**Supplementary Materials for:**

*Development of a Probability Calculator for Psychosis Risk in*

*Children, Adolescents, and Young Adults*

**Description of Tests Composing the Penn Computerized Neurocognitive Battery (Penn CNB)**

The tests included in the CNB have been described in detail in Gur et al (2010, 2012). These broad domains were selected because they represent well-established brain systems, and each test in the battery was used in functional neuroimaging studies to probe aspects of these domains (Roalf et al., 2013). The tests are briefly described below by domain:

1. Executive:

The Penn Conditional Exclusion Test (PCET; Kurtz et al., 2004) measures the executive functions of abstraction and mental flexibility (ABF), critical for effective problem solving. It assesses the ability to derive principles and concepts from feedback, as well as the ability to detect and adjust to changing rules. The PCET uses the “odd man out” paradigm, in which participants must determine which object in a group does not belong. The exclusion rule can be based on the shape or configuration of the objects (e.g. a square would not fit in with three stars), the size of the objects, or the thickness of the lines outlining the objects. The participant is given feedback (“correct” or “incorrect”) after each response, and the test-administration program automatically changes the exclusion rule after ten consecutively correct responses (without informing the participant). The participant must then use the feedback to determine what the new exclusion rule is, and after ten consecutively correct responses, the rule is changed again. The test is scored based on demonstrated learning (proportion of correct responses multiplied by the number of learned rules; 1 is added to accommodate participants who were unable to discover any rule).

The Penn Continuous Performance Test (PCPT) measures vigilance and visual attention (ATT) independent of working memory or perceptual factors. Vertical and horizontal lines in 7-segment displays appear on the screen (at a rate of one second each), and the participant must press the spacebar when the lines are configured as complete numbers (first half of task) or complete letters (second half of task). Each half lasts 1.5 minutes, and during each one-second response window, the stimulus is presented for only 300 milliseconds (leaving 700 milliseconds of blank screen).

The Penn Letter N-Back Test measures working memory (WM), the ability to keep and refresh goal-related information. Participants attend to a continual series of letters that flash on the screen (one at a time) and press the spacebar according to three different rules (called the 0-back, 1-back, and 2-back). During the 0-back condition, the participant must simply respond to a currently present target (“X”). During the 1-back condition, he/she must press the spacebar when the letter on the screen is the same as the previous letter. During the 2-back condition, he/she must press the spacebar when the letter on the screen is the same as the letter before the previous letter (i.e. 2 letters back). In all trials, the inter-stimulus interval (ISI) is 2.5 seconds, and the stimuli (letters) themselves are presented for 0.5 seconds each. The participant practices all three principles before testing.

2. Episodic memory:

The Penn Word Memory Test (PWMT) measures episodic memory for verbal material (VMEM). In the first part of this test, participants are shown 20 words (for one second each) that they will be asked to identify later. During the recognition phase, participants are shown a series (one at a time) of 40 words, 20 of which are the stimuli they were asked to memorize, and the other 20 are distractors (matched for length, imageability and concreteness). For each word, the participant must decide whether he/she has seen the word in the memorization phase on a four-choice scale (“definitely not,” “probably not,” “probably yes,” or “definitely yes”).

The Penn Facial Memory Test (PFMT) measures episodic memory for faces (FMEM). The task is identical to the Penn Word Memory Test (above), except that the participant is asked to memorize faces instead of words. PFMT distractor faces are matched for age, ethnicity, and gender.

The Visual Object Learning Test (VOLT) measures episodic memory for shapes (SMEM). The task is nearly identical to the PWMT and PFMT (above), except that the participant is asked to memorize 10 Euclidean shapes instead of 20 words or faces.

3. Complex cognition:

The Penn Verbal Reasoning Test (PVRT) measures language-mediated complex cognition ability (LAN). The task involves a series of analogy problems patterned after Educational Testing Service factor-referenced test kit.

