Machine learning-optimized perinatal depression screening: Maximum impact, minimal burden

Eric Hurwitz[1], Caroline Shell[2], Kritika Chugh[3], Veerle Bergink[4], Rena C. Patel[5], Crystal Schiller[3], Melissa A. Haendel[1], on behalf of the National Clinical Cohort Collaborative (N3C)

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]Department of OBGYN, Sinai Hospital of Baltimore, Baltimore, MD, United States

[3]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[4]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[5]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

Abstract

Introduction: Perinatal depression affects up to 30% of pregnant and postpartum women, which has increased since the COVID-19 pandemic, making rapidly identifying affected women a high clinical priority. While screening tools like the Edinburgh Postnatal Depression Scale (EPDS) are widely used, brevity is important for busy clinical practice to reduce administration time and patient burden. Current methods to shorten assessments rely on traditional psychometric approaches, rather than machine learning (ML) methods that could optimize predictive accuracy.

Methods: We developed an ML framework using National Clinical Cohort Collaborative (N3C) data to predict full 10-item EPDS scores from shortened question subsets (n=22,924). We evaluated all 2-5 item combinations using linear regression, validating performance across multiple cohorts including postpartum women (n=7,750) and external pregnancy populations (n=1,217). For additional validation, we applied our approach to the PHQ-9 (n=398,606) to test generalizability. Binary

classification models using clinical thresholds (≥13) determined EPDS screening accuracy. Decision curve analysis was performed to assess the clinical utility of our ML method.

Results: The optimal 2-question EPDS combinations Q4+Q8 (anxiety/sadness) and Q5+Q8 (scared/sadness) both achieved R^2=0.70. Binary classification demonstrated strong performance (sensitivity=0.68-0.72, specificity=0.98-0.99). The framework generalized across postpartum subsets, external pregnancy cohorts, and PHQ-9 validation (R^2=0.64-0.73). Adding covariates did not improve performance. Decision curve analysis showed our ML approach had superior clinical benefit (0.01-0.03) versus traditional additive scoring.

Conclusion/Implications: Our ML framework suggests a reduced assessment burden with two EPDS questions maintains predictive accuracy as the full-item EPDS. With ~3.6 million annual U.S. births, this approach could identify additional positive perinatal depression screens, enhancing screening implementation across clinical settings.

Introduction

Standardized mental health assessments, such as the Edinburgh Postnatal Depression Scale (EPDS) and Patient Health Questionnaire (PHQ)-9, are well-established self-reported questionnaires for mental health screening, diagnosis, and assessing symptom severity[1,2]. However, prior work has shown that assessments with longer length have lower response rates and quality[3]. While shortening assessments can reduce patient and clinical burden, it is important to ensure the shortened assessment is still valid for diagnostic and treatment purposes[3].

Abbreviated versions of standardized mental health assessments are widely used, such as the PHQ-2, a 2-item subset of the PHQ-9 for assessing depressive symptom severity, and the GAD-2, a shortened form of the GAD-7 used to measure anxiety symptom severity[4,5]. To create shortened assessments, researchers typically choose between two main approaches: 1) administering a static short form, where all participants receive the same fixed subset of items from the full questionnaire, or 2) using a tailored test, in which different items are administered to different participants based on their responses[3]. When developing a static short form, researchers commonly rely on psychometric methods to select the most informative and representative items from the full assessment[3]. Psychometrics is a specialization in the field of psychology that uses statistical models to quantify psychological constructs and draw inferences from the data[6]. Examples of psychometrics include item response theory, computerized adaptive testing, and classical test theory[7–11]. Item response theory models unobservable traits (i.e., latent variables) and the probability of responding in particular ways to items based on mathematical functions accounting for item properties and person characteristics[7,8]. Computerized adaptive testing selects the most informative items for an individual based on their prior responses[9]. Classical test theory, developed before item response theory, views an observed test score as comprising a true score plus random error, focusing on overall test reliability and using straightforward statistics to identify and select the most psychometrically sound items[10].

The EPDS is the most commonly used assessment for screening perinatal mental health disorders, one of the most frequent complications during pregnancy and postpartum periods[12]. The EPDS is a 10-item self-administered questionnaire, with response range from 0-3 for each question and a total score ≥13 indicating a positive screen for perinatal depression, that lacks a widely-accepted abbreviated version[13]. Untreated perinatal mental health disorders contribute significantly to maternal morbidity and mortality, with suicide representing a leading cause of pregnancy-related death[14]. The scale of perinatal mental health challenges has been underscored since the COVID-19 pandemic, with studies reporting depression prevalence rates of 30% compared to pre-pandemic estimates of 10-20%[15]. The American College of Obstetricians and Gynecologists (ACOG) recommends screening for depression during pregnancy at the first obstetric visit, between 24-28 weeks gestation, and at the typical 6-week postpartum visit[16]. Recognizing the value of screening, the American Academy of Pediatrics (AAP) also recommends screening for perinatal depression and anxiety during the postpartum period at 1-, 2-, 4-, and 6-month well-child pediatrician visits due to pediatric touchpoints with clinical care[17]. Despite these clinical guidelines, perinatal mental health remains underrecognized and undertreated with screening rates reported at less than 20%[18]. This discrepancy highlights the significant utility of developing shortened, more accessible screening assessments.

Previous efforts to shorten the EPDS have used psychometric techniques described above as well as other analytic methods[19–24]. These include 1) Pearson correlation analysis and multiple linear regression to identify a 2-item subset of the EPDS (Q5:scared + Q8:sad), n=494)[19], 2) stepwise multiple regression analysis to create a 5-item subset (Q2:enjoyment + Q3:blame + Q4:anxious + Q7:sleep + Q8:sad, n=2,730)[20], 3) exploratory and confirmatory factor analysis to identify a 7-item subset (Q1:funny + Q2:enjoy + Q6:overwhelmed + Q7:sleep + Q8:sad + Q9:cry + Q10:harm, n=15,172)[21], 4) logistic and least absolute shrinkage and selection operator (LASSO) regressions to create a 3-item subset (Q2:enjoy + Q5:scared + Q7:sleep, n=305)[22], 5) a Rasch measurement model,

differential item functioning, and tests of local independence to identify an 8-question subset (Q1:funny + Q2:enjoy + Q3:blame + Q4:anxious + Q5:scared + Q6:overwhelmed + Q9:cry + Q10:harm, n=324)[23], 6) a priori question selection of a 7-item depressive subscale (Q1:funny + Q2:enjoy + Q6:overwhelmed + Q7:sleep + Q8:sad + Q9:cry + Q10:harm - the 7 depression-related EPDS questions), and 7) the 2 questions that most resemble the PHQ-2 (Q2:enjoy + Q8:sad) from the EPDS to create a 2-item assessment (n=106)[24]. Harel et al. built upon these previously developed shortened versions of the EPDS and applied item response theory and optimal test assembly to create the most informative shortened forms, ranging from 1 to 9 questions. Their analysis resulted in a 5-item shortened version of the EPDS (Q1:funny + Q2:enjoy + Q8:sad + Q9:cry + Q10:harm), which demonstrated a Cronbach's alpha of 0.82, sensitivity of 0.83 (95% CI: 0.73-0.89), and specificity of 0.86 (95% CI: 0.80-0.90) using a cutoff score of 4 (n=5,157)[25]. Although previous studies have created shortened versions of the EPDS, none have employed a machine learning (ML) approach. Furthermore, only two studies identified a 2-item subset[19,24], and prior efforts were limited by smaller sample sizes, which may have contributed to a lack of ML attempt. Additionally, existing studies examining shortened EPDS versions used data prior to the COVID-19 pandemic, which may not reflect current symptom patterns and presentations of perinatal mental health disorders and presents a gap in the literature[19–24]. Finally, most of these studies focused on a single assessment (e.g., the EPDS) rather than developing generalizable methods.

In this study, we developed a generalizable ML approach to shorten mental health assessments using data from the National Clinical Cohort Collaborative (N3C), a large electronic health record dataset. We applied this method to the EPDS and validated our approach through multiple analyses: a sensitivity analysis only in a postpartum cohort (from the combined pregnancy and postpartum cohorts), validation in an external pregnancy cohort, and application to a second mental health assessment (PHQ-9)[26]. Our ML framework predicts full assessment total scores from shortened question subsets, enabling efficient screening while maintaining predictive accuracy.

Beyond statistical validation across multiple cohorts, we evaluated the clinical utility of our ML approach compared with traditional additive scoring methods to assess real-world implementation value. This framework provides a scalable, data-driven methodology that can be readily applied to other standardized mental health assessments, offering significant potential to reduce screening burden while improving detection rates across multiple clinical and research settings.

Methods

Study population

The N3C Enclave consists of a limited dataset of COVID-19 cases and controls and includes their EHR data from a significant number of clinical sites. The harmonized data is incorporated in N3C from individuals who are COVID-19 positive and negative from January 1, 2020 to present with weekly updates to include near real-time procedures and clinical visits, with historical data dating back to January 1, 2018. Ingestion of data into N3C is matched with COVID-19 positive individual with two COVID-19 negative controls and matched on up to four sociodemographic variables (age, sex, race, and ethnicity) when available from the data partner site. Data in this study from the Enclave uses release up to December 16, 2024 (v187 with 85 data partner sites) using a Limited Data Set.

The National Institutes of Health (NIH) Institutional Review Board (IRB) approved N3C Enclave. Individual data partner sites maintain their own local IRB-approved data transfer or under a Johns Hopkins University Reliance Protocol (IRB00249128). Each investigator with access to the Enclave receives IRB approval from their institution. This project (RP-E39D65) was approved by the N3C Data Access Committee.

Edinburgh Postnatal Depression Scale (EPDS) and PHQ-9 phenotyping

Item-level responses to the EPDS and PHQ-9 in N3C were obtained using OMOP concepts listed in Table S1 and came from the Observation table. EPDS symptom classifications followed Harel et al., while PHQ-9 items were based on the Diagnostic and Statistical Manual of Mental Disorders (DSM)-5 criteria for major depressive disorder[25,27]. We then filtered for complete assessments where individuals answered all items on the same date and calculated total scores by summing all item-level responses. For the EPDS, we filtered on individuals who were female sex at birth. We did not perform any sex-based exclusion for the PHQ-9 since the goal was for methodological validation rather than creating a shortened PHQ-9 assessment. When multiple assessments existed for the same

individual, we filtered on one row per person ID selecting the highest total score, as the cohort was expected to contain a larger proportion of individuals with lower scores and wanted to optimize representation of individuals with higher scores.

