sensitivity (recall 1)	Specificity (recall 0)	PPV[a] (precision 1)	NPV[b] (precision 0)
PHQ-9[c] [29,31,43-65]	80%	84%	39%	96%
GAD-7[d] [24,32,64-68]	71%	80%	29%	99%
PHQ-2[e] [29,40,48,50,51,60,62,65,69-74]	75%	79%	53%	88%
GAD-2[f] [32,51,64-66,68,75-77]	76%	75%	62%	88%

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GAD-7: 7-item Generalized Anxiety Disorder questionnaire.

[e]PHQ-2: 2-item Patient Health Questionnaire.

[f]GAD-2: 2-item Generalized Anxiety Disorder Questionnaire.

The aforementioned metrics were used to examine the trade-off between false positives and negatives, as the main goal of the models was to minimize the false-negative predictions. In other words, we paid close attention to how many false positives should be allowed, while simultaneously minimizing false negatives through iterative consultation with clinician experts on the team. The final goal of the model was to signpost students to appropriate treatment recommendations using their self-report data within a stepped care approach (ie, not to provide clinical triage). We used the metrics from Table 3 and 4 as a guide to define the minimum sensitivity that the system needs based on how these questionnaires are used in the literature.

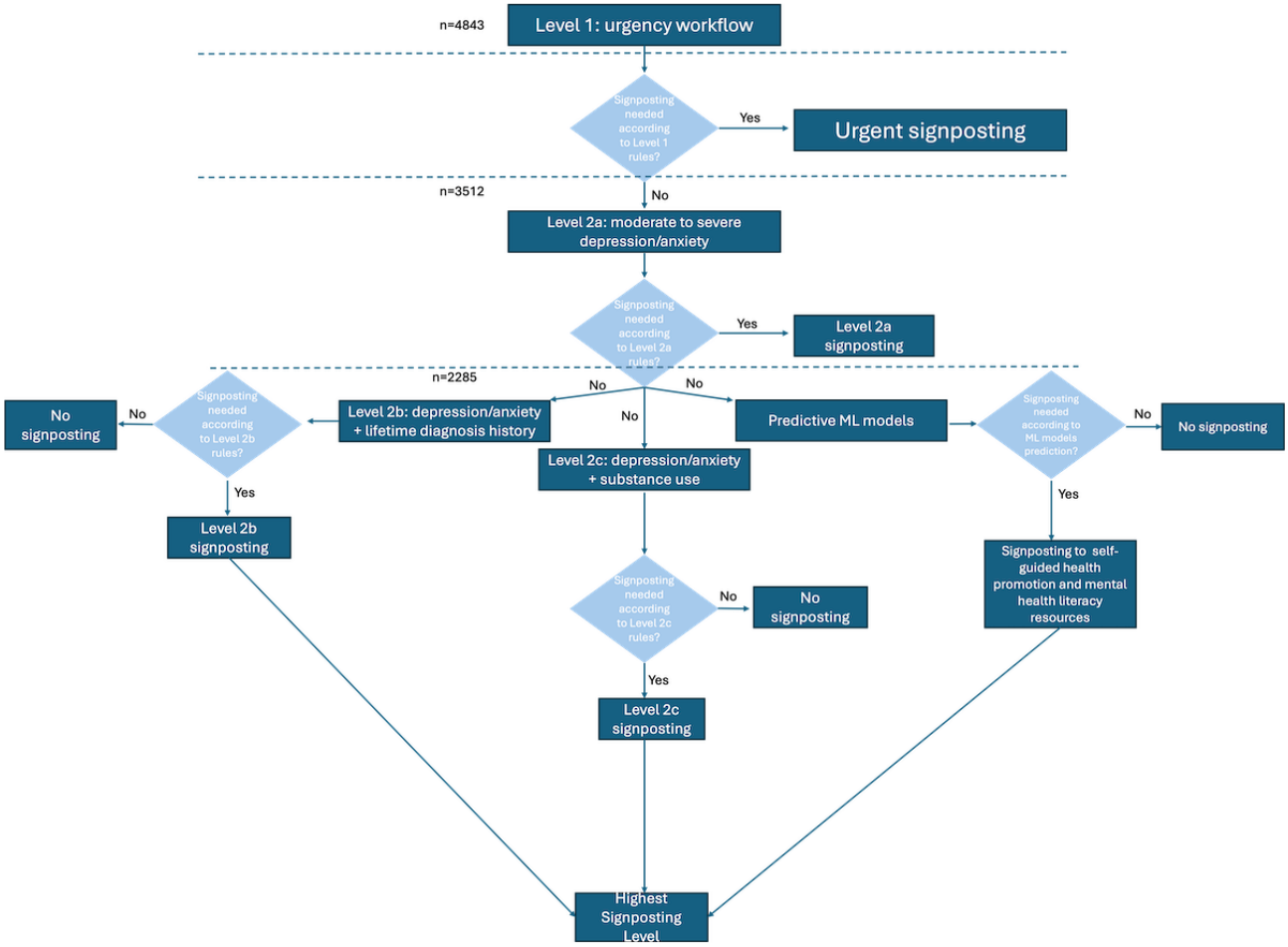
In this context, we explored adjusting the probability threshold used to classify observations. Instead of the usual 50% cutoff, we tested different thresholds to match sensitivity levels reported in Tables 3 and 4. We used the *yellowbrick* [78] library to visualize how different thresholds affect precision, recall, and the number of true and false positives or negatives. Such numbers were validated with clinicians during the iterative discussion sessions.

Integrating the ML Models Into the Recommendation Engine

Some students may not be currently experiencing clinically significant symptoms of anxiety and depression, but could be at future risk of developing said symptoms. The clinically defined rules capture students *currently* experiencing mental health challenges. The ML models predicts *future* risk of worsening mental health. As such, the clinically defined rules and ML models offer 2 complementary approaches to prediction models which are integrated into the recommendation engine as follows (Figure 7).

First, students are evaluated according to the level 1 rules (urgency-level workflow). The highest level of signposting requires timely in-person assessment. Students who do not meet the urgency workflow criteria are evaluated on level 2a (presence of depressive and anxiety symptoms). Levels 1 and 2a screen for immediate risk (eg, suicidality) as well as moderate to severe levels of anxiety and depression; if students do not meet these thresholds, they are then evaluated according to level 2b (presence of depressive and anxiety symptoms in those with lifetime diagnosis history), level 2c (presence of depressive and anxiety symptoms in those who use substances), and the predictive ML models in parallel.

