Machine learning-based prediction of substance use in adolescents: Derivation and validation in multinational datasets in South Korea, USA, and Norway

Hojae Lee, Hyejun Kim, Seokjun Kim, Ahmed Hammoodi, Yujin Choi, Hyeon Jin Kim, Jiseung Kang, Lee Smith, Guillaume Fond, Laurent Boyer, Sung Wook Baik, Hayeon Lee, Jaeyu Park, Rosie Kwon, Selin Woo, Dong Keon Yon

Submitted to: Journal of Medical Internet Research
on: May 31, 2024

Table of Contents

Machine learning-based prediction of substance use in adolescents: Derivation and validation in multinational datasets in South Korea, USA, and Norway

Hojae Lee[1]; Hyejun Kim[1]; Seokjun Kim[1]; Ahmed Hammoodi[2]; Yujin Choi[1]; Hyeon Jin Kim[1]; Jiseung Kang[3]; Lee Smith[4]; Guillaume Fond[5]; Laurent Boyer[5]; Sung Wook Baik[6]; Hayeon Lee[1]; Jaeyu Park[1]; Rosie Kwon[1]; Selin Woo[1]; Dong Keon Yon[7]

[1]Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center Seoul KR
[2]Department of Business Administration, Kyung Hee University School of Management Seoul KR
[3]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital Boston US
[4]Centre for Health, Performance and Wellbeing, Anglia Ruskin University Cambridge GB
[5]Research Centre on Health Services and Quality of Life, Assistance Publique-Hopitaux de Marseille, Aix Marseille University Marseille FR
[6]Department of Software, Sejong University College of Electronics and Information Engineering Seoul KR
[7]Department of Pediatrics, Kyung Hee University College of Medicine Seoul KR

Corresponding Author:
Dong Keon Yon
Department of Pediatrics, Kyung Hee University College of Medicine
23 Kyungheedae-ro, Dongdaemun-gu
Seoul
KR

Abstract

Background: We aimed to address gaps in global understanding of cultural and social variations by employing a high-performance machine learning model to predict adolescent substance use across three national datasets.

Objective: This study aims to develop a predictive model for adolescent substance use using multinational datasets and machine learning(ML).

Methods: The study utilized the Korea Youth Risk Behavior Web-Based Survey(KYRBS) from South Korea(n=1,145,178) to train ML models. For external validation, we employed the Youth Risk Behavior Survey(YRBS) from the USA(n=1,690,108) and Norwegian nationwide Ungdata surveys(Ungdata) from Norway(n=793,879). After developing diverse tree-based models, we further evaluated feature importance.

Results: The study utilized nationwide adolescent datasets for ML model development and validation, analyzing data from 1,145,178 KYRBS adolescents, 54,709 YRBS Asian subset participants, and 720,812 from Ungdata. The random forest model was the top performer on the KYRBS, achieving an AUROC of 80.8%(95% CI, 80.7-80.8) with sensitivity of 72.9%(72.8-73.0), specificity of 72.9%(72.9-73.0), accuracy of 72.9%(72.8-73.0), and balanced accuracy of 72.9%(72.8-73.0). The model's AUROC scores were 73.2% for YRBS and 75.7% for Ungdata in external validation. The top features for predicting substance use were smoking status, body mass index (BMI), and alcoholic consumption.

Conclusions: With multinational datasets from South Korea, USA, and Norway, the findings of this study underscore the potential efficacy of ML models in predicting adolescent substance use with smoking status, body mass index, and alcoholic consumption identified as key predictors. The random forest model exhibited notable performance in this prediction. These findings could be a basis for future studies exploring more comprehensive factors influencing adolescent substance use or developing intervention strategies based on these predictors.

(JMIR Preprints 31/05/2024:62805)
DOI: https://doi.org/10.2196/preprints.62805

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Article

Machine learning-based prediction of substance use in adolescents: Derivation and validation in multinational datasets in South Korea, USA, and Norway

Running title: **Machine learning and substance use**

Hojae Lee,[1,2‖] Hyejun Kim,[1,3‖] Seokjun Kim,[1,4‖] Ahmed Hammoodi,[5] Yujin Choi,[1,6] Hyeon Jin Kim,[1,2] Jiseung Kang,[7,8] Lee Smith,[9] Guillaume Fond,[10] Laurent Boyer,[10] Sung Wook Baik,[11] Hayeon Lee[1], Jaeyu Park,[1,2] Rosie Kwon,[1,2*] Selin Woo,[1*] Dong Keon Yon[1,2,12*]

1. Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea

2. Department of Regulatory Science, Kyung Hee University, Seoul, South Korea

3. Department of Applied Information Engineering, Yonsei University, Seoul, South Korea

4. Department of Medicine, Kyung Hee University College of Medicine, Seoul, South Korea

5. Department of Business Administration, Kyung Hee University School of Management, Seoul, South Korea

6. Department of Korean Medicine, Kyung Hee University College of Korean Medicine, Seoul, South Korea

7. Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA.