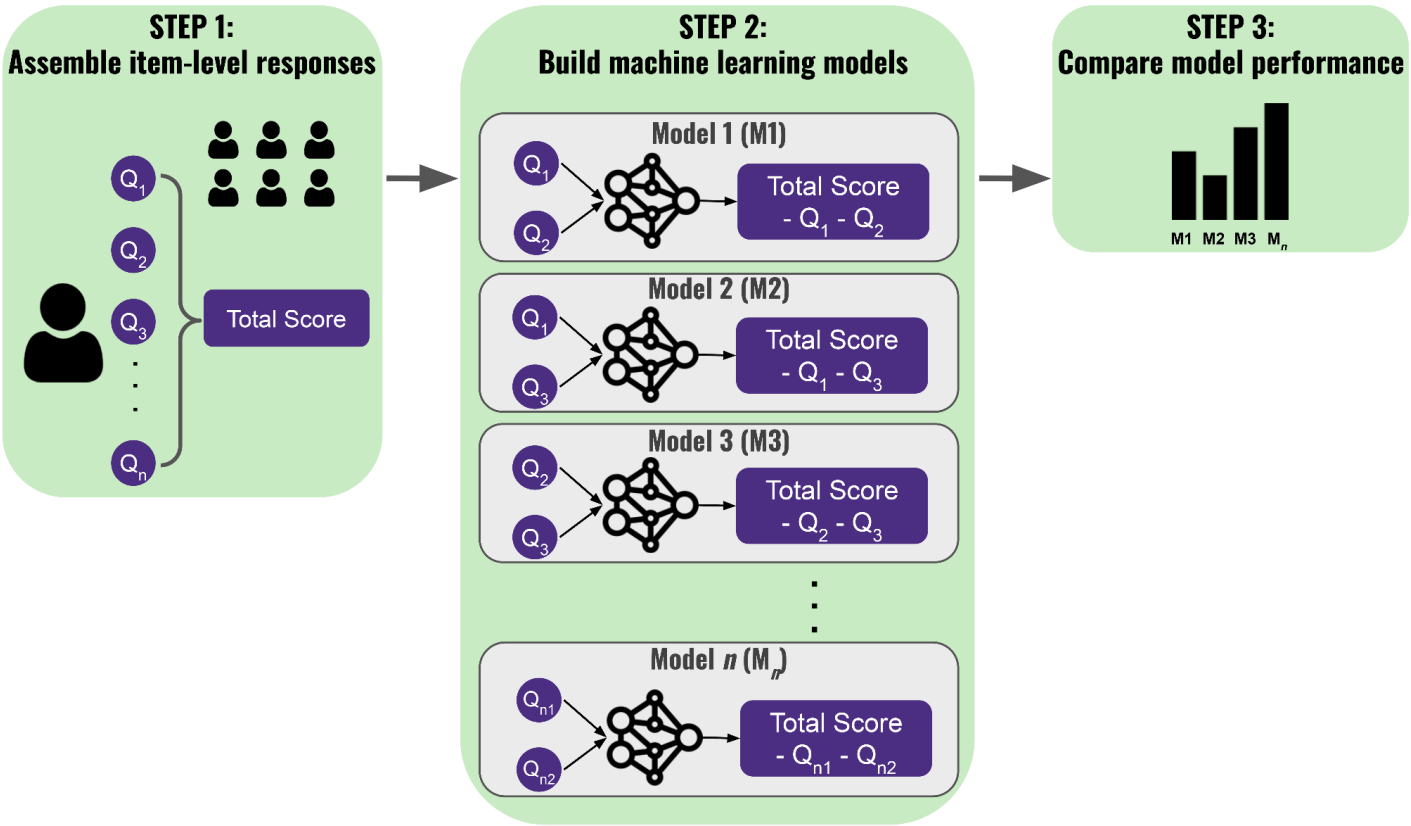
Calculating the average total score per question response

The mean and standard deviation of the total score for individuals who responded 0-3 for each of the individual questions was calculated to determine the relationship between responses to individual questions in the EPDS and PHQ-9 and the total score. To determine if the average score per individual question response was significantly different, we performed linear regression models assessing the relationship between the total score of individual questions and we repeated the analysis using response 0, response 1, and response 2 as the reference at a significance level of 0.05 with Bonferroni correction[28].

Developing a ML method for shortening psychometric assessments

We developed our ML approach to shorten psychometric assessments based on the hypothesis that an individual's responses to a subset of questions could reliably predict their total score. Therefore, each model was built using a framework that involved selecting a subset of questions, subtracting the responses to the selected subset of questions from the total score, and then using linear regression to predict the remaining portion of the score[29]. This process was repeated for every combination of two, three, four, and five questions from the psychometric assessments used in this study, which were the EPDS and PHQ-9 (Figure 1). Our primary focus was on two-question assessments, as commonly used screeners like the PHQ-2 are limited to two items to reduce respondent and clinician burden[4]. We then identified the top five combinations of two questions with the best predictive performance, assessed by R^2, root mean square error (RMSE), and mean absolute error (MAE)[30]. All ML analyses were performed using the Caret package in R[31].

Figure 1: A schematic displaying our ML method with 2-item subsets.



Our three-step machine learning (ML) framework identifies the most predictive question subsets within mental health assessments (e.g., the EPDS). First, we gather item-level responses to a mental health assessment and calculate the total score by summing the individual responses (Step 1). Next, we develop ML models using each permutation of two items as inputs to predict the remaining score (total score minus the two input items) (Step 2). Finally, we compare performance metrics across all models to identify which two-item combination most accurately predicts the remaining score, effectively determining the optimal brief assessment version (Step 3).

After developing our ML approach to predict the remaining EPDS total score as a continuous outcome, we extended the method to a binary classification model with the EPDS using logistic regression with clinical cutoffs to assess whether responses to a subset of questions could accurately predict a positive or negative screening result[32]. Binary classification models were constructed using top-performing question pairs identified from the continuous outcome model results or a priori based on clinical practice (e.g., using Q1:enjoy + Q2:sad from the PHQ-2)[4]. Model performance metrics included accuracy (95% confidence interval [CI]), kappa, sensitivity, specificity, positive predictive value, negative predictive value, prevalence, detection rate, detection prevalence, balanced accuracy, precision, recall, and F1 score[33].

<u>A sensitivity analysis for shortening the EPDS and predicting the total score among women in the postpartum period.</u>

We aimed to evaluate whether the top-performing question pairs identified from the full EPDS cohort, which included any female in N3C with a recorded EPDS, also performed well among women specifically in the postpartum period. To identify women in the postpartum period, we leveraged the Hierarchy and rule-based pregnancy episode Inference integrated with Pregnancy Progression Signatures (HIPPS) algorithm that has been developed and validated in N3C[34]. We refined our EPDS cohort to include women who completed the assessment between 4 weeks and 12 months postpartum, as assessments conducted before 4 weeks may be confounded by “postpartum blues” (“baby blues”), which affect up to 85% of new mothers, and our goal was to capture EPDS testing within the first year postpartum[35]. We then reapplied our continuous ML approach, as outlined above, to the postpartum cohort to identify the top five pairs of questions with the highest predictive performance for the remaining total score. Binary classification models were then built using select top-performing pairs of questions.

Validating our method in an external EPDS cohort of pregnant women

For an additional layer of validation and to assess the generalizability of our method, we assessed our ML approach in a cohort external from N3C. The cohort comprised 1,217 pregnant women who participated in a prospective study conducted by Washington University in St. Louis between January 2017 and January 2020, followed by secondary analysis by Stanford University, which we refer to hereafter as the Wash U cohort[36,37]. Briefly, Institutional Review Board approval for this study was obtained from Washington University School of Medicine Human Research Protection Office. The study enrolled English-speaking pregnant individuals aged ≥18 years with singleton pregnancies ≤20 weeks gestation who planned to deliver at Barnes-Jewish Hospital. Participants were seen longitudinally at study visits throughout each trimester, wore actigraphy devices continuously, and completed validated questionnaires about sleep habits and lifestyle. This cohort differs from N3C in its prospective design with active longitudinal follow-up and standardized data collection protocols, compared to N3C's retrospective electronic health record data from multiple healthcare systems. Additional details about the cohort can be seen in prior publications[37]. This cohort provided item-level responses to the EPDS in a different study setting and different demographic distribution compared to the N3C EPDS cohort. We then re-applied our continuous and binary classification ML models as described earlier. Demographic categorical variables between the N3C entire and Wash U cohorts were compared using Chi-square tests[38]. Two separate tests were performed: one comparing EPDS screening rates and one comparing race/ethnicity distributions between cohorts. Age was compared using an unpaired, two-sided t-test.

Covariates

Demographics of age and race/ethnicity were obtained using standardized N3C-wide definitions. Pregnancy outcomes were identified for each person using the HIPPS algorithm. Prior history of mental disorders were determined using condition diagnoses mapped to phecodes, a group of codes

for high-throughput EHR-based phenotyping based on International Classification of Diseases (ICD) codes to quickly identify diseases and conditions, such as mental health disorders[39,40], which are listed in Table S2.

Comparing addition and ML approach performance

Validated shortened assessments like the PHQ-2 use simple additive scoring, where responses to two questions are summed and compared against a threshold (typically ≥3 for positive screening)[41]. To evaluate our ML approach against this established method, we compared model performance using: 1) the optimal question combinations identified through our ML approach, and 2) the EPDS equivalent of the PHQ-2 questions (Q2:enjoy + Q8:sad). We evaluated four classification strategies: 1) simple addition with ≥3 threshold, 2) ML classification, 3) ensemble "OR" logic (positive if either ML or addition methods are positive), and 4) ensemble "AND" logic (positive if both ML and addition methods are positive). Performance was assessed using confusion matrices with thresholds of ≥3 for the additive approach and ≥13 for ML classification (standard EPDS cutoff)[13,42]. We compared F1 score, kappa, precision, and recall across all four approaches for each question combination. The F1 score, calculated as the harmonic mean of precision and recall, provides a single balanced metric that equally weighs both the ability to correctly identify true positive cases (recall/sensitivity) and minimize false positives (precision/positive predictive value)[33]. This makes it particularly useful for comparing overall screening effectiveness when there are trade-offs between these two competing priorities[33].

Decision curve analysis

We conducted decision curve analysis to evaluate the clinical utility of different EPDS screening strategies by comparing the net benefit of four prediction methods across threshold probabilities ranging from 0.01 to 0.99[43,44]. Analyses were performed separately for three two-question

combinations selected based on: 1) optimal predictive performance identified through our ML framework, and 2) clinical relevance as the EPDS equivalent to established brief screening tools (i.e., PHQ-2). Depression was defined as a binary outcome based on a total EPDS score ≥13, as this threshold identifies moderate to severe depression and minimizes false positive results[42]. For each question pair, we computed a simple sum score and generated binary indicators using a cutoff of ≥3[41]. The four prediction approaches included: 1) a simple addition method (sum of responses scaled to probabilities), 2) logistic regression models using 10-fold cross-validation, 3) an "OR" logic method (maximum probability from the addition and ML models), and 4) an "AND" logic method (minimum probability from both models). Net benefit was calculated using the standard formula: (TP/n)–(FP/n) x (p_t/(1–p_t)), where TP and FP represent true and false positives, n is the total sample size, and p_t is the threshold probability. Reference strategies included "treat all" (equal to the observed prevalence) and "treat none" (net benefit=0). Methods with net benefit exceeding both reference strategies across thresholds were considered to have potential clinical utility[43,44].