Figure 7. Recommendation engine—flow of rules and machine learning models for signposting to treatment recommendations. Summary of the recommendation engine's flow of rules and machine learning (ML) models. First, students (n=4843) are screened based on the level 1 rule (Figure 3). If they meet the level 1 rule criteria, they are signposted to urgent or timely in-person assessment. If not, they are first screened by level 2a rule (n=3512). If they meet the level 2a rule criteria, they are signposted to a treatment recommendation based on the symptoms displayed (Figure 2a). If not, they (n=2285) are screened in parallel by the level 2b rule (Figure 5), level 2c rule (Figure 6), and by the ML model that predicts future risk. Each of these will generate a signposting to a treatment recommendation, the highest of which will be selected. ML: machine learning.



Ethical Considerations

The U-Flourish study was reviewed for ethical compliance and approved by the Queen's University and Affiliated Teaching Hospitals Research Ethics Board (HSREB PSIY-609-18). Informed consent was obtained from all participants, who are also made aware that participation is voluntary and does not impact academic standing. Students completing the Fall surveys receive a free coffee or tea valued at CAD $2 (US $1.45) from the local campus serving to be redeemed every Friday over the month the survey is running by showing proof of survey completion to the cashier. At the completion of both annual surveys (Fall and Spring), students are entered into a draw to win 1 of 5 iPads, valued at CAD $450 (US $325). Only deidentified data are shared.

Results

Descriptive Analyses

Table 5 provides the sample demographics of the total 4843 individuals in the dataset, the remaining 3512 individuals who were not screened by level 1 rules, and the 2285 individuals who were part of the ML models.

Consistently, they were predominantly female and white, aged between 16 and 20 years. When looking at the total population (n=4843), the PHQ-9 score had a mean PHQ-9 value of 7.72 (SD 5.92; scores range from 1 to 27) and mean GAD-7 value of 8.01 (SD 5.55; scores range from 1 to 21). These scores decrease per group as expected, since each group moves from higher need of assessment to lowest.

Table 5. Sample demographics.

Characteristics	Participant groups		
	In the dataset (n=4843)	Not screened by level 1 rules (n=3512)	Part of the ML models (n=2285)
Score, mean (SD)			
PHQ-9[a]	7.72 (5.92)	6.00 (4.72)	3.48 (2.56)
GAD-7[b]	8.01 (5.55)	6.92 (5.09)	3.96 (2.65)
Age (years), n (%)			
16-20	4552 (94)	4552 (94)	2125 (93)
21-25	242 (5)	242 (5)	114 (5)
26+	49 (1)	49 (1)	46 (2)
Sex, n (%)			
Male	1293 (26.7)	1011 (28.8)	779 (34.1)
Female	3477 (71.8)	2469 (70.3)	1499 (65.6)
Nonbinary	53 (1.1)	25 (0.7)	5 (0.2)
Prefer not to say	20 (0.4)	7 (0.2)	2 (0.1)
Ethnicity, n (%)			
White	3438 (71)	2494 (71)	1622 (71)
East/Southern Asian	1017 (21)	702 (20)	480 (21)
South Asian	339 (7)	246 (7)	137 (6)
Black	145 (3)	105 (3)	69 (3)
Latin	97 (2)	70 (2)	46 (2)
Indigenous	97 (2)	35 (1)	23 (1)
Middle Eastern	194 (4)	141 (4)	69 (3)
Other	48 (1)	35 (1)	23 (1)
International status, n (%)			
International student	436 (9)	421 (12)	201 (9)

[a]PHQ-9: 9-item Patient Health Questionnaire.

[b]GAD-7: 7-item Generalized Anxiety Disorder Questionnaire.

Rule-Based Algorithm for Categorizing Current Level of Treatment Needed

Overview of Levels

This section describes the results of validating the rules-based algorithm on the entire dataset. As described in the previous section, this algorithm entails 2 levels: level 1, which signposts to in-person urgent or timely level assessment, and level 2 comprising nonurgent assessments (this level is also divided into 3 rules and used to signpost students to treatment recommendations based on a stepped care approach, as seen in Figures 4-6). Of note, this study was meant to determine how the algorithms and models would behave with real student data; individual student data were deidentified and students were not contacted about recommendations.

Level 1: In-Person Urgent or Timely Level Rules

Out of a total of 4843 people, 1331 (27.5%) were recommended to the timely assessment or urgent care levels based on the application of these rules. The remaining 3512 individuals were then evaluated by the other rules (symptom-based, lifetime history, and substance misuse).

Level 2: Nonurgent Recommendation–Level Rules

Level 2a

Out of the 3512 individuals, 797 (22.7%) were signposted with moderate symptoms, 358 (10.2%) with severe symptoms, and 72 (2%) with very severe symptoms. A total of 2285 (65%) individuals were not signposted, that is, did not display at least moderate symptoms of anxiety or depression, and were thus eligible for the ML aspect of the recommendation system. In other words, such individuals are not at current risk based on anxiety or depressive symptoms but will still be assessed for future risk. As a reminder, an individual can be eligible for both the ML system and still be assessed by the lifetime history and substance misuse rules (levels 2b and c).

Out of the 1227 individuals signposted to a stepped care level, 53 (1.5%) were signposted to self-guided support and

counseling, 753 (21.4%) to counseling and primary care, 349 (9.9%) to primary care and mental health team, 1 (0.02%) to mental health team, and 71 (2%) to mental health team and crisis/same day appointment with wellness services.

Level 2b

In total, 7.2% (253/3512) of individuals are signposted with moderate symptoms, 4.2% (146/3512) with severe symptoms, and 1.1% (39/3512) with very severe symptoms. In total 11.2% (394/3512) of individuals are signposted based on their lifetime history alone without associated screen positives for depression or anxiety. In total, 76.3% (2680/3512) of individuals were not signposted.

Out of the 832 individuals signposted to a stepped care level, 394 (11.2%) were signposted to self-guided support, 27 (0.8%) to counseling, 226 (6.4%) to counseling and primary care, 4 (0.1%) to primary care and mental health team, 143 (4.1%) to mental health team, and 38 (1.1%) to mental health team and crisis/same day appointment with wellness services.

Level 2c

For this level, we have the largest number of individuals signposted based on substance use alone (1674/3512, 47.7%). The great majority of these, 1147 individuals, were signposted due to reporting binge drinking behavior (5 drinks or more per week on 1 occasion). In total, 14.6% (511/3512) are signposted based on moderate screen, 6% (211/3512) with severe screens for anxiety or depression, and 1.6% (56/3512) with very severe screens. In total, 30.2% (1060/3512) of individuals are not signposted.

Out of the 2452 individuals signposted, the 1674 individuals (47.7%) screened due to binge drinking were signposted to self-guided support, 25 (0.7%) to counseling, 486 (13.8%) to counseling and primary care, 5 (0.1%) to primary care and mental health team, 206 (5.9%) to mental health team, 1 (0.03%) to mental health team and addiction services community, and 55 (1.6%) to addiction services community and crisis/same day appointment with wellness services.