8. Department of Anesthesia, Harvard Medical School, Boston, MA, USA.

9. Centre for Health, Performance and Wellbeing, Anglia Ruskin University, Cambridge, UK

10. Research Centre on Health Services and Quality of Life, Assistance Publique-Hopitaux de Marseille, Aix Marseille University, Marseille, France

11. Department of Software, Sejong University College of Electronics and Information Engineering,

Seoul, South Korea

12. Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea

‖ These authors contributed equally.

***Corresponding authors**

Rosie Kwon, PhD

Department of Pediatrics, Kyung Hee University College of Medicine, 23 Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, South Korea.

Email: rosiekwon514@gmail.com

Selin Woo, PhD

Department of Pediatrics, Kyung Hee University College of Medicine, 23 Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, South Korea.

Email: dntpfls@naver.com

Dong Keon Yon, MD, FACAAI, FAAAAI

Department of Pediatrics, Kyung Hee University College of Medicine, 23 Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, South Korea.

Tel: +82-2-6935-2476

Fax: +82-504-478-0201

Email: yonkkang@gmail.com

Words Count: 3039

ABSTRACT

Background: We aimed to address gaps in global understanding of cultural and social variations by employing a high-performance machine learning model to predict adolescent substance use across three national datasets.

Objectives: This study aims to develop a predictive model for adolescent substance use using multinational datasets and machine learning(ML).

Methods: The study utilized the Korea Youth Risk Behavior Web-Based Survey(KYRBS) from South Korea(n=1,145,178) to train ML models. For external validation, we employed the Youth Risk Behavior Survey(YRBS) from the USA(n=1,690,108) and Norwegian nationwide Ungdata surveys(Ungdata) from Norway(n=793,879). After developing diverse tree-based models, we further evaluated feature importance.

Results: The study utilized nationwide adolescent datasets for ML model development and validation, analyzing data from 1,145,178 KYRBS adolescents, 54,709 YRBS Asian subset participants, and 720,812 from Ungdata. The random forest model was the top performer on the KYRBS, achieving an AUROC of 80.8%(95% CI, 80.7-80.8) with sensitivity of 72.9%(72.8-73.0), specificity of 72.9%(72.9-73.0), accuracy of 72.9%(72.8-73.0), and balanced accuracy of 72.9%(72.8-73.0). The model's AUROC scores were 73.2% for YRBS and 75.7% for Ungdata in external validation. The top features for predicting substance use were smoking status, body mass index (BMI), and alcoholic consumption.

Conclusions: With multinational datasets from South Korea, USA, and Norway, the findings of this study underscore the potential efficacy of ML models in predicting adolescent substance use with smoking status, body mass index, and alcoholic consumption identified as key predictors. The random forest model exhibited notable performance in this prediction. These findings could be a basis for future studies exploring more comprehensive factors influencing adolescent substance use or developing intervention strategies based on these predictors.

Keywords: adolescents; machine learning; substance; prediction; random forest

Introduction

Substance use among adolescents remains a global concern, with a myriad of health challenges.[1, 2] When initiated at an early age, these behaviors can escalate to more serious health disorders.[3] As globalization increases and cultural integration continues, substance use patterns varies widely, making paramount to understand these patterns across a diverse cultural landscape.[4] Conventional statistical methods have long been employed to examine the predictors and outcomes of adolescent substance use.[5, 6] With recent advancements, machine learning (ML) is emerging as a promising tool to provide a fresh perspective on this intricate matter.[7]

Existing studies provide insights into the epidemiology and sociocultural correlates of substance use among adolescents in distinct landscapes.[5, 6] However, there remains a paucity of studies employing ML techniques to predict substance use across multinational datasets, which could offer more granular, accurate, and potentially actionable insights.[8]

In this study, our primary focus was on developing a ML-based prediction model for adolescent substance use. We began the model development process using comprehensive datasets from South Korea and extended our validation process by utilizing datasets from the United States and Norway.[9] This validation process, followed by model refinement, ensured applicability of our model beyond a specific region but achieved accuracy in predicting substance use across distinct cultural and national contexts. By integrating these global datasets, we developed a predictive model that reflects our collaborative international research. Our novel approach equips stakeholders with a sophisticated tool informed by global data. This helps address and preempt adolescent substance use effectively across different