Use of generative AI

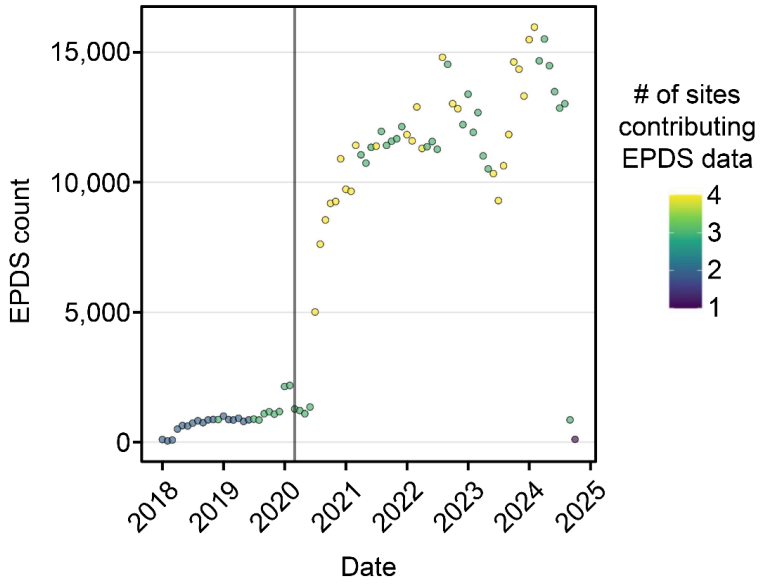
Chat-GPT, developed by Open-AI[45], and Claude, developed by Anthropic[46], were used to edit some portions of the manuscript, such as grammar, sentence structure, and synonyms. All content generated from these models were reviewed by authors and were not used for idea generation.

Results

EPDS testing over time and demographics of study participants

Within the N3C database, EPDS data were contributed by a range of one to four data partner sites between January 2018 and October 2024. The frequency of EPDS administration demonstrated a substantial increase over time, reaching a peak of >15,000 administrations in early 2024 across four contributing sites. Notably, a marked escalation in EPDS utilization was observed coinciding with the onset of the COVID-19 pandemic in March 2020 and an additional one to two data partner sites contributing EPDS data (Figure 2). This increase likely reflects the expansion of data partner sites with item-level EPDS reporting capabilities rather than solely changes in clinical practice, with the majority of available data collected in the post-pandemic period. After filtering for individuals with EPDS data who 1) had all item-level responses on the same date and 2) were female sex at birth, the N3C cohort of all women who took the EPDS included 22,924 women. Of those, 2,626 (11.5%) screened positive (total score ≥13). Among all women, the majority were White non-Hispanic (n=11,190 [48.8%]); however there was a larger percentage of Black non-Hispanic women in the positive screen group (n=708 [27.0%]) compared to those with a negative screen (n=4,443 [21.9%]). The median age of women who completed the EPDS was 32 years old (interquartile range [IQR] = 27-37) (Table 1).

Figure 2: EPDS data availability in N3C over time.



The total number of EPDS assessments in N3C by month from January 2018 to October 2024 in N3C. Data points demonstrate an upward trend over time and are colored according to the number of contributing data partner sites. The vertical line at March 2020 represents the beginning of the COVID-19 pandemic.

Table 1: Demographics of the N3C cohort with any EPDS data.

Group	Count (%) or Median (IQR)
Total cohort	**22,924**
EPDS screening: Positive	2,626 (11.5%)
Negative	20,298 (88.5%)
Race/ethnicity: White NH	11,190 (48.8%)
Black NH	5151 (22.5%)
Hispanic/Latinx of any race	4,226 (18.4%)
Other NH	1,318 (5.7%)
Unknown	1,039 (4.5%)
Age (years)	32 (27-37)

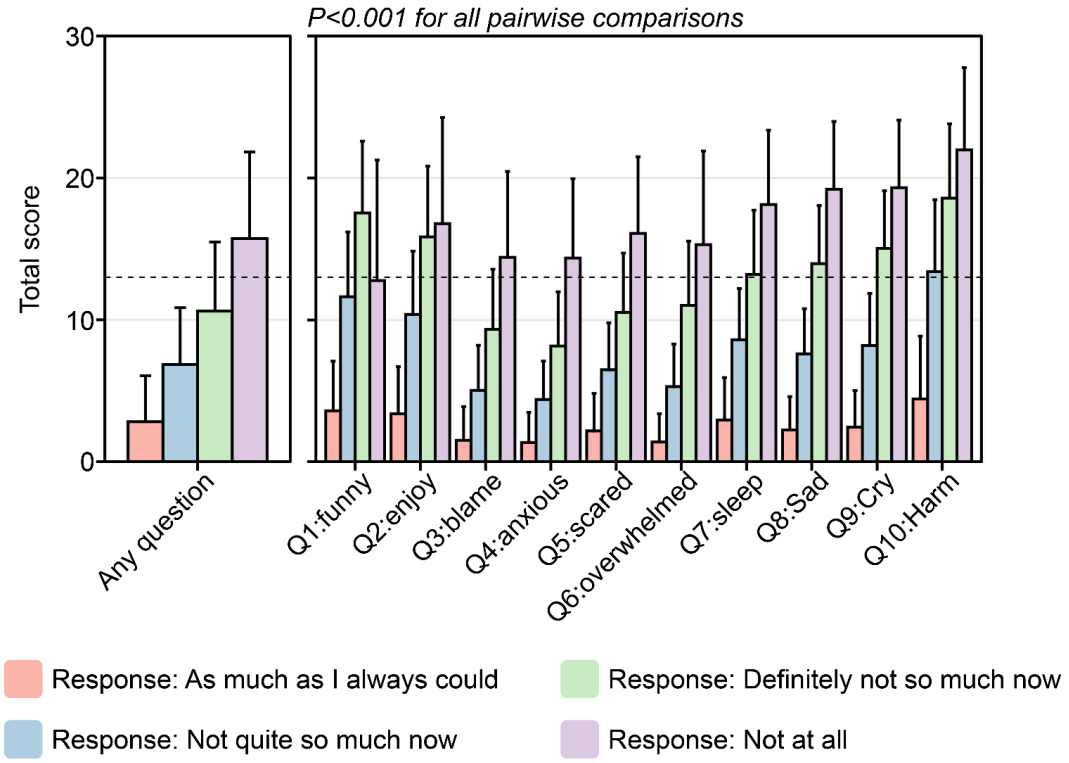
NH: non-Hispanic

We developed a generalizable ML method to shorten psychometric assessments applied to the EPDS.

We conceptualized the idea of shortening a psychometric assessment by exploring whether a person's total score could be predicted using responses to only a subset of questions rather than using a new cutoff score. We calculated the mean and standard deviation of the total EPDS score for each individual item response and used linear regression to determine whether total scores varied significantly across responses. We observed a consistent pattern where women selecting more severe response categories for individual items had correspondingly higher total EPDS scores, with all pairwise comparisons being statistically significant ($P<0.001$) (Figure 3). Interestingly, total EPDS scores based on responses to Q1 ("I have been able to laugh and see the funny side of things") mostly followed the same pattern, with the exception of those who responded "Nearly every day" (Figure 3). Building on these results, we hypothesized a subset of EPDS question responses could have predictive value for the total score. In other words, we developed a ML algorithm and evaluated model performance if the responses to a subset of questions could predict the total score, excluding responses to the questions included in the model. We built linear regression ML models using the response to each permutation of 2 EPDS questions to predict the remaining total score. The results demonstrated the 2 questions most predictive of the remaining EPDS total score were Q4 (I have been anxious or worried for no good reason) + Q8 (I have felt sad or miserable), followed by Q5 (I have felt scared or panicky for no very good reason) + Q8:sad, measured by R^2, RMSE, and MAE. Model results for Q4:anxious + Q8:sad were R^2=0.70, RMSE = 2.17, MAE = 1.61, and Q5:anxious+Q8 were R^2=0.70, RMSE=2.23, MAE=1.72 (Figure 4). Although the top five models demonstrated relatively similar performance, our analysis of all two-question combinations revealed considerable variability; for instance, the lowest-performing pair (Q1:funny + Q2:enjoy) yielded an R^2 of 0.37, RMSE of 3.66, and MAE of 2.92 (Table S3). We then assessed the performance of models when categorizing the EPDS total score as a positive or negative screen using a threshold of 13,

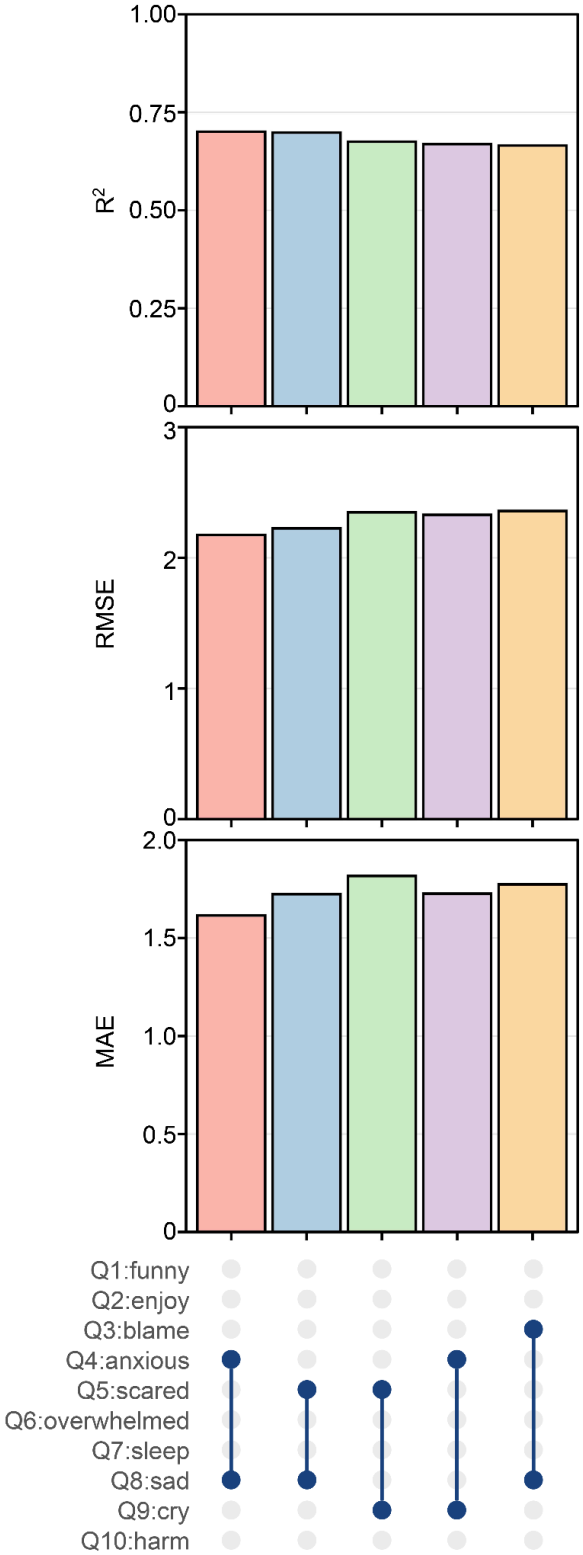
consistent with clinical literature[47], and ran a binomial classification model using Q4:anxious + Q8:sad in addition to Q5:scared + Q8:sad. Both models similarly demonstrated high performance (Table 2).

Figure 3: Increased individual question responses are associated with increased EPDS total scores.