ML Model

As previously mentioned, the ML models were created using the XGBoost algorithm, taking as input a priori features (Table 2). We show the results of the ML models for predicting anxiety or depression, using PHQ-9, GAD-7, PHQ-2, and GAD-2 as inputs in Table 6. All models had sensitivity of 90% or above, aligning with Tables 3 and 4.

More details on the contingency table and evaluation metrics can be seen in Tables S1 through S8 in Multimedia Appendix 1. We call 0 or 1 the class labels representing the negative and positive classes, respectively. In the context of anxiety, these labels represent someone not having clinically significant anxiety, that is, GAD-7 score less than 10 (negative class), or someone having clinically significant anxiety, that is, GAD-7 score of 10 or more (positive class). In the context of depression, they represent someone not having clinically significant depression, that is, PHQ-9 score less than 10 (negative class), or someone having clinically significant depression, that is, PHQ-9 score of 10 or more (positive class).

The final prediction probability threshold for the GAD-7 model was 14%, as opposed to the usual cutoff of 50%. The threshold for the other models was 11%. This was done to adjust the sensitivity to align with Tables 3 and 4; in other words, we reduced the number of false negatives by increasing the sensitivity of the model, which also increased the number of false positives (decreasing the specificity). The optimal "scale_pos_weight" parameter for all models was set to 1, indicating that there was no scaling in the weight of errors predicting the minority class.

Table 6. Results of machine learning models.

	Sensitivity (recall 1)	Specificity (recall 0)	PPV[a] (precision 1)	NPV[b] (precision 0)
PHQ-9[c]	94%	18%	24%	93%
GAD-7[d]	90%	24%	24%	90%
PHQ-2[e]	90%	20%	23%	89%
GAD-2[f]	92%	19%	43%	90%

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GAD-7: 7-item Generalized Anxiety Disorder questionnaire.

[e]PHQ-2: 2-item Patient Health Questionnaire.

[f]GAD-2: 2-item Generalized Anxiety Disorder Questionnaire.

Feature Importance of ML Models

We used a SHAP summary plot to understand how different feature values influence the model's predictions. Refer to Table 2 for feature naming, definitions, and scoring. The following sections detail the results of the feature importance analysis.

Anxiety

GAD-7 Model

We found that items 3 and 5 of the GAD-7, lifetime anxiety disorder, item 3 of the PSS-4, and a family history of mood disorder were associated with a greater likelihood of screening positive for clinically significant anxiety at time 2. Looking

beyond these top 5 features, we can see that several PHQ-9 items, additional GAD-7 and PSS-4 items, family history of anxiety disorder, functional impairment, and diagnosis of disorders (eg, eating and psychotic) also contribute to the prediction. Figure S1 in Multimedia Appendix 1 shows the SHAP summary plot for the GAD-7 model.

GAD-2 Model

This model takes only the first 2 GAD-7 items as input rather than the entire questionnaire. Lifetime history of anxiety is the most important feature of this model, followed by item 3 of the PSS-4, item 2 of the GAD-7, family history of anxiety, and item 4 of the PSS-4. When looking into the importance of other features, we see that a history of abuse (most prominently bullying), functional impairment, other disorders such as mood and eating, self-harm, and additional items of PSS-4 and PHQ-9 are all important contributors to the prediction. Figure S2 in in Multimedia Appendix 1 shows the SHAP summary plot for the GAD-2 model

Depression

PHQ-9 Model

PHQ-9 items 3, 5, and 9 appear in the top 5 most important features and contribute toward a positive depression prediction. A family history of mood disorders, as well as item 4 of the PSS-4, is the remaining feature in the top 5. Other PHQ-9 items appear prominently as top contributors, as well as functional impairments. Like the GAD-2 model, bullying seems to contribute to clinically significant depression risk. GAD-7 items also appear as important predictors, indicating that anxiety and depression are highly correlated. Additional PSS-4 items, abuse, family history of disorders, and self-harm are all important predictors of this model. Figure S3 in Multimedia Appendix 1 shows the SHAP summary plot for the PHQ-9 model.

PHQ-2 Model

When changing the input to include only the first 2 items of the PHQ-9 and GAD-7, the topmost feature becomes the first PSS-4 item, followed by early experiences with sexual abuse, item 2 of the GAD-7, a visit to an emergency department due to a mental health condition, and functional impairment associated with depression. In addition to these features, additional PSS-4 items, family history of mental disorders (mood in particular), history of abuse (eg, bullying or rejection), suicide attempt, self-harm, and disorder diagnosis (eg, psychotic) also influence the prediction. Figure S4 in Multimedia Appendix 1 shows the SHAP summary plot for the PHQ-2 model.

Discussion

Principal Findings

This paper summarizes the development of a mental health calculator and support recommendation engine tailored for undergraduate students. U-Flourish longitudinal student survey data were used to develop the prediction models and corresponding recommendation engine. The prediction models, developed as part of the recommendation engine, were created using a hybrid approach, leveraging clinically defined rules and predictive models to improve early detection and accurate signposting. Findings from this study form the first major step in development of a digital application to provide accurate and useful feedback to students about their mental health, currently and over the course of the year, and to suggest indicated early intervention and prevention resources based on data they share. The aim is to improve prevention and early intervention of common mental health concerns and rationalize service utilization, using the 2 approaches of clinically defined rules and ML models, to assess current mental health status and predict future risk, respectively.

Main findings included that just more than 25% of students met the clinically defined rules for probable anxiety or depression and were signposted to in-person assessment through urgent, crisis, or student health primary care. Furthermore, ML models adjusted for optimal sensitivity were able to successfully predict the risk of students developing clinically significant anxiety or depression over the academic year. The most important predictors of these models included current depressive and anxiety symptoms, stress, early history of abuse, and family history of disorders. We showed that the use of the first 2 items of PHQ-9 and GAD-7 (PHQ-2 and GAD-2, respectively) can successfully predict risk, when coupled with other features, indicating that the full questionnaire data do not need to be collected for successful predictions, reducing the student burden.

Rule-Based Algorithm

Out of all students, 27.5% were considered appropriate for a timely or urgent signposting level for further assessment based on the level 1 workflow. When considering the next tier of less immediate recommendations (level 2a, Figure 4), more people were filtered by *moderate* depression and anxiety (797 individuals) than severe or very symptom levels (430 individuals). This is expected from a clinical standpoint, as each category in level 2a progressively evaluates higher levels of reported anxiety and depression. In general, the results reveal that the rules are consistent with other data and patterns in student mental health, which further support the validity of the rule-based algorithm.

In addition, when considering the substance use rules, almost 50% of students were signposted to treatment recommendations on the basis of substance use alone, without any associated mental health symptoms, specifically binge drinking (defined as 5 drinks on 1 occasion, weekly). This is consistent with other Canadian data on university students [79] and points to a concerning culture of alcohol abuse among undergraduate students.