The Penn Matrix Reasoning Task (PMRT) measures nonverbal reasoning ability (NVR) using matrix reasoning problems as used in the Raven’s Progressive Matrices Test (Raven, 1989; Raven, 2000; Raven, Raven, & Court, 2000) and the Matrix Reasoning subscale of the WAIS-III (Raven, Raven, & Court, 2003).

The Penn Line Orientation Test (PLOT) measures the complex reasoning domain of spatial ability (SPA). The participant is shown two lines on the computer screen that differ in length and orientation, and must press a button to rotate one of the lines until its orientation (angle relative to a horizontal line) is the same as the other (non-rotating) line.

4. Social Cognition:

The Penn Emotion Identification Test (EMI) measures the social cognition domain of emotion identification - specifically, the ability to decode and correctly identify facial expressions of emotion. Participants are shown 40 faces (one at a time), and must determine whether the emotion expressed by the actor’s face is happiness, sadness, anger, fear, or none at all. There are 4 female 4 male faces for each emotion (4 x 2 x 5 = 40).

The Penn Emotion Differentiation Test (EMD) measures the social cognition domain of emotion intensity differentiation - the ability to decode the intensity of facial expressions of emotion. Participants are shown two faces at a time, both expressing the same emotion, and must determine which of the two faces expresses the emotion more intensely. Differential intensity was obtained by morphing a neutral face to one of four emotions (happy, sad, anger, fear).

The Penn Age Differentiation Test (AGD) measures the social cognition domain of the ability to decode the age of a face. Participants are shown two faces at a time, both neutral, and must determine which of the two faces is older. The stimuli were constructed from young faces morphed into old faces, providing graded levels of difficulty.

**Expanded Summary of Variable-Selection Techniques.**

1. Lasso regression. Lasso regression is a type of regularized regression that assesses a “penalty” (forced downward bias of coefficients) for both the number of predictors used in the model and the collinearity among them (Tibshirani, 1996). Usually, the penalty causes most coefficients to become exactly zero, retaining a confined set of non-redundant predictors for prediction (i.e. features with non-zero coefficients are “good”). Here, we randomly split our sample in half, ran lasso predicting psychosis spectrum status at final time point, and saved the variables selected as optimal. This was repeated 10,000 times, and variables were ranked for importance based on what percentage of the 10,000 random splits they were selected.

2. Random Forest importance. The random forest algorithm (Liaw & Wiener, 2002) leaves the realm of conventional linear modeling and incorporates decision trees. The first step of these decision trees is to determine which single variable best predicts PS in the training sample. Once that is determined, the algorithm splits the sample into those who are high and those who are low on the “important” variable. In these split sub-samples, the algorithm then looks for the most important variable (omitting the first variable because the sample has just been split by that variable). Those sub-samples are then further split based on their “most important” variables, which can be different between the two sub-samples. The above splitting continues until some stopping criterion is met – e.g. if all the persons in one sub-sample have the same value for a variable (preventing it from being split further). The above is repeated for 500 trees, where each new tree is created using a bootstrapped training sample (re-sampled with replacement for each tree). Variable importance is determined by permuting the values of each variable in the trees one-at-a-time to determine how much the prediction suffers. For example, if a variable is randomly scrambled across participants (permuted) and that doesn’t have an effect on the prediction accuracy, then that variable could not have been very important (prediction is fine without it). Conversely, if the predictions become inaccurate after scrambling the variable, then that variable must have been important. As for #1, the above was repeated 10,000 times, and the variable importance metrics were saved.

