A machine learning model using in-game data for predicting unhealthy substance use among adolescents.

Kammarauche Aneni, MBBS MHS[1], Ching-Hua Chen PhD RN[2], Gaoqianxue Liu BS[1], Saatvik Kher[3], and Lynn Fiellin MD[1]
[1] Yale University, [2] IBM, [3]Pomona College

Background. Despite the substance use epidemic in the United States, less than 10% of adolescents needing treatment receive it. Adolescents with unhealthy substance use are at elevated risk of developing a substance use disorder. Thus, early identification of adolescents with unhealthy substance use can lead to timely interventions. However, barriers to screening for unhealthy substance use preclude treatment access. These barriers include lack of time for screening, provider discomfort, and lack of privacy in healthcare settings. Using digital tools such as games for screening that can be employed without needing personnel or trained providers may mitigate these barriers. Games collect considerable data during play that may be used for prediction. Performance in a game is captured by in-game metrics such as reaction time, speed of task completion, or choice move in a game. These in-game metrics may reflect digital biomarkers of health outcomes. Digital biomarkers are "physiological and/or behavioral measures generated by persons that may explain, influence, or predict health outcomes." Performance in a game is also influenced by cognitive processes such as working memory, inhibitory control, and decision making. These cognitive processes are implicated in the development of unhealthy substance use and, in turn, are impacted by unhealthy substance use. As such, in-game metrics may represent digital biomarkers of cognitive processes that predict unhealthy substance use. However, it is not yet known if in-game metrics predict unhealthy substance use among adolescents. To utilize in-game metrics in screening for unhealthy substance use among adolescents, it is critical to investigate whether these metrics reflect underlying physiological processes associated with unhealthy substance use. This study aimed to develop a predictive machine learning model for unhealthy substance use among adolescents using in-game data from an existing videogame. The first part of this study (reported here) developed a machine learning model for unhealthy substance use among adolescents. A subsequent study will investigate whether metrics identified in this study correlate with cognitive processes implicated in the development of unhealthy substance use.

Methods. We used data previously collected from a randomized controlled trial (RCT) of a videogame, *PlayForward*, an evidence-based videogame developed by the play2PREVENT Lab for HIV prevention and reduction of high-risk behaviors such as unhealthy substance use. During gameplay, players earn points for making good decisions around risky behaviors such as engaging in unhealthy substance use. In the *PlayForward* RCT, 333 adolescents aged 11-14 years were enrolled, 166 of whom played the videogame. At baseline, self-report data on unhealthy substance use (outcome 1) and self-efficacy to refuse substances (outcome 2) were collected. The game software collected data related to adolescent performance in the game and stored these data as log files. Through a review of the literature and the reference files developed by the game developers for the log data, we identified in-game metrics that may be predictive of unhealthy substance use and extracted these as features for training machine learning models. ***Feature selection:*** To select features used to train our models, we normalized all features and then removed those that were either highly correlated with other features (i.e., correlation > 0.95) or which had low variance (i.e., variance < 0.01). Classification models for our outcomes were constructed using six binary classification models: SVM, logistic regression, Gradient boosting, Neural Networks, Decision Trees, and Random Forest. ***Analysis:*** The analytic sample was split into a 70/30 train/test set. Using 5-fold cross-validation, we computed the Area Under the Receiver Operating Characteristics (AUROC) curve values, sensitivity, and specificity to determine which models best fit our outcomes.

Results. We dropped six participant log files due to file errors resulting in a sample of 160 adolescents with 33% reporting unhealthy substance use and 49% reporting poor self-efficacy to refuse substances. We extracted 285 in-game performance metrics from in-game log files. We dropped 28 in-game metrics highly correlated with other metrics and 18 in-game metrics with low variance, resulting in 239 in-game metrics for 160 participants. Our neural network model best predicted unhealthy substance use with the following mean (SD): AUC, 0.61 (0.04); sensitivity, 0.49 (0.09); and specificity, 0.55 (0.13). The logistic regression model performed best in predicting adolescents with poor self-efficacy to refuse substances with the following mean (SD) scores: AUC, 0.63 (0.09), sensitivity, 0.61 (0.06); specificity, 0.55 (0.15). The in-game performance metrics that contributed most to model performance were metrics captured during the beginning levels of the game.

Conclusion. The rising substance use crises call for innovative methods for early identification and intervention delivery. Our preliminary findings suggest that in-game performance metrics may predict an adolescent's substance use and efficacy to refuse substances with significant but weak correlations. Limitations were that the machine learning methods used were non-temporal and required hand-crafted features. Alternative methods that can take advantage of patterns in time-series data could help overcome these limitations. Despite these limitations and although further validation is needed, machine learning models may identify adolescents with risky substance use behaviors using game-based tools and, therefore, identify adolescents who will benefit from timely interventions. Given the ability for games to be delivered without the need for trained providers and across multiple sites, such as primary care or school-based settings, these game-based tools hold promise for early identification and monitoring of risky substance use behaviors among adolescents.