ML Models

Four ML models were developed, considering as inputs the PHQ-9, GAD-7, PHQ-2, and GAD-2 (in addition to the other inputs), each predicting clinically significant depression or anxiety at time 2. When looking into the models, optimized also by probability classification threshold, results using only PHQ-2 and GAD-2 items as inputs were comparable with results from models using the complete questionnaires (Table 6) in terms of sensitivity. These results suggest that models trained on the abbreviated questionnaires can be used for successful prediction of depression and anxiety rather than relying on the full versions.

Of note, during the cross-validation process, the optimal models indicated that weight scaling was not needed. This suggests that, while there is a class imbalance, the ratio from majority to minority classes was not large enough to require a weighted correction. As previously discussed, we explored different probability thresholds as opposed to the traditionally used 50% cutoff of an item belonging to a certain class, in order to achieve a target sensitivity similar to the ones in Tables 3 and 4. The thresholds between all models were similar, with 0.14 for the GAD-7 and 0.11 for the other models.

With these adjustments, all models had a sensitivity higher than the ones obtained in the literature for Table 3, except the model that takes GAD-2 as input—with model sensitivity of 92% compared with the sensitivity of 94% in Table 3. However, with the exception of the PHQ-9 model, few studies among more than 40 peer-reviewed papers applied questionnaires specifically to student populations. In particular, only 1 study did so for the GAD-2, and no study reported evaluation metrics for the PHQ-2. A more accurate representation of the questionnaire metrics, therefore, can be found in Table 4, especially GAD-7, GAD-2, and PHQ-2. Although studies included as part of Table 4 focus on general adult populations rather than student ones, their variety and sample size make it a more realistic and reliable estimate.

Given this, the sensitivity for all 4 models was higher than the sensitivity achieved by each questionnaire in Table 4. The NPV of each model was slightly smaller. However, an expected consequence of adjusting the thresholds to increase sensitivity is a drop in specificity—as fewer false negatives are predicted, the number of false positives increases. This is reflected in the low specificity, PPV, and accuracy values. Absolute values for false positives can be seen in Tables S1, S3, S5, and S7 in Multimedia Appendix 1 in the upper right corner. The end goal of our system is not to conduct clinical triage but to provide students with a tool that recommends stepped care levels based on current needs. The ML models assess future risk to recommend lower-intensity, self-guided resources to students—which could mitigate risk through proactive prevention to guard against the development of symptoms and need for more intensive support. In this manner, having a higher number of false positives was deemed acceptable by clinician partners in the study, as this would not cause an undue burden on health care resources and might potentially reduce it. At worst, students who might not need any recommendation (false positives) would be recommended self-guided or nonclinical (eg, peer support) interventions. In other words, since these recommendations do not involve the use of clinical services, this would not increase the burden on already strained resources—in fact, if effective, the burden downstream on services would be expected to decrease with improved drinking behavior from the low-intensity interventions.

The features that contribute to the GAD-7 model also contribute to the GAD-2, albeit in a different order—for example, item 4 of the PSS-4 is the fifth most important contributor in the GAD-2 model but is in the 10th position on the GAD-7 model—suggesting that while these features are stable contributors to anxiety, interaction between features may also play a crucial role in determining their importance. As previously noted, the results of the GAD-2 and GAD-7 models were similar, indicating that other variables “compensate” for the removal of GAD-7 items. Indeed, for the GAD-2 model, the 2 most important features are similar to the previous model when removing items 3 and 5 of the GAD-7 as inputs. For example, the second most important feature of the GAD-7 model, lifetime anxiety history, became the most important feature of the GAD-2 model once item 3 of GAD-7 was not included as an input. Interestingly, the first item of GAD-7 (feeling nervous, anxious, or on edge) appears as the 13th most important feature, suggesting that the second GAD-7 item (not being able to stop or control worrying) is a much better predictor of future anxiety. The profile of the SHAP plot for PHQ-9 and PHQ-2 is different, with several features that did not strongly contribute to the PHQ-9 model appearing as the top contributors for the PHQ-2 model (such as sexual abuse and visit to emergency due to mental health problems). However, the PHQ-2 model still performed well, suggesting that the interaction between remaining features played a crucial role in predictions.

As outcomes vary, feature importance also changes. However, some similarities can be found when looking into the models grouped by depression or anxiety. Factors such as diagnosis of anxiety disorder, overwhelming stress as measured by PSS-4 items (in particular, items 3 and 4), and a family history of mood and anxiety disorders seem to be stable predictors of student anxiety levels. When predicting depression, PSS-4 items (in particular, item 4), history of childhood abuse, and emergency mental health visits also were stable predictors. These features could be collected by different sites trying to implement a similar survey and recommendation infrastructure.

It is important to note that to calculate the SHAP values in this paper, an average of all predictions was used, representing data from different users. In the future, if a platform was able to collect enough data from 1 student, it would be interesting to calculate SHAP values using an average of predictions from that specific student. This would provide a personalized ranking of the most important features for that student and, as such, would allow more personalized and tailored recommendations. As an example, in case a student is signposted to self-guided support, and in their personalized feature ranking item 2 of GAD-7 seems to be a major predictor, self-guided resources focusing on this item could be provided. On the same token, if a student indicated family history of a mental disorder, specific resources and services that deal with the issue can be recommended. This, however, would be dependent on having enough individual student data.

Limitations and Future Work

Of note, this work relied on self-report data to develop rule-based algorithm and prediction models. Therefore, for future refinement and validation the prediction models will need to be compared with clinical assessment (ie, structured clinical interviews) and the concordance of the automated recommendations compared with consensus clinical judgment. Future work in the system should also focus on tier 2 levels (Figure 2), which take into account additional features that could further personalize treatment recommendations. For example, if a student is recommended for counseling and has a history

of abuse, a counselor who specializes in this area could be highlighted. Future research should also investigate in more detail the longitudinal impact of mental health and the proposed intervention, that is, beyond the academic year.

Students with lived experience collaborated on the U-Flourish Survey study, defining and customizing the stepped care model, and drafting appropriate messaging as part of the U-Flourish research program. Future research on the recommendation engine focusing on refined development of accuracy and independent validation will continue to partner with student users, specifically, in refining the design, customization aspects (tier 2), and field testing. Furthermore, students from minoritized subgroups will continue to be involved in the creation of proper student-centric language for the recommendations to ensure culturally sensitive and engaging resources and that messaging is clear around automated suggestions based on shared, anonymized data. Since a conscious decision was made in this study to prioritize sensitivity over specificity, proper student-centric language is also essential to ensure no unnecessary stress to students who are classified at a potential future risk with the ML models.