3. Relieff and STIR. Relieff is an algorithm designed specifically for feature selection and known for being especially sensitive to interactions among features (Le et al., 2019). Given n cases and p variables, Relieff first chooses a random case (e.g. n100). In p-dimensional Euclidean space, the algorithm finds the nearest neighbor that is the same as the random case on the DV (a “hit”) and the nearest neighbor that is different from the random case on the DV (a “miss”). For any given variable, if the value of that variable for the randomly drawn case is closer to the “hit” case than to the “miss” case, the variable importance goes up; otherwise, it goes down. Consider the example of trying to classify vehicles as “bus” or “not bus”, where the only variables you have are the vehicle’s weight and model year. If the random case you choose happens to be a bus, chances are the nearest neighbor that is a bus (a “hit”) will have more similar weight to your random case than will the nearest not-bus (a “miss”). This would demonstrate the importance of knowing vehicle weight in the classification. Conversely, you would not expect the same phenomenon for model year—i.e. there is no reason to think the nearest “hit” will have a model year near or far from the random case. The stir process was repeated 10,000 times, and variables were selected only if their stir-produced p-value was significant. Note that here we used a variation of relieff called “MultiSURF”, but the principle is the same.

**Public Interface for Psychosis Risk Calculator**

The risk calculator developed here is available to the public at the following web address: <https://www.mooremetrics.com/psychosis-risk-calculator/>

References

Gur, R. C., Richard, J., Hughett, P., Calkins, M. E., Macy, L., Bilker, W. B.,... Gur, R. E. (2010). A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. *Journal of Neuroscience Methods, 187*, 254 –262.

Gur, R. C., Richard, J., Calkins, M. E., Chiavacci, R., Hansen, J. A., Bilker, W. B., Loughead, J., Connolly, J. J., Qiu, H., Mentch, F. D., Abou-Sleiman, P. M., Hakonarson, H., & Gur, R. E. (2012). Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. *Neuropsychology, 26*, 251-265.

Kurtz, M. M., Ragland, J. D., Moberg, P. J., & Gur, R. C. (2004). The Penn Conditional Exclusion Test: A new measure of executive-function with alternate forms for repeat administration. *Archives of Clinical Neuropsychology*, *19*(2), 191-201.

Le, T. T., Urbanowicz, R. J., Moore, J. H., & McKinney, B. A. (2019). Statistical inference Relief (STIR) feature selection. *Bioinformatics, 35*(8), 1358-1365.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 21*(1), 1-16.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual Section 1. General Overview (1998 Edition, Updated 2003)*. San Antonio, TX: Pearson Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000). Standard progressive matrices. *Psychology Press.*

Roalf, D. R., Ruparel, K., Verma, R., Elliott, M. A., Gur, R. E., & Gur, R. C. (2013). White matter organization and neurocognitive performance variability in schizophrenia. *Schizophrenia Research, 143*, 172-178.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

|  |  |
| --- | --- |
| Table S1. Variables Used in NAPLS-2 and PNC-based Psychosis Risk Calculators | |
| NAPLS-2 | PNC Equivalent |
| Age | Age |
| Sum of SOPS items P1 and P2 | Total score on PS-R |
| BACS symbol coding raw score | Overall CNB Speed |
| Hopkins Verbal Learning Test | Penn Word Memory Test |
| Stressful life events | n/a |
| Family history of psychosis | Family history of psychosis |
| GF:Social (decline in functioning) | CGAS (overall) total score |
| Traumatic events (> 1) | Traumatic events |
| Note. CNB = computerized neurocognitive battery; GF = Global Functioning: Social; CGAS = Children’s Global Assessment Scale; BACS = Brief Assessment of Cognition in Schizophrenia; \*participant sex was not included in the original Cannon et al. calculator but was added here because it is collected by all participants at no additional cost. | |