The average EPDS total score based on responses to individual questions (Q1-Q10). The left panel shows overall scores by response to all questions, while the right panel breaks down scores by specific question. Four response categories are shown: "As much as I always could" (pink), "Not quite so much now" (blue), "Definitely not so much now" (green), and "Not at all" (purple). Data are expressed as mean and standard deviation. Statistical comparisons between each pair of response categories demonstrated significant differences in total EPDS scores across all comparisons (P<0.001). The horizontal dashed line represents a clinical cutoff score of 13 for positive depression screening.

Figure 4: Q4+Q8 and Q5+Q8 models displayed the highest performance for predicting the remaining EPDS total score.



Model performance metrics across the top five different models (represented by colored bars) using 2-item EPDS question subsets to predict the EPDS total score from the entire cohort of women who took the EPDS in N3C. The top panel displays R^2 values (coefficient of determination), the middle panel shows RMSE, while the bottom panel shows MAE. The dot-and-line diagram at the bottom indicates which variables (Q1-Q10) were included in each model configuration, with blue dots representing included variables connected by vertical lines.

Table 2: Binary classification models using Q4:anxious + Q8sad and Q5:scared + Q8:sad can predict EPDS screening results.

Metric	Q4:anxious + Q8:sad	Q5:scared + Q8:sad
Accuracy	0.95	0.95
95% confidence interval	(0.948, 0.953)	(0.949, 0.954)
Kappa	0.74	0.74
Sensitivity	0.72	0.68
Specificity	0.98	0.99
Positive predictive value	0.83	0.86
Negative predictive value	0.96	0.96
Prevalence	0.11	0.11
Detection rate	0.08	0.08
Detection prevalence	0.10	0.09
Balanced accuracy	0.85	0.83
Precision	0.83	0.86
Recall	0.72	0.68
F1 score	0.77	0.76

Our method shortening the EPDS remained stable for women in the postpartum period and an external pregnancy cohort

To validate our ML EPDS shortening approach, we tested our method in two additional cohorts: 1) a postpartum-specific subset within N3C (n=7,750) identified using the Hierarchy and rule-based pregnancy episode Inference integrated with Pregnancy Progression Signatures (HIPPS) algorithm and filtered to 4 weeks-12 months post-delivery[34], and 2) an external cohort of pregnant women from the Wash U cohort (n=1,217)[36,37]. The postpartum N3C cohort showed a slightly higher positive screening rate but otherwise similar demographics to the full N3C cohort (13.6% vs 11.5%). Although positive screening results require clinical confirmation for PPD diagnosis, our observed screening rates are consistent with published literature indicating PPD prevalence of approximately one in six postpartum women[48]. The Wash U cohort was demographically distinct compared to the full N3C cohort, with a higher percentage of Black NH participants (55.2% vs 22.5%, χ^2=589.92, degrees of freedom=2, P<0.001), younger median age (28 vs 32 years, P<0.001), and lower positive screening rate (7.6% vs 11.5%, χ^2=17.17, degrees of freedom=1, P<0.001) (Table 3). We observed a similar overall pattern of higher total scores corresponding to higher individual item responses in both the N3C Postpartum and Wash U cohorts, with a few exceptions that were not statistically significant (Figure S1). Despite demographic and clinical differences, both validation cohorts demonstrated the same fundamental patterns: individual question responses strongly correlated with total scores, and Q4:anxious + Q8:sad and Q5:scared + Q8:sad consistently emerged as the top-performing question pairs across all three cohorts (N3C full, N3C postpartum, Wash U external). While Q5:scared + Q8:sad performed slightly better in the postpartum cohort and Q4:anxious + Q8:sad in the full cohort, differences were minimal, confirming the robustness and generalizability of our ML approach across diverse populations and clinical contexts (Figure 5).

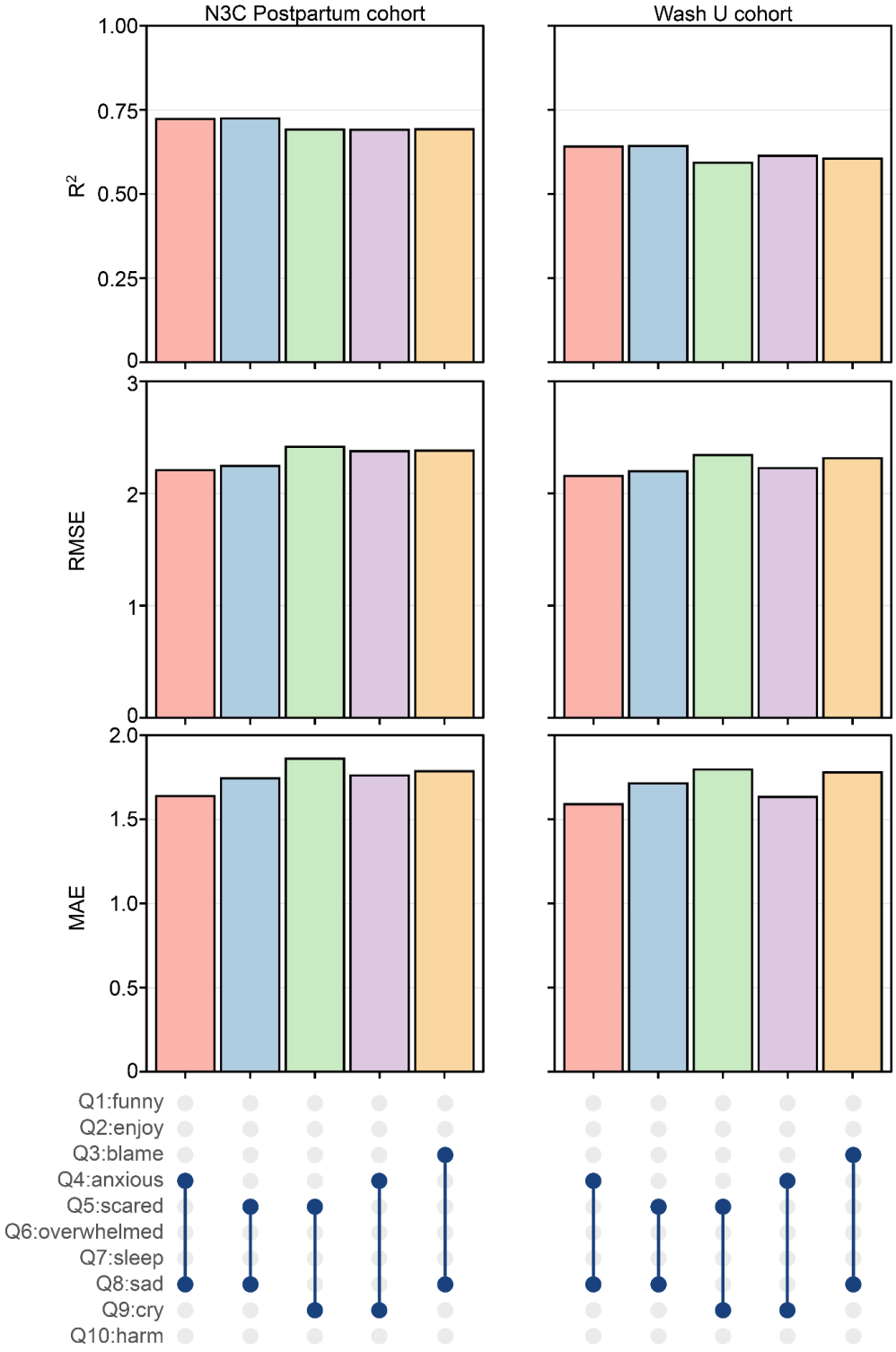
Table 3: Demographics of the postpartum cohort with EPDS data.

Group	N3C Entire cohort	N3C Postpartum cohort	Wash U cohort	Statistical comparison N3C Entire vs Wash U cohorts
Total cohort	**22,924**	**7,750**	**1,217**	
EPDS screening: Positive	2,626 (11.5%)	1,051 (13.6%)	92 (7.6%)	χ^2=17.17, df=1, P<0.001
Negative	20,298 (88.5%)	6,699 (86.4%)	1,125 (92.4%)	
Race/ethnicity: White NH	11,190 (48.8%)	3,322 (49.6%)	495 (40.7%)	χ^2=589.92, df=2, P<0.001
Black NH	5151 (22.5%)	1,387 (20.7%)	672 (55.2%)	
Hispanic/Latinx of any race	4,226 (18.4%)	1,172 (17.5%)	50 (4.1%)	
Other NH	1,318 (5.7%)	512 (7.6%)	0 (0%)	
Unknown	1,039 (4.5%)	306 (4.6%)	0 (0%)	
Age (years)	32 (27-37)	30 (26-34)	28 (24-32)	P<0.001

NH: non-Hispanic, df: degrees of freedom

Age is reported as median and interquartile range. Categorical variables between the N3C entire and Wash U cohorts were compared using Chi-square tests. Two separate tests were performed: one comparing EPDS screening rates and one comparing race/ethnicity distributions between cohorts. Age was compared using an unpaired, two-sided t-test.

Figure 5: Q4:anxious + Q8:sad and Q5:scared + Q8:sad models displayed the highest performance for predicting the remaining EPDS total score in a postpartum cohort.



Model performance metrics across the top five different models (represented by colored bars) using 2-item EPDS question subsets to predict the EPDS total score among women in the postpartum period in N3C (left) and Wash U cohort (right). The top panel displays R^2 values (coefficient of determination), the middle panel shows RMSE, while the bottom panel shows MAE. The dot-and-line diagram at the bottom indicates which variables (Q1-Q10) were included in each model configuration, with blue dots representing included variables connected by vertical lines.

Our ML method generalized to an assessment beyond the EPDS.