When looking into the demographics (Table 5), the majority of respondents in the dataset were White and female for every group in the analysis, limiting generalizability. Therefore, we plan to test the algorithm and models on similar datasets collected from other universities and in different regions to ensure that they can be used accurately in populations with different characteristics. In particular, we are currently evaluating our system with a similar dataset to U-Flourish collected at the University of Oxford in United Kingdom [80]. Harmonized datasets will then be enriched with other data from subsets of students including clinical structured interviews, unstructured qualitative interviews, and linked medical records service use data. Once the application is further co-designed with students and validated with this larger harmonized and enriched datasets, we plan to carry out a rigorous multisite validation and implementation study in new cohorts of students at different institutions in different countries. Depending on privacy regulations, data could be stored locally at institutional servers as the infrastructure required to run these models would be minimal. Once implemented, we plan to monitor and investigate the longitudinal impact of the use of the application on early detection and prevention of anxiety and depression. Secondary outcomes will include effect on academic performance and well-being. Data collected as part of this deployment, properly anonymized, could also be used to improve prediction models and obtain a better balance between sensitivity and specificity and tailored recommendations.

Finally, it should be noted that while the prediction and recommendation engine models currently rely on self-report, the measures used are well validated and widely used in literature. We believe that their use is justified, given that the goal of the recommendation engine is to connect students to appropriate resources based on their symptoms without increasing the burden on health care services (ie, it is not intended as a diagnostic tool but as a screening measure to help connect high-risk students to services). That said, the measures could be further refined in terms of appropriate thresholds and the relative weighting of particular items in this target population. Furthermore, validation of the mental health calculator and recommendation engine is required in field studies using structured clinical interviews, which would further inform weighting and potential bias mitigation strategies. Field studies must also consider security and privacy risks. While outside the scope of this paper, a real-world implementation of the recommendation engine must account for secure data storage and collection. The predictive models should be housed in secure servers and receive only deidentified data as input to improve user safety. Any real-world deployment must also be constantly evaluated to identify and act on model drift (ie, changes in conditions that could lead to degradation in model performance) if it occurs.

Conclusions

Based on data from the U-Flourish longitudinal successive cohort survey study, we were able to develop prediction models and a support recommendation engine that integrates pragmatic clinical rule-based algorithms and ML models to identify current and future risk of anxiety and depression and signpost students to the indicated level of mental health support, improving early detection and proactive prevention. Using a stepped care model and signposting to indicated level of support will thereby rationalize service use and improve support capacity. resources. Collectively, the U-Flourish mental health calculator (classification, prediction, and recommendation engine) is an accessible and sustainable digital solution addressing barriers to care and tailored for university students with common mental health concerns.

Acknowledgments

The authors would like to thank the Rossy Family Foundation, McCall MacBain Foundation, Mach Gaensslen Foundation, Canadian Institutes of Health Research, TransformHF, and the Ted Rogers Centre for Heart Research for their support. They also thank P1Vital Products (Jonathan Kingslake and R&D team) for collaborating on adapting their mobile solution for this project, and all of the students at Queen's University for having their say and for partnering with them on the U-Flourish student-led engagement project.

Authors' Contributions

Conceptualization: QP (lead), AD (equal), PV (supporting), SP (supporting), JS (supporting), CKS (supporting)
Data curation: AD (lead), NK (supporting), PV (supporting), CKS (supporting)
Formal analysis: PV (lead), CKS (supporting), JS (supporting)

Funding acquisition: QP (lead), AD (equal)
Investigation: PV (lead)
Methodology: QP (lead), AD (equal), PV (supporting), CKS (supporting), SP (supporting), JHG (supporting), SS (supporting)
Project administration: QP (lead), AD (equal), PV (supporting)
Resources: AD (lead), CKS (supporting), NK (supporting)
Supervision: QP (lead), AD (equal), PV (supporting)
Validation: PV (lead), AD (supporting), CKS (supporting)
Visualization: PV (lead), AD (supporting)
Writing—original draft: PV (lead), AD (supporting), KJP (supporting), SP (supporting), CKS (supporting), JHG (supporting), NK (supporting)
Writing—review & editing: PV (lead), AD (supporting), KJP (supporting), SP (supporting), CKS (supporting), JHG (supporting), NK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Machine learning models additional material.
[DOCX File , 455 KB-Multimedia Appendix 1]

References

1. Adams KL, Saunders KE, Keown-Stoneman CDG, Duffy AC. Mental health trajectories in undergraduate students over the first year of university: a longitudinal cohort study. BMJ Open. 2021;11(12):e047393. [FREE Full text] [doi: 10.1136/bmjopen-2020-047393] [Medline: 34848401]
2. Alegría M, NeMoyer A, Falgàs Bagué I, Wang Y, Alvarez K. Social determinants of mental health: where we are and where we need to go. Curr Psychiatry Rep. 2018;20(11):95. [FREE Full text] [doi: 10.1007/s11920-018-0969-9] [Medline: 30221308]
3. Hale DR, Viner RM. How adolescent health influences education and employment: investigating longitudinal associations and mechanisms. J Epidemiol Community Health. 2018;72(6):465-470. [FREE Full text] [doi: 10.1136/jech-2017-209605] [Medline: 29615474]
4. Bruffaerts R, Mortier P, Kiekens G, Auerbach RP, Cuijpers P, Demyttenaere K, et al. Mental health problems in college freshmen: prevalence and academic functioning. J Affect Disord. 2018;225:97-103. [FREE Full text] [doi: 10.1016/j.jad.2017.07.044] [Medline: 28802728]
5. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry. 2010;10(1):113. [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
6. Salaheddin K, Mason B. Identifying barriers to mental health help-seeking among young adults in the UK: a cross-sectional survey. Br J Gen Pract. 2016;66(651):e686-e692. [FREE Full text] [doi: 10.3399/bjgp16X687313] [Medline: 27688518]
7. King N, Pickett W, McNevin SH, Bowie CR, Rivera D, Keown-Stoneman C, et al. U-Flourish Student Well-Being Academic Success Research Group. Mental health need of students at entry to university: baseline findings from the U-flourish student well-being and academic success study. Early Interv Psychiatry. 2021;15(2):286-295. [doi: 10.1111/eip.12939] [Medline: 32048460]
8. Center for Collegiate Mental Health. Center for collegiate mental health annual report. PennState. 2024. URL: https://ccmh.psu.edu/annual-reports [accessed 2025-08-12]
9. Patten SB, King N, Munir A, Bulloch AGM, Devoe D, Rivera D, et al. Transitions to campus mental health care in university students: determinants and predictors. J Am Coll Health. 2024;72(8):2455-2462. [doi: 10.1080/07448481.2022.2115303] [Medline: 36194448]
10. King N, Pickett W, Pankow K, Dimitropoulos G, Cullen E, McNevin S, et al. Access to university mental health services: understanding the student experience: L'accès aux services universitaires de santé mentale : comprendre l'expérience des étudiants. Can J Psychiatry. 2024;69(12):841-851. [FREE Full text] [doi: 10.1177/07067437241295640] [Medline: 39497426]
11. Adewuya AO, Ola BA, Afolabi OO. Validity of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. J Affect Disord. 2006;96(1-2):89-93. [doi: 10.1016/j.jad.2006.05.021] [Medline: 16857265]
12. Al-Ghafri G, Al-Sinawi H, Al-Muniri A, Dorvlo AS, Al-Farsi YM, Armstrong K, et al. Prevalence of depressive symptoms as elicited by Patient Health Questionnaire (PHQ-9) among medical trainees in Oman. Asian J Psychiatr. 2014;8:59-62. [doi: 10.1016/j.ajp.2013.10.014] [Medline: 24655629]