|  |  |  |
| --- | --- | --- |
| Table S2. Cox Proportional Hazards Model Predicting Conversion to Psychosis in PNC Sample | | |
| Predictor | Odds Ratio | p-value |
| Age (years) | 1.32 | < 0.0005 |
| Family History | 1.20 | 0.788 |
| PWMT† | 1.03 | 0.870 |
| Trauma | 0.91 | 0.466 |
| CGASǂ | 0.99 | 0.279 |
| Overall Speed† | 0.88 | 0.510 |
| PS-R Totalǂ | 1.04 | 0.012 |
| Note. †Units are standard deviation; ǂunits are points—e.g. a 1-point increase in CGAS score (out of 100) decreases the odds of transition by 1% (1.00 – 0.99 = 0.01 = 1%); PWMT = Penn Word Memory Test; CGAS = Children’s Global Assessment Scale; PS-R = Prevention through Risk Identification, Management, and Education Screen - Revised. | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table S3. Prediction Performance Diagnostics for PNC Psychosis Risk Calculator, with 95% CIs. | | | | | | | |
| Max Vars in Model | AUC | Balanced Accuracy [(PPV + NPV)/2] | Sensitivity | Specificity | Positive Predictive Value (PPV) | Negative Predictive Value (NPV) | Threshold |
| 1 | 0.656 (0.634 - 0.678) | 0.644 (0.624 - 0.664) | 0.640 (0.584 - 0.695) | 0.672 (0.607 - 0.736) | 0.778 (0.753 - 0.803) | 0.510 (0.483 - 0.536) | 0.656 (0.634 - 0.678) |
| 2 | 0.676 (0.666 - 0.687) | 0.669 (0.658 - 0.680) | 0.725 (0.704 - 0.746) | 0.628 (0.616 - 0.640) | 0.778 (0.772 - 0.784) | 0.560 (0.543 - 0.577) | 0.676 (0.666 - 0.687) |
| 3 | 0.655 (0.641 - 0.669) | 0.654 (0.637 - 0.671) | 0.747 (0.710 - 0.785) | 0.563 (0.531 - 0.594) | 0.754 (0.745 - 0.764) | 0.554 (0.525 - 0.583) | 0.655 (0.641 - 0.669) |
| 4 | 0.680 (0.667 - 0.694) | 0.671 (0.657 - 0.684) | 0.713 (0.686 - 0.740) | 0.648 (0.627 - 0.668) | 0.784 (0.775 - 0.794) | 0.557 (0.536 - 0.578) | 0.680 (0.667 - 0.694) |
| 5 | 0.675 (0.660 - 0.690) | 0.667 (0.653 - 0.682) | 0.721 (0.689 - 0.753) | 0.630 (0.595 - 0.665) | 0.778 (0.765 - 0.791) | 0.557 (0.534 - 0.580) | 0.675 (0.660 - 0.690) |
| 6 | 0.697 (0.686 - 0.708) | 0.687 (0.676 - 0.697) | 0.730 (0.713 - 0.748) | 0.663 (0.641 - 0.685) | 0.796 (0.786 - 0.805) | 0.578 (0.563 - 0.593) | 0.697 (0.686 - 0.708) |
| 7 | 0.693 (0.680 - 0.706) | 0.680 (0.667 - 0.693) | 0.695 (0.670 - 0.720) | 0.692 (0.665 - 0.718) | 0.802 (0.790 - 0.814) | 0.558 (0.540 - 0.576) | 0.693 (0.680 - 0.706) |
| 8 | 0.693 (0.680 - 0.706) | 0.680 (0.668 - 0.692) | 0.696 (0.676 - 0.716) | 0.690 (0.661 - 0.718) | 0.801 (0.788 - 0.814) | 0.558 (0.544 - 0.573) | 0.693 (0.680 - 0.706) |
| 9 | 0.687 (0.672 - 0.702) | 0.676 (0.661 - 0.691) | 0.711 (0.676 - 0.747) | 0.662 (0.628 - 0.696) | 0.791 (0.778 - 0.804) | 0.561 (0.537 - 0.586) | 0.687 (0.672 - 0.702) |
| 10 | 0.678 (0.661 - 0.695) | 0.666 (0.649 - 0.683) | 0.685 (0.637 - 0.734) | 0.670 (0.621 - 0.720) | 0.789 (0.771 - 0.807) | 0.543 (0.516 - 0.570) | 0.678 (0.661 - 0.695) |

A close up of text on a white background

Description automatically generated

Supplementary Figure S1. ROC Curve for within-sample Prediction, Compared to within-sample Prediction of Randomly Permuted Cases.