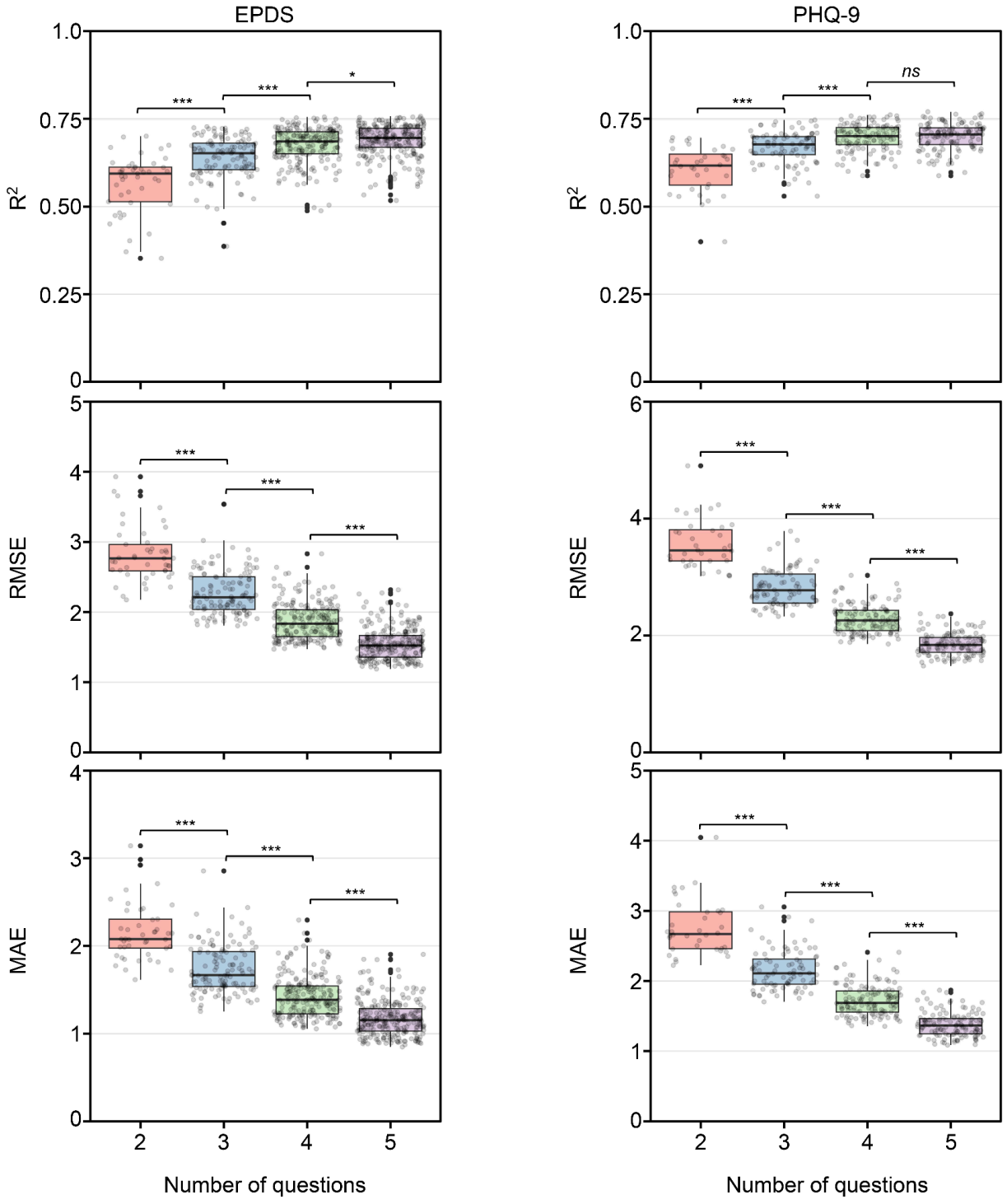
As methodological validation for our ML framework's generalizability beyond the EPDS, we applied it to the PHQ-9. We selected the PHQ-9 for validation because: 1) it provided a large sample size (n=398,606) for robust statistical testing, and 2) the existence of the established PHQ-2 allowed us to benchmark our ML-identified optimal questions against the current clinical standard[4]. This validation was designed to test whether our framework could consistently identify predictive question subsets across different psychometric instruments. The PHQ-9 cohort consisted of 398,606 distinct individuals with 151,106 (37.9%) individuals who screened positive (total score ≥10)[1,49,50]. The cohort was predominantly White Non-Hispanic (n=274,421, 68.8%) and the median age was 50 years old (IQR=31-68) (Table S4).

Similar to our approach with the EPDS, we first evaluated the average total PHQ-9 score based on individual question responses and observed the same pattern as with the EPDS; total scores increased as the responses to individual questions increased (all pairwise comparisons were statistically significant: $P<0.001$) (Figure S2). We then performed our ML method and identified the optimal 2 questions to predict the remaining total PHQ-9 score assessed by R^2, RMSE, and MAE. The results indicated that Q2:sad + Q7:concentration was the best-performing model based on R^2 and RMSE, and ranked as the 4th-best combination when assessed by MAE (Figure S3). Interestingly, our ML approach identified Q2:sad + Q7:concentration rather than the PHQ-2's Q1:enjoy + Q2:sad. When we compared Q2:sad + Q7:concentration against the established PHQ-2 using binary classification with a threshold of 10, both models demonstrated comparable performance, with slightly better results observed for our ML-identified combination, validating the robustness of our methodology (Table S6).

Model performance improved with more questions included.

Shortening psychometric assessments is valuable for improving efficiency, but it comes with a trade-off in predictive accuracy. Our goal was to find the optimal balance between reducing assessment burden (fewer questions) while maintaining the ability to accurately predict the full assessment score. To systematically evaluate this trade-off, we examined how predictive performance changed when using 2, 3, 4, or 5 questions to predict the total score that would be obtained from completing all items on the EPDS and in the PHQ-9 for validation and compared performance by R^2, RMSE, and MAE. As expected, there was a dose-dependent relationship where including additional questions improved model performance (higher R^2, lower RMSE and MAE), with the exception that R^2 improvement from 4 to 5 questions in the PHQ-9 was not statistically significant (Figure 6 and Table S7).

Figure 6: Adding additional questions improved model performance predicting the remaining total score of psychometric assessments.



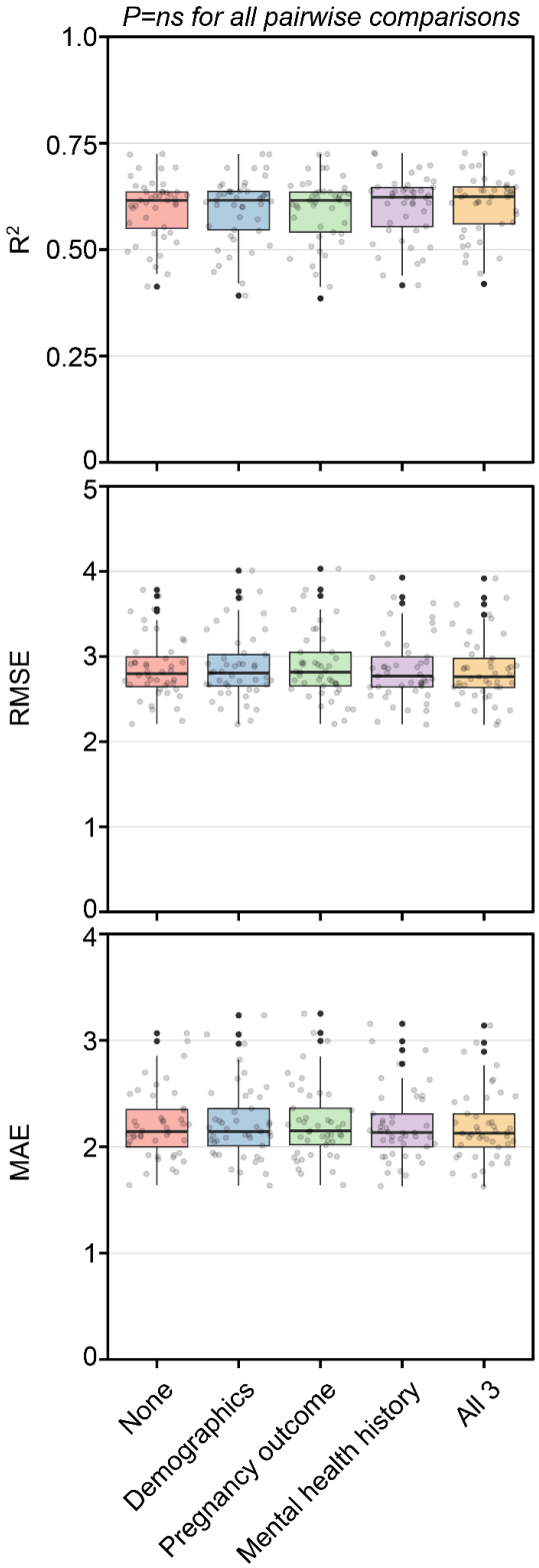
* = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$ after Bonferroni correction

Box-plots showing model performance metrics for predicting the total scores using shortened versions of the EPDS (left) and PHQ-9 (right) questionnaires (represented by colored bars) from women who took the EPDS and all individuals who took the PHQ-9 in N3C. Each gray dot represents an individual ML model testing a different combination of 2, 3, 4, or 5 questions to predict the total score for the EPDS and PHQ-9. The top panel displays R^2 values (coefficient of determination), the middle panel shows RMSE, while the bottom panel shows MAE. Overall, both EPDS and PHQ-9 assessments demonstrated improved model performance with additional questions, evidenced by higher R^2 values and lower RMSE and MAE.

Adding covariates did not improve EPDS model performance.

Covariates, such as demographics (i.e., age and race/ethnicity), pregnancy outcomes, and prior history of mental disorders, are known to increase risk for PPD[51]. To investigate whether these covariates impacted ML model performance for EPDS total score prediction using a subset of questions, we built ML models using each permutation of 2 EPDS questions in our N3C postpartum cohort adding covariates to models to assess their impact on model performance: 1) no covariates, 2) demographics (i.e., age and race/ethnicity), 3) pregnancy outcome (i.e., live birth, stillbirth, spontaneous abortion, induced abortion, ectopic pregnancy, delivery record only, and missing pregnancy outcome), 4) prior mental health history (i.e., any mental health diagnosis prior to the EPDS assessment date), and 5) all 3 (i.e., demographics, pregnancy outcome, and prior mental health history). The results revealed models that included covariates did not perform better than those without covariates *(P>0.05)* (Figure 7). For validation, we repeated this process using 3, 4, and 5 question subsets of the EPDS and observed similar findings *(P>0.05)* (Figure S4).

Figure 7: Adding covariates to ML models did not improve EPDS total score prediction performance.



ns = not significant

Box plots showing model performance metrics (R^2, RMSE, MAE) across all 2-question EPDS combinations in the N3C postpartum cohort. Models were evaluated with: no covariates (None), demographics only, pregnancy outcomes only, mental health history only, and all covariates combined (All 3). Each gray dot represents an individual 2-question model. Results demonstrate that predictive performance was statistically equivalent across all covariate combinations, with no significant pairwise differences observed (all $P>0.05$).

Clinical trade-offs between ML and other screening approaches.

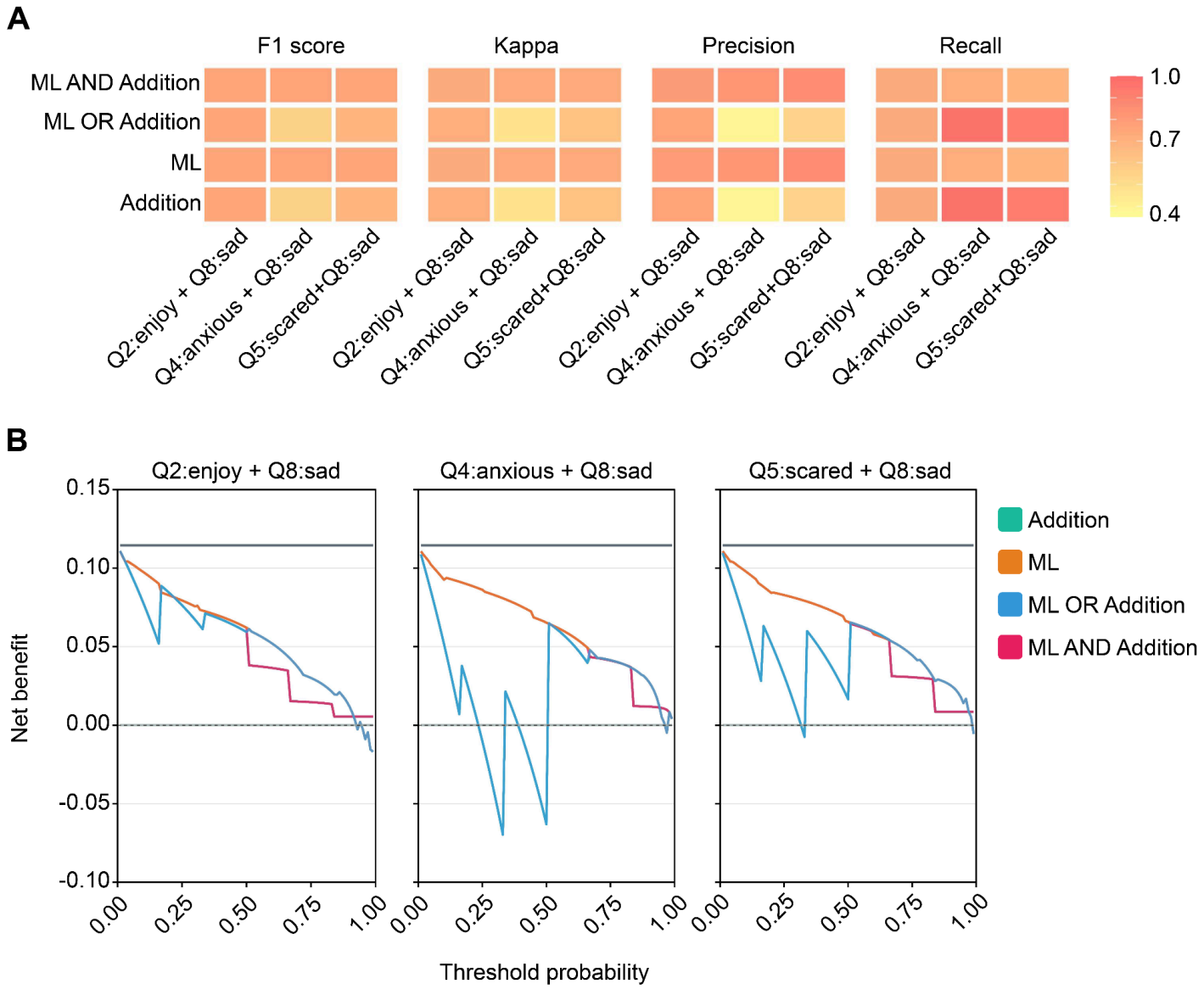
Given that the PHQ-2 exists as a shortened form of the PHQ-9 and comprises questions assessing anhedonia and sadness, we selected the equivalent two questions from the EPDS (Q2:enjoy + Q8:sad) and compared model performance with the question combinations previously identified through our ML approach (Q4:anxious + Q8:sad and Q5:scared + Q8:sad). In clinical screening with the PHQ-2, a threshold of ≥3 is used to balance sensitivity and specificity for positive depression screens[52]. Therefore, we directly compared our ML approach against simple additive scoring using this established threshold for the EPDS. Overall, we evaluated four distinct approaches: 1) simple addition (≥3 threshold), 2) ML classification, 3) ensemble "OR" logic (positive if either ML or addition methods are positive), and 4) ensemble "AND" logic (positive if both ML and addition methods are positive).

A key challenge in implementing screening is balancing the competing clinical priorities of identifying as many true cases as possible (sensitivity/recall) versus avoiding unnecessary follow-ups from false positives (precision). The results revealed distinct clinical trade-offs that could guide implementation decisions based on practice setting and resources. Traditional additive scoring acts like a "wide net" approach, catching 84-91% of depression cases but flagging many patients who don't actually have depression (precision=54-77%) (Figure 8A). From a pure screening perspective, this high sensitivity aligns with clinical principles of minimizing missed cases, particularly given the serious risks associated with untreated perinatal depression; however, the false positive rate may burden healthcare systems with unnecessary follow-up evaluations[53]. Our ML approach offers an alternative framework with higher precision (80-86%) but reduced sensitivity (recall=68-73%) (Figure 8A). The ensemble approaches amplified these trade-offs, with the OR logic maximizing case detection (recall=92-97%) at the cost of the lowest precision (precision=39-54%), while the AND logic achieved the same precision as the ML approach (precision=80-86%) with similar case detection rates (recall=68-73%) (Figure 8A). Rather than asserting one approach as universally superior, we

emphasize that the optimal strategy depends on implementation context and could potentially involve staged screening, similar to what is done in gestational diabetes testing[54].

We used decision curve analysis to assess clinical utility across different treatment thresholds, which evaluates whether using a screening test provides more benefit than harm compared to treating all patients or no patients[43,44]. The ML approach demonstrated consistently improved net benefit across all question combinations and most clinically relevant probability thresholds (0-80%), meaning it would result in better patient identification regardless of individual physician risk tolerance or institutional decision-making preferences. The simple addition method showed comparable but slightly lower performance, remaining clinically useful across a wide range of thresholds. Notably, the ensemble “OR” approach, despite its high sensitivity, frequently resulted in negative net benefit, particularly for Q4:anxious + Q8:sad, indicating potential harm compared to treating all patients. The ensemble “AND” approach showed limited clinical utility, with minimal net benefit across most thresholds (Figure 8B and Figure S5).

Figure 8: Our ML approach demonstrated superior clinical utility across question combinations for PPD screening.



A. *Performance comparison of EPDS question combinations using different screening methods. ML-based approaches achieved higher F1 scores, kappa values, and precision compared to the other methods.*

B. *Decision curve analysis evaluating clinical utility of different EPDS question combinations and screening approaches. Our ML method demonstrated superior net benefit over traditional approaches, with the greatest advantage observed at lower threshold probabilities. Decision curves for individual methods are also presented in Figure S5 for improved visibility.*

Discussion

In this work, we developed a generalizable ML method that enables accurate prediction of the full-length EPDS score from just two questions with improved clinical utility compared to traditional additive scoring approaches. Applied to a total of >23,000 individuals across multiple cohorts, our method consistently identified optimal question pairs in EPDS of Q4 (I have been anxious or worried for no good reason) + Q8 (I have felt sad or miserable) and Q5 (I have felt scared or panicky for no very good reason) + Q8:sad that maintained robust predictive performance while reducing respondent and clinician burden. Decision curve analysis revealed our ML approach provided consistently better clinical utility than simple addition methods across clinically relevant treatment thresholds, with meaningful improvements in patient outcomes. Validated retrospectively across demographically diverse populations including postpartum women, pregnant women, and general populations, this framework offers practical value for clinical perinatal mental health screening while providing a scalable methodology for optimizing any standardized mental health assessment that can be tested in prospective studies.

Shortened psychometric assessments are critical for reducing participant burden in research and clinical practice[3]. Methods to shorten assessments have primarily used psychometric approaches, such as item-response theory, computerized adaptive testing, and classical test theory[7–11]. ML approaches have also been used to shorten clinical assessments; however, they differ in that their primary goal is maximizing the accuracy of predicting the total score. Previous research has directly compared psychometric and ML approaches[3]. In this case, a genetic algorithm was used due to the impracticality of evaluating all possible question combinations, and a regression tree method was employed to identify the most effective subsets of items[3]. However, in our study, since the EPDS and PHQ-9 consist of only 10 and 9 questions, respectively, we were able to evaluate all possible question combinations. Additionally, we prioritized identifying two-question combinations from the outset to support practical implementation. Another study compared psychometric and ML

approaches for diagnostic purposes using Monte Carlo simulations to examine how classification accuracy is influenced by factors such as sample size, diagnosis prevalence, the correlation between diagnosis and test scores, and the structure of the assessment[11]. While screening tools cannot provide a diagnosis, our goal was to develop the shortest assessment with sufficient predictive power to identify those who warrant further evaluation for early diagnosis. Because individuals with higher assessment scores are more likely to receive a diagnosis, our method could be used to predict total scores to find those at greatest risk for the mental health disorder being assessed[41].

When comparing our study to prior work shortening the EPDS[19–24], our study includes the largest sample size, with nearly 8,000 more individuals than the next largest study and approximately 2,300 times the size of the smallest cohort among the listed studies. Interestingly, one of the top two performing pairs of questions with the best predictive performance was Q5:scared + Q8:sad, which was also identified by Choi et al[19]. Our other best-performing model included Q4:anxious + Q8:sad. In previous literature, Q4:anxious appeared in 2 out of 8 other shortened assessments, Q5:scared in 3 out of 8, and Q8:sad in 6 out of 8. Given that Q8 captures sadness, a core symptom of depression, its consistent presence across shortened EPDS versions is expected. The selection of Q4:anxious or Q5:scared alongside sadness reflects the EPDS's dual focus on both depressive and anxiety symptomatology. The two-question combination asking about anxiety and sadness mirrors clinical practice to ask about these two domains. The high prevalence of anxiety symptoms in women with perinatal depression supports why our ML method identified these anxiety-related questions as optimal predictors alongside the core depression symptom[55,56]. In terms of methodology, our study was the only one that used a ML approach to select the optimal two EPDS questions for shortening the assessment. While ML in psychometrics has gained attention in other fields, its use in mental health assessments has been limited[3,11,57].

During the investigation of our ML method to shorten the EPDS, we observed an increasing total score associated with individual question scores. However, an interesting and unexpected

finding was when examining this association at the individual question level, individuals who answered "Not at all" for Q1:funny did not have significantly higher total scores compared to those who responded "Definitely not so much now", which did not follow the pattern for the other nine EPDS questions. This question assesses the phenotype of anhedonia, and our findings may reflect the fact that there is likely little difference between individuals who responded "Not at all" and those who responded "Definitely not so much now," suggesting a minimal distinction in their ability to find humor[58]. Another unexpected finding was when applying our ML method to the PHQ-9, the combination of Q1:enjoy + Q2:sad (i.e., the PHQ-2[4]), did not appear in the top 5 most predictive models. However, when comparing the performance of the best model identified through our data-driven approach (Q2:sad + Q7:concentration) to the established PHQ-2, both models performed similarly overall, with the PHQ-2 showing slightly higher sensitivity and our Q2:sad + Q7:concentration pair exhibiting greater specificity. The balance between sensitivity and specificity is a common trade-off in ML, requiring optimization based on the specific goals and context of the use case[59]. Given the focus of this work is on perinatal mental health and the EPDS, we did not investigate this subset of questions from the PHQ-9, though it may be of interest for future research.

The most clinically relevant finding from our study demonstrated that our ML approach had the best net benefit based on decision curve analysis, which provides a framework for evaluating the clinical net benefit of screening tools by weighing the benefits of correctly identifying patients who need treatment against the harms of unnecessary interventions, offering a more clinically meaningful assessment than traditional ML performance metrics alone[43,44,60]. Evidence-based assessment tools must be practical for resource-limited settings, requiring instruments that are brief, validated, and easily implementable. Our approach addresses these implementation challenges by reducing assessment burden from 10 questions to just 2 while maintaining predictive accuracy and superior clinical utility. While the net benefit improvement appears modest (0.01-0.03), this translates to 1-3 additional positive screened patients identified per 100 screened which is highly impactful at

population scale, particularly for underdiagnosed conditions like perinatal depression. With approximately 3.6 million births annually in the United States[61], implementation of this ML approach would theoretically identify up to an additional ~100,000 positive perinatal depression screened women per year compared to simple additive scoring, representing a substantial public health impact where every additional case detected matters for maternal and infant outcomes. While the increased number of positive screens raises concerns given existing shortages in clinicians for diagnostic assessments and available therapists, the growing availability of digital health technologies, which have demonstrated improved perinatal depression outcomes, offers a promising complementary pathway for addressing this capacity gap[62–64]. Importantly, the consistent stability of our ML approach across different threshold probabilities provides clinicians with reliable performance regardless of specific cutoff selection, eliminating the unpredictable variability that characterizes traditional scoring methods and ensuring robust clinical implementation across multiple healthcare settings.

This study has several strengths, including the development of a generalizable ML framework for shortening psychometric assessments and predicting full-length scores from shortened versions, as well as being the first to apply ML methods to shorten the EPDS with the largest sample size. One limitation is that EPDS data were only contributed from a maximum of four data partner sites in N3C, which may limit generalizability. However, our cohort demographics demonstrate a mix of individuals with White non-Hispanic, Black non-Hispanic, Hispanic/Latinx, and other participants, suggesting some representation across racial and ethnic groups despite the limited number of contributing sites. Another limitation is that we only used linear and binomial logistic regression ML models, which were initially chosen for their simplicity and interpretability. Because these models demonstrated strong performance, we did not pursue more complex models like random forest or XGBoost[65,66]. Future research may explore additional ML models and hyperparameter tuning to optimize predictive power. Finally, for longer assessments, it may not be computationally feasible to generate a model and

evaluate performance for every combination of questions. Future studies could consider using a genetic algorithm alongside our ML approach to identify the best question subsets[3].

Our ML framework addresses a critical gap in mental health screening by providing a generalizable, data-driven methodology that substantially reduces assessment burden while maintaining clinical utility. Beyond the immediate application to perinatal depression screening, where our approach could identify up to 100,000 positive screens annually in the U.S. alone, this framework sets the stage for future prospective validation using real world data to ensure consistent performance stability across probability thresholds, ensuring reliable implementation regardless of clinical decision cutoffs. The ability to accurately predict full assessment scores from just two questions enables efficient longitudinal monitoring and reduces patient fatigue in research and clinical settings. Furthermore, the increased use of telehealth since the COVID-19 pandemic presents an opportunity to combine shortened EPDS assessments with our ML approach to improve remote screening capabilities[67]. For researchers, this methodology provides a systematic approach to optimize any standardized mental health instrument, potentially improving how we approach assessment development and validation. Most importantly, by making effective screening more accessible and less burdensome, this framework directly addresses the persistent gap between clinical guidelines recommending routine mental health screening and real-world implementation, offering a scalable solution that could improve mental health outcomes across diverse clinical settings and populations.

N3C Attribution

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS Contract No. 75N95023D00001 and the Center for Data to Health U24TR002306. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [https://doi.org/10.1093/jamia/ocaa196].

Disclaimer: This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the N3C program.

IRB: The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources.

Individual Acknowledgements For Core Contributors

We gratefully acknowledge the many core contributors to N3C as listed at covid.cd2h.org/core-contributors.

Data Partners with Released Data

The following institutions whose data is released or pending:

Available: Advocate Health Care Network — UL1TR002389: The Institute for Translational Medicine (ITM) • Aurora Health Care Inc — UL1TR002373: Wisconsin Network For Health Research • Boston

University Medical Campus — UL1TR001430: Boston University Clinical and Translational Science Institute • Brown University — U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Carilion Clinic — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Case Western Reserve University — UL1TR002548: The Clinical & Translational Science Collaborative of Cleveland (CTSC) • Charleston Area Medical Center — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Children's Hospital Colorado — UL1TR002535: Colorado Clinical and Translational Sciences Institute • Columbia University Irving Medical Center — UL1TR001873: Irving Institute for Clinical and Translational Research • Dartmouth College — None (Voluntary) Duke University — UL1TR002553: Duke Clinical and Translational Science Institute • George Washington Children's Research Institute — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • George Washington University — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Harvard Medical School — UL1TR002541: Harvard Catalyst • Indiana University School of Medicine — UL1TR002529: Indiana Clinical and Translational Science Institute • Johns Hopkins University — UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • Louisiana Public Health Institute — None (Voluntary) • Loyola Medicine — Loyola University Medical Center • Loyola University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Maine Medical Center — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Mary Hitchcock Memorial Hospital & Dartmouth Hitchcock Clinic — None (Voluntary) • Massachusetts General Brigham — UL1TR002541: Harvard Catalyst • Mayo Clinic Rochester — UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Medical University of South Carolina — UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • MITRE Corporation — None (Voluntary) • Montefiore Medical Center — UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Nemours — U54GM104941: Delaware CTR ACCEL Program • NorthShore University HealthSystem —

UL1TR002389: The Institute for Translational Medicine (ITM) • Northwestern University at Chicago — UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • OCHIN — INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • Oregon Health & Science University — UL1TR002369: Oregon Clinical and Translational Research Institute • Penn State Health Milton S. Hershey Medical Center — UL1TR002014: Penn State Clinical and Translational Science Institute • Rush University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Rutgers, The State University of New Jersey — UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Stony Brook University — U24TR002306 • The Alliance at the University of Puerto Rico, Medical Sciences Campus — U54GM133807: Hispanic Alliance for Clinical and Translational Research (The Alliance) • The Ohio State University — UL1TR002733: Center for Clinical and Translational Science • The State University of New York at Buffalo — UL1TR001412: Clinical and Translational Science Institute • The University of Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Iowa — UL1TR002537: Institute for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine — UL1TR002736: University of Miami Clinical and Translational Science Institute • The University of Michigan at Ann Arbor — UL1TR002240: Michigan Institute for Clinical and Health Research • The University of Texas Health Science Center at Houston — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • The University of Texas Medical Branch at Galveston — UL1TR001439: The Institute for Translational Sciences • The University of Utah — UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center — UL1TR002544: Tufts Clinical and Translational Science Institute • Tulane University — UL1TR003096: Center for Clinical and Translational Science • The Queens Medical Center — None (Voluntary) • University Medical Center New Orleans — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Alabama at Birmingham — UL1TR003096: Center for Clinical and Translational Science • University of Arkansas for Medical Sciences —

UL1TR003107: UAMS Translational Research Institute • University of Cincinnati — UL1TR001425: Center for Clinical and Translational Science and Training • University of Colorado Denver, Anschutz Medical Campus — UL1TR002535: Colorado Clinical and Translational Sciences Institute • University of Illinois at Chicago — UL1TR002003: UIC Center for Clinical and Translational Science • University of Kansas Medical Center — UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • University of Kentucky — UL1TR001998: UK Center for Clinical and Translational Science • University of Massachusetts Medical School Worcester — UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University Medical Center of Southern Nevada — None (voluntary) • University of Minnesota — UL1TR002494: Clinical and Translational Science Institute • University of Mississippi Medical Center — U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center — U54GM115458: Great Plains IDeA-Clinical & Translational Research • University of North Carolina at Chapel Hill — UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Oklahoma Health Sciences Center — U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • University of Pittsburgh — UL1TR001857: The Clinical and Translational Science Institute (CTSI) • University of Pennsylvania — UL1TR001878: Institute for Translational Medicine and Therapeutics • University of Rochester — UL1TR002001: UR Clinical & Translational Science Institute • University of Southern California — UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • University of Vermont — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • University of Virginia — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Washington — UL1TR002319: Institute of Translational Health Sciences • University of Wisconsin-Madison — UL1TR002373: UW Institute for Clinical and Translational Research • Vanderbilt University Medical Center — UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • Virginia Commonwealth University — UL1TR002649: C. Kenneth and

Dianne Wright Center for Clinical and Translational Research • Wake Forest University Health Sciences — UL1TR001420: Wake Forest Clinical and Translational Science Institute • Washington University in St. Louis — UL1TR002345: Institute of Clinical and Translational Sciences • Weill Medical College of Cornell University — UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • West Virginia University — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) Submitted: Icahn School of Medicine at Mount Sinai — UL1TR001433: ConduITS Institute for Translational Sciences • The University of Texas Health Science Center at Tyler — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • University of California, Davis — UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, Irvine — UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, Los Angeles — UL1TR001881: UCLA Clinical Translational Science Institute • University of California, San Diego — UL1TR001442: Altman Clinical and Translational Research Institute • University of California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute NYU Langone Health Clinical Science Core, Data Resource Core, and PASC Biorepository Core — OTA-21-015A: Post-Acute Sequelae of SARS-CoV-2 Infection Initiative (RECOVER) Pending: Arkansas Children's Hospital — UL1TR003107: UAMS Translational Research Institute • Baylor College of Medicine — None (Voluntary) • Children's Hospital of Philadelphia — UL1TR001878: Institute for Translational Medicine and Therapeutics • Cincinnati Children's Hospital Medical Center — UL1TR001425: Center for Clinical and Translational Science and Training • Emory University — UL1TR002378: Georgia Clinical and Translational Science Alliance • HonorHealth — None (Voluntary) • Loyola University Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • Medical College of Wisconsin — UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • MedStar Health Research Institute — None (Voluntary) • Georgetown University — UL1TR001409: The Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) •

MetroHealth — None (Voluntary) • Montana State University — U54GM115371: American Indian/Alaska Native CTR • NYU Langone Medical Center — UL1TR001445: Langone Health's Clinical and Translational Science Institute • Ochsner Medical Center — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • Regenstrief Institute — UL1TR002529: Indiana Clinical and Translational Science Institute • Sanford Research — None (Voluntary) • Stanford University — UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • The Rockefeller University — UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute — UL1TR002550: Scripps Research Translational Institute • University of Florida — UL1TR001427: UF Clinical and Translational Science Institute • University of New Mexico Health Sciences Center — UL1TR001449: University of New Mexico Clinical and Translational Science Center • University of Texas Health Science Center at San Antonio — UL1TR002645: Institute for Integration of Medicine and Science • Yale New Haven Hospital — UL1TR001863: Yale Center for Clinical Investigation

Data Availability

The National Clinical Cohort Collaborative (N3C) data transfer to the National Center for Advancing Translational Sciences is performed under a Johns Hopkins University Reliance Protocol (IRB00249128) or individual site agreements with the National Institutes of Health (NIH). The N3C Data Enclave is managed under the authority of the NIH; information can be found[68]. N3C Enclave data are protected and can be accessed for COVID-19–related research with an NIH-approved 1) IRB protocol and 2) institutional Data Use Request (DUR ID: RP-E39D65). A detailed accounting of data protections and access tiers is found at the NCATS website[68]. N3C Enclave and data access instructions can be found at the N3C website[69]; all codes used to produce the analyses in this

manuscript are available within the N3C Enclave to users with valid log-in credentials to support reproducibility.

Code availability

Code is available for use upon request to the corresponding author (EH).

Funding

Individual authors were supported by NIMH R01131542 (RC), NHGRI U24HG011449 (EH and MAH), and NHGRI RM1HG010860 (EH and MAH).

Author contributions

EH conceived the study and analyzed the data. EH, CS, KC, VB, CS, and MAH interpreted the results. MAH was responsible for resources. EH was responsible for the original manuscript draft preparation. EH was responsible for visualization. MAH was responsible for supervision and acquired funding. EH, CS, KC, VB, RCP, CS, and MAH were responsible for reviewing and editing the manuscript for publication.