13. Pranckeviciene A, Saudargiene A, Gecaite-Stonciene J, Liaugaudaite V, Griskova-Bulanova I, Simkute D, et al. Validation of the Patient Health Questionnaire-9 and the generalized anxiety disorder-7 in Lithuanian student sample. PLoS One. 2022;17(1):e0263027. [FREE Full text] [doi: 10.1371/journal.pone.0263027] [Medline: 35085349]
14. Byrd-Bredbenner C, Eck K, Quick V. GAD-7, GAD-2, and GAD-mini: psychometric properties and norms of university students in the United States. Gen Hosp Psychiatry. 2021;69:61-66. [doi: 10.1016/j.genhosppsych.2021.01.002] [Medline: 33571925]
15. King N, Pickett W, Keown-Stoneman CDG, Miller CB, Li M, Duffy A. Changes in sleep and the prevalence of probable insomnia in undergraduate university students over the course of the COVID-19 pandemic: findings from the U-Flourish cohort study. BJPsych Open. 2023;9(6):e210. [FREE Full text] [doi: 10.1192/bjo.2023.597] [Medline: 37933532]
16. Appleby JA, King N, Saunders KE, Bast A, Rivera D, Byun J, et al. Impact of the COVID-19 pandemic on the experience and mental health of university students studying in Canada and the UK: a cross-sectional study. BMJ Open. 2022;12(1):e050187. [FREE Full text] [doi: 10.1136/bmjopen-2021-050187] [Medline: 35074809]
17. King N, Pickett W, Rivera D, Byun J, Li M, Cunningham S, et al. The impact of the COVID-19 pandemic on the mental health of first-year undergraduate students studying at a major Canadian university: a successive cohort study. Can J Psychiatry. 2023;68(7):499-509. [FREE Full text] [doi: 10.1177/07067437221094549] [Medline: 35450455]
18. Abrams Z. Student mental health is in crisis. Campuses are rethinking their approach. Oct 12, 2022. URL: https://www.apa.org/monitor/2022/10/mental-health-campus-care [accessed 2025-02-03]
19. University of Toronto. Student mental health resource. 2020. URL: https://mentalhealth.utoronto.ca/ [accessed 2025-05-06]
20. YOU at college. A comprehensive tool for student success. 2022. URL: https://youatcollege.com/ [accessed 2025-05-06]
21. Carleton University. Wellness services navigator. URL: https://wellness.carleton.ca/navigator/ [accessed 2025-05-06]
22. Goodday SM, Rivera D, Foran H, King N, Milanovic M, Keown-Stoneman CD, et al. U-Flourish university students well-being and academic success longitudinal study: a study protocol. BMJ Open. 2019;9(8):e029854. [doi: 10.1136/bmjopen-2019-029854] [Medline: 31455708]
23. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
24. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
25. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry. 2011;168(12):1266-1277. [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
26. van Buuren S. Flexible Imputation of Missing Data, Second Edition. London, England. Taylor & Francis; 2021.
27. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations inR. J. Stat. Softw. 2011;45(3):1-67. [doi: 10.32614/cran.package.mice]
28. The R project for statistical computing. R Core Team. 2023. URL: https://www.r-project.org/ [accessed 2025-02-10]
29. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Ann Fam Med. 2010;8(4):348-353. [FREE Full text] [doi: 10.1370/afm.1139] [Medline: 20644190]
30. Sun Y, Fu Z, Bo Q, Mao Z, Ma X, Wang C. The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital. BMC Psychiatry. 2020;20(1):474. [FREE Full text] [doi: 10.1186/s12888-020-02885-6] [Medline: 32993604]
31. Chibanda D, Verhey R, Gibson LJ, Munetsi E, Machando D, Rusakaniko S, et al. Validation of screening tools for depression and anxiety disorders in a primary care population with high HIV prevalence in Zimbabwe. J Affect Disord. 2016;198:50-55. [doi: 10.1016/j.jad.2016.03.006] [Medline: 27011359]
32. Christensen H, Batterham PJ, Grant JB, Griffiths KM, Mackinnon AJ. A population study comparing screening performance of prototypes for depression and anxiety with standard scales. BMC Med Res Methodol. 2011;11:154. [FREE Full text] [doi: 10.1186/1471-2288-11-154] [Medline: 22103584]
33. Velmovitsky PE, Alencar P, Leatherdale ST, Cowan D, Morita PP. Application of a mobile health data platform for public health surveillance: a case study in stress monitoring and prediction. Digit Health. 2024;10:20552076241249931. [FREE Full text] [doi: 10.1177/20552076241249931] [Medline: 39281042]
34. Velmovitsky PE. Use of Smart Technology Tools for Supporting Public Health Surveillance: From Development of a Mobile Health Platform to Application in Stress Prediction. Ontario. University of Waterloo; 2023.
35. XGBoost 2.1.1 documentation. XGBoost Parameters. URL: https://xgboost.readthedocs.io/en/stable/parameter.html [accessed 2024-10-21]
36. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. 2017.
37. Ghazisaeedi M, Mahmoodi H, Arpaci I, Mehrdar S, Barzegari S. Validity, reliability, and optimal cut-off scores of the WHO-5, PHQ-9, and PHQ-2 to screen depression among university students in Iran. Int J Ment Health Addict. 2022;20(3):1824-1833. [FREE Full text] [doi: 10.1007/s11469-021-00483-5] [Medline: 33495691]

38. Keenan L, Ingram Y, Green B, Daltry R, Harenberg S. Validation and clinical utility of the Patient Health Questionnaire-9 and center for epidemiologic studies depression scale as depression screening tools in collegiate student-athletes. J Athl Train. 2023;58(10):821-830. [FREE Full text] [doi: 10.4085/1062-6050-0558.22] [Medline: 37459388]
39. Mohamadian R, Khazaie H, Ahmadi SM, Fatmizade M, Ghahremani S, Sadeghi H, et al. The psychometric properties of the persian versions of the Patient Health Questionnaires 9 and 2 as screening tools for detecting depression among university students. Int J Prev Med. 2022;13:116. [FREE Full text] [doi: 10.4103/ijpvm.IJPVM_213_20] [Medline: 36276890]
40. Zhang Y, Liang W, Chen Z, Zhang H, Zhang J, Weng X, et al. Validity and reliability of Patient Health Questionnaire-9 and Patient Health Questionnaire-2 to screen for depression among college students in China. Asia Pac Psychiatry. 2013;5(4):268-275. [doi: 10.1111/appy.12103] [Medline: 24123859]
41. Khubchandani J, Brey R, Kotecki J, Kleinfelder J, Anderson J. The psychometric properties of PHQ-4 depression and anxiety screening scale among college students. Arch Psychiatr Nurs. 2016;30(4):457-462. [doi: 10.1016/j.apnu.2016.01.014] [Medline: 27455918]
42. Tran AGTT. Using the GAD-7 and GAD-2 generalized anxiety disorder screeners with student-athletes: empirical and clinical perspectives. Sport Psychol. 2020;34(4):300-309.
43. Azah MNN, Shah MEM, Shaaban J, Bahri LS, Rushidi WMWM, Yaacob MJ. Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. J Int Med Res. 2005;12(4):259-263.
44. Baron EC, Davies T, Lund C. Validation of the 10-item Centre for Epidemiological Studies Depression Scale (CES-D-10) in Zulu, Xhosa and Afrikaans populations in South Africa. BMC Psychiatry. 2017;17(1):6. [FREE Full text] [doi: 10.1186/s12888-016-1178-x] [Medline: 28068955]
45. Cassiani-Miranda CA, Cuadros-Cruz AK, Torres-Pinzón H, Scoppetta O, Pinzón-Tarrazona JH, López-Fuentes WY, et al. Validity of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adult primary care users in Bucaramanga, Colombia. Rev Colomb Psiquiatr (Engl Ed). 2021;50(1):11-21. [doi: 10.1016/j.rcp.2019.09.001] [Medline: 33648690]
46. Cholera R, Gaynes BN, Pence BW, Bassett J, Qangule N, Macphail C, et al. Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. J Affect Disord. 2014;167:160-166. [FREE Full text] [doi: 10.1016/j.jad.2014.06.003] [Medline: 24972364]
47. Cumbe VFJ, Muanido A, Manaca MN, Fumo H, Chiruca P, Hicks L, et al. Validity and item response theory properties of the Patient Health Questionnaire-9 for primary care depression screening in Mozambique (PHQ-9-MZ). BMC Psychiatry. 2020;20(1):382. [FREE Full text] [doi: 10.1186/s12888-020-02772-0] [Medline: 32698788]
48. Errazuriz A, Beltrán R, Torres R, Passi-Solar A. The validity and reliability of the PHQ-9 and PHQ-2 on screening for major depression in Spanish Speaking Immigrants in Chile: a cross-sectional study. Int J Environ Res Public Health. 2022;19(21):13975. [FREE Full text] [doi: 10.3390/ijerph192113975] [Medline: 36360856]
49. Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, et al. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatry Res. 2013;210(2):653-661. [FREE Full text] [doi: 10.1016/j.psychres.2013.07.015] [Medline: 23972787]
50. González-Sánchez A, Ortega-Moreno R, Villegas-Barahona G, Carazo-Vargas E, Arias-LeClaire H, Vicente-Galindo P. New cut-off points of PHQ-9 and its variants, in Costa Rica: a nationwide observational study. Sci Rep. 2023;13(1):14295. [FREE Full text] [doi: 10.1038/s41598-023-41560-0] [Medline: 37652965]
51. Hanlon C, Medhin G, Selamu M, Breuer E, Worku B, Hailemariam M, et al. Validity of brief screening questionnaires to detect depression in primary care in Ethiopia. J Affect Disord. 2015;186:32-39. [doi: 10.1016/j.jad.2015.07.015] [Medline: 26226431]
52. Indu PS, Anilkumar TV, Vijayakumar K, Kumar K, Sarma PS, Remadevi S, et al. Reliability and validity of PHQ-9 when administered by health workers for depression screening among women in primary care. Asian J Psychiatr. 2018;37:10-14. [doi: 10.1016/j.ajp.2018.07.021] [Medline: 30096447]
53. Belhadj H, Jomli R, Ouali U, Zgueb Y, Nacef F. Validation of the Tunisian version of the patient health questionnaire (PHQ-9). Eur psychiatr. 2020;41(S1):S523-S523. [doi: 10.1016/j.eurpsy.2017.01.695]
54. Kiely KM, Butterworth P. Validation of four measures of mental health against depression and generalized anxiety in a community based sample. Psychiatry Res. 2015;225(3):291-298. [doi: 10.1016/j.psychres.2014.12.023] [Medline: 25578983]
55. Mihić L, Knežević G, Lazarević LB, Marić NP. Screening for depression in the Serbian general population sample: an alternative to the traditional Patient Health Questionnaire-9 cut-off score. J Public Health (Oxf). 2024;46(1):e15-e22. [doi: 10.1093/pubmed/fdad204] [Medline: 37934963]
56. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Tanaka Y, Hosaka M, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry. 2018;52:64-69. [doi: 10.1016/j.genhosppsych.2018.03.007] [Medline: 29698880]
57. Bian C, Li C, Duan Q, Wu H. Reliability and validity of Patient Health Questionnaire: depressive syndrome module for outpatients. Sci. Res. Essays. 2011;6(2):278-282.
58. Carballeira Y, Dumont P, Borgacci S, Rentsch D, de Tonnac N, Archinard M, et al. Criterion validity of the French version of Patient Health Questionnaire (PHQ) in a hospital department of internal medicine. Psychol Psychother. 2007;80(Pt 1):69-77. [doi: 10.1348/147608306X103641] [Medline: 17346381]