References

1. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
2. Berwick, D. M. *et al.* Performance of a five-item mental health screening test. *Med. Care* **29**, 169–176 (1991).
3. Gonzalez, O. Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behav. Res.* **56**, 903–919 (2021).
4. Löwe, B., Kroenke, K. & Gräfe, K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J. Psychosom. Res.* **58**, 163–171 (2005).
5. Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O. & Löwe, B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Ann. Intern. Med.* **146**, 317–325 (2007).
6. Jones, L. V. & Thissen, D. 1 A history and overview of psychometrics. in *Handbook of Statistics* vol. 26 1–27 (Elsevier, 2006).
7. Embretson, S. E. & Reise, S. P. *Item Response Theory*. (Psychology Press, London, England, 2013). doi:10.4324/9781410605269.
8. The Theory and Practice of Item Response Theory. *Google Books* https://books.google.com/books/about/The_Theory_and_Practice_of_Item_Response.html?id=NchXAQAAQBAJ.
9. Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F. & Mislevy, R. J. *Computerized Adaptive Testing: A Primer*. (Routledge, 2000). doi:10.4324/9781410605931.
10. Hambleton, R. K. & Jones, R. W. Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice* **12**, 38–47 (1993).
11. Gonzalez, O. Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychol. Methods* **26**, 236–254 (2020).
12. Khadka, N. *et al.* Trends in postpartum depression by race, ethnicity, and prepregnancy body mass index. *JAMA Netw. Open* **7**, e2446486 (2024).
13. Cox, J. L., Holden, J. M. & Sagovsky, R. Detection of postnatal depression. Development of the 10-item

Edinburgh Postnatal Depression Scale. *Br. J. Psychiatry* **150**, 782–786 (1987).

14. Pearlstein, T., Howard, M., Salisbury, A. & Zlotnick, C. Postpartum depression. *Am. J. Obstet. Gynecol.* **200**, 357–364 (2009).
15. Caffieri, A. *et al.* Global prevalence of perinatal depression and anxiety during the COVID-19 pandemic: An umbrella review and meta-analytic synthesis. *Acta Obstet. Gynecol. Scand.* **103**, 210–224 (2024).
16. Implementing Perinatal Mental Health Screening. https://www.acog.org/programs/perinatal-mental-health/implementing-perinatal-mental-health-screening.
17. Earls, M. F., Yogman, M. W., Mattson, G., Rafferty, J. & COMMITTEE ON PSYCHOSOCIAL ASPECTS OF CHILD AND FAMILY HEALTH. Incorporating recognition and management of perinatal depression into pediatric practice. *Pediatrics* **143**, e20183259 (2019).
18. Britt, R. U.S. Maternal Depression Screening Rates Released for the First Time Through HEDIS. *Policy Center for Maternal Mental Health* https://policycentermmh.org/u-s-maternal-depression-screening-rates-released-for-the-first-time-through-hedis/ (2022).
19. Choi, S. K. *et al.* The simplified Edinburgh Postnatal Depression Scale (EPDS) for antenatal depression: is it a valid measure for pre-screening? *Int. J. Med. Sci.* **9**, 40–46 (2012).
20. Eberhard-Gran, M., Eskild, A., Samuelsen, S. O. & Tambs, K. A short matrix-version of the Edinburgh Depression Scale. *Acta Psychiatr. Scand.* **116**, 195–200 (2007).
21. Gollan, J. K. *et al.* Generating an efficient version of the Edinburgh Postnatal Depression Scale in an urban obstetrical population. *J. Affect. Disord.* **208**, 615–620 (2017).
22. Martínez, P., Magaña, I., Vöhringer, P. A., Guajardo, V. & Rojas, G. Development and validation of a three-item version of the Edinburgh Postnatal Depression Scale. *J. Clin. Psychol.* **76**, 2198–2211 (2020).
23. Pallant, J. F., Miller, R. L. & Tennant, A. Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BMC Psychiatry* **6**, 28 (2006).
24. Venkatesh, K. K., Zlotnick, C., Triche, E. W., Ware, C. & Phipps, M. G. Accuracy of brief screening tools for identifying postpartum depression among adolescent mothers. *Pediatrics* **133**, e45–53 (2014).
25. Harel, D. *et al.* Shortening the Edinburgh postnatal depression scale using optimal test assembly methods:

Development of the EPDS-Dep-5. *Acta Psychiatr. Scand.* **143**, 348–362 (2021).

26. Haendel, M. A. *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J. Am. Med. Inform. Assoc.* **28**, 427–443 (2021).
27. Abuse, S. Bookshelf. *Substance Abuse and Mental Health Services Administration (US)* https://www.ncbi.nlm.nih.gov/books/NBK519712/table/ch3.t5/ (2016).
28. Bland, J. M. & Altman, D. G. Multiple significance tests: the Bonferroni method. *BMJ* **310**, 170 (1995).
29. Maulud, D. & Abdulazeez, A. M. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends* **1**, 140–147 (2020).
30. Tatachar, A. V. Comparative assessment of regression models based on model evaluation metrics. *Impact Factor value:* **7**, 529 (2008).
31. Kuhn, M. Building Predictive Models inRUsing thecaretPackage. *J. Stat. Softw.* **28**, 1–26 (2008).
32. Applied Logistic Regression. *Google Books* https://books.google.com/books/about/Applied_Logistic_Regression.html?id=bRoxQBIZRd4C.
33. Erickson, B. J. & Kitamura, F. Magician's corner: 9. Performance metrics for machine learning models. *Radiol. Artif. Intell.* **3**, e200126 (2021).
34. Jones, S. E. *et al.* Who is pregnant? Defining real-world data-based pregnancy episodes in the National COVID Cohort Collaborative (N3C). *JAMIA Open* **6**, ooad067 (2023).
35. McKelvey, M. M. & Espelin, J. Postpartum depression: Beyond the 'baby blues'. *Nurs. Made Incred. Easy!* **16**, 28–35 (2018).
36. Ravindra, N. G. *et al.* Deep representation learning identifies associations between physical activity and sleep patterns during pregnancy and prematurity. *NPJ Digit. Med.* **6**, 171 (2023).
37. Stout, M. J. *et al.* A multidisciplinary Prematurity Research Cohort Study. *PLoS One* **17**, e0272155 (2022).
38. Ugoni, A. & Walker, B. F. The Chi square test: an introduction. *COMSIG Rev.* **4**, 61–64 (1995).
39. Bastarache, L. Using phecodes for research with the electronic health record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).
40. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, e0175508 (2017).

41. Manea, L., Gilbody, S. & McMillan, D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ* **184**, E191–6 (2012).

42. McCabe-Beane, J. E., Segre, L. S., Perkhounkova, Y., Stuart, S. & O'Hara, M. W. The identification of severity ranges for the Edinburgh Postnatal Depression Scale. *J. Reprod. Infant Psychol.* **34**, 293–303 (2016).

43. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).

44. Vickers, A. J., Cronin, A. M., Elkin, E. B. & Gonen, M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med. Inform. Decis. Mak.* **8**, 53 (2008).

45. ChatGPT. https://chatgpt.com/?utm_source=google&utm_medium=paidsearch_brand&utm_campaign=DEPT_SEM_Google_Brand_Acquisition_NAMER_US_Consumer_CPA_BAU_Mix&utm_term=chatgpt&gad_source=1&gad_campaignid=21714513245&gbraid=0AAAAA-IW-UVFk-hpJoMk5cXUoRwk50Tkb&gclid=Cj0KCQjw5ubABhDIARIsAHMighZT7_afTK4rSWfz34gPuX87_uWooKyn06qmNvMBGD8VeQZhbmDdZ0IaApi3EALw_wcB.

46. Claude. https://claude.ai/login?returnTo=%2F%3F.

47. Meltzer-Brody, S. *et al.* Elevated corticotropin releasing hormone (CRH) during pregnancy and risk of postpartum depression (PPD). *J. Clin. Endocrinol. Metab.* **96**, E40–7 (2011).

48. Shorey, S. *et al.* Prevalence and incidence of postpartum depression among healthy mothers: A systematic review and meta-analysis. *J. Psychiatr. Res.* **104**, 235–248 (2018).

49. Gilbody, S., Richards, D., Brealey, S. & Hewitt, C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J. Gen. Intern. Med.* **22**, 1596–1602 (2007).

50. Wittkampf, K. A., Naeije, L., Schene, A. H., Huyser, J. & van Weert, H. C. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen. Hosp. Psychiatry* **29**, 388–395 (2007).

51. Katon, W., Russo, J. & Gavin, A. Predictors of postpartum depression. *J. Womens. Health (Larchmt)* **23**, 753–759 (2014).

52. 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* **22**, 262–266 (2016).

53. Gavin, N. I. *et al.* Perinatal depression: a systematic review of prevalence and incidence. *Obstet. Gynecol.* **106**, 1071–1083 (2005).

54. Committee on Practice Bulletins-Obstetrics. ACOG practice bulletin no. 190: Gestational diabetes mellitus. *Obstet. Gynecol.* **131**, e49–e64 (2018).

55. Hurwitz, E. *et al.* Unlocking the potential of wearable device wear time to enhance postpartum depression screening and detection. *medRxiv* 2024.10.07.24315026 (2024) doi:10.1101/2024.10.07.24315026.

56. Thurgood, S., Avery, D. & Williamson, L. Postpartum Depression (PPD). (2010) doi:10.1007/978-0-387-79061-9_5723.

57. Orrù, G., Monaro, M., Conversano, C., Gemignani, A. & Sartori, G. Machine Learning in psychometrics and psychological research. *Front. Psychol.* **10**, 2970 (2019).

58. Putnam, K. T. *et al.* Clinical phenotypes of perinatal depression and time of symptom onset: analysis of data from an international consortium. *Lancet Psychiatry* **4**, 477–485 (2017).

59. Shreffler, J. & Huecker, M. R. Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

60. Zhao, L. *et al.* Understanding decision curve analysis in clinical prediction model research. *Postgrad. Med. J.* **100**, 512–515 (2024).

61. Births and Natality. https://www.cdc.gov/nchs/fastats/births.htm (2025).

62. Ko, J. Y. & Haight, S. C. Addressing perinatal mental health and opportunities for public health. *Am. J. Public Health* **110**, 765–767 (2020).

63. Lau, Y., Cheng, J.-Y., Wong, S.-H., Yen, K.-Y. & Cheng, L.-J. Effectiveness of digital psychotherapeutic intervention among perinatal women: A systematic review and meta-analysis of randomized controlled trials. *World J. Psychiatry* **11**, 133–152 (2021).

64. Arifin, S. R. M. *et al.* An evaluation of digital intervention for perinatal depression and anxiety: A systematic review. *AIMS Public Health* **11**, 499–525 (2024).

65. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

66. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.

67. Shaver, J. The state of telehealth before and after the COVID-19 pandemic. *Prim. Care* **49**, 517–530 (2022).

68. N3C Data Access Forms and Resources. https://ncats.nih.gov/research/research-activities/n3c/resources/data-access.

69. N3C - For Researchers. https://covid.cd2h.org/for-researchers/.