59. Chen S, Fang Y, Chiu H, Fan H, Jin T, Conwell Y. Validation of the nine-item Patient Health Questionnaire to screen for major depression in a Chinese primary care population. Asia Pac Psychiatry. 2013;5(2):61-68. [doi: 10.1111/appy.12063] [Medline: 23857806]
60. Graham AK, Minc A, Staab E, Beiser DG, Gibbons RD, Laiteerapong N. Validation of the computerized adaptive test for mental health in primary care. Ann Fam Med. 2019;17(1):23-30. [FREE Full text] [doi: 10.1370/afm.2316] [Medline: 30670391]
61. Henkel V, Mergl R, Kohnen R, Maier W, Möller H-J, Hegerl U. Identifying depression in primary care: a comparison of different methods in a prospective cohort study. BMJ. 2003;326(7382):200-201. [FREE Full text] [doi: 10.1136/bmj.326.7382.200] [Medline: 12543837]
62. Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, Oikawa Y, et al. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. Gen Hosp Psychiatry. 2013;35(6):592-597. [doi: 10.1016/j.genhosppsych.2013.08.001] [Medline: 24029431]
63. Kim M, Jung S, Park JE, Sohn JH, Seong SJ, Kim B, et al. Validation of the Patient Health Questionnaire-9 and Patient Health Questionnaire-2 in the general Korean population. Psychiatry Investig. 2023;20(9):853-860. [FREE Full text] [doi: 10.30773/pi.2023.0100] [Medline: 37794667]
64. Vrublevska J, Trapencieris M, Rancans E. Adaptation and validation of the Patient Health Questionnaire-9 to evaluate major depression in a primary care sample in Latvia. Nord J Psychiatry. 2018;72(2):112-118. [doi: 10.1080/08039488.2017.1397191] [Medline: 29105551]
65. Villarreal-Zegarra D, Barrera-Begazo J, Otazú-Alfaro S, Mayo-Puchoc N, Bazo-Alvarez JC, Huarcaya-Victoria J. Sensitivity and specificity of the Patient Health Questionnaire (PHQ-9, PHQ-8, PHQ-2) and General Anxiety Disorder scale (GAD-7, GAD-2) for depression and anxiety diagnosis: a cross-sectional study in a Peruvian hospital population. BMJ Open. 2023;13(9):e076193. [FREE Full text] [doi: 10.1136/bmjopen-2023-076193] [Medline: 37714674]
66. Ahn J, Kim Y, Choi K. The psychometric properties and clinical utility of the Korean version of GAD-7 and GAD-2. Front Psychiatry. 2019;10:127. [FREE Full text] [doi: 10.3389/fpsyt.2019.00127] [Medline: 30936840]
67. Sidik SM, Arroll B, Goodyear-Smith F. Validation of the GAD-7 (Malay version) among women attending a primary care clinic in Malaysia. J Prim Health Care. 2012;4(1):5-11, A1. [FREE Full text] [Medline: 22377544]
68. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med. 2007;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
69. Aljabr QM, AlMubarak SS, Alsaqar FH, Alhelal AA, bin Obaid MR, Alshhakhs HH, et al. Prevalence of depression in primary health care patient in Saudi Arabia, Al Hassa using PHQ 9. Ann Med Health Sci Res. 2021;11(S3):185-192.
70. Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A, et al. Validity and utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in rural Chiapas, Mexico: a cross-sectional study. J Clin Psychol. 2017;73(9):1076-1090. [FREE Full text] [doi: 10.1002/jclp.22390] [Medline: 28195649]
71. Gelaye B, Wilson I, Berhane HY, Deyessa N, Bahretibeb Y, Wondimagegn D, et al. Diagnostic validity of the Patient Health Questionnaire-2 (PHQ-2) among Ethiopian adults. Compr Psychiatry. 2016;70:216-221. [FREE Full text] [doi: 10.1016/j.comppsych.2016.07.011] [Medline: 27567282]
72. Shaff J, Kahn G, Wilcox HC. An examination of the psychometric properties of the Patient Health Questionnaire-9 (PHQ-9) in a multiracial/ethnic population in the United States. Front Psychiatry. 2023;14:1290736. [FREE Full text] [doi: 10.3389/fpsyt.2023.1290736] [Medline: 38293592]
73. Bhana A, Mntambo N, Gigaba SG, Luvuno ZPB, Grant M, Ackerman D, et al. Validation of a brief mental health screening tool for common mental disorders in primary healthcare. S Afr Med J. 2019;109(4):278-283. [FREE Full text] [doi: 10.7196/SAMJ.2019.v109i4.13664] [Medline: 31084695]
74. Carey M, Boyes A, Noble N, Waller A, Inder K. Validation of the PHQ-2 against the PHQ-9 for detecting depression in a large sample of Australian general practice patients. Aust J Prim Health. 2016;22(3):262-266. [doi: 10.1071/PY14149] [Medline: 26306421]
75. Basaraba CN, Stockton MA, Sweetland A, Medina-Marino A, Lovero KL, Oquendo MA, et al. Does it matter what screener we use? A comparison of ultra-brief PHQ-4 and E-mwTool-3 screeners for anxiety and depression among people with and without HIV. AIDS Behav. 2023;27(4):1154-1161. [FREE Full text] [doi: 10.1007/s10461-022-03852-w] [Medline: 36209180]
76. Cano-Vindel A, Muñoz-Navarro R, Medrano LA, Ruiz-Rodríguez P, González-Blanch C, Gómez-Castillo MD, et al. PsicAP Research Group. A computerized version of the Patient Health Questionnaire-4 as an ultra-brief screening tool to detect emotional disorders in primary care. J Affect Disord. 2018;234:247-255. [doi: 10.1016/j.jad.2018.01.030] [Medline: 29549826]
77. Luo Z, Li Y, Hou Y, Zhang H, Liu X, Qian X, et al. Adaptation of the two-item generalized anxiety disorder scale (GAD-2) to Chinese rural population: a validation study and meta-analysis. Gen Hosp Psychiatry. 2019;60:50-56. [doi: 10.1016/j.genhosppsych.2019.07.008] [Medline: 31326672]
78. Yellowbrick v1.5 documentation. Yellowbrick: Machine Learning Visualization. URL: https://www.scikit-yb.org/en/latest/ [accessed 2025-02-03]

79. Summary: Canadian postsecondary education alcohol and drug use survey. Public Health Agency of Canada. 2021. URL: https://health-infobase.canada.ca/alcohol/cpads/ [accessed 2025-02-03]
80. Student Well-being. Nurture-U. URL: https://www.nurtureuniversity.co.uk/ [accessed 2025-05-05]

Abbreviations

C-SSRS: Columbia-Suicide Severity Rating Scale
GAD-2: 2-item Generalized Anxiety Disorder Questionnaire
GAD-7: 7-item Generalized Anxiety Disorder Questionnaire
ML: machine learning
NPV: negative predictive value
PHQ-2: 2-item Patient Health Questionnaire
PHQ-9: 9-item Patient Health Questionnaire
PPV: positive predictive value
PSS-4: Perceived Stress Scale–4
SHAP: Shapley additive explanations

Edited by J Sarvestan; submitted 14.02.25; peer-reviewed by D Gyimah, LP Gorrepati, VSK Kancharla; comments to author 11.03.25; revised version received 23.05.25; accepted 26.05.25; published 17.09.25

Please cite as:
Velmovitsky P, Keown-Stoneman C, J Pfisterer K, Hews-Girard J, Saliba J, Saha S, Patten S, King N, Duffy A, Pham Q
Development of a Recommendation Engine to University Student Mental Health Support Aligned With Stepped Care: Longitudinal Cohort Study
J Med Internet Res 2025;27:e72669
URL: https://www.jmir.org/2025/1/e72669
doi: 10.2196/72669
PMID: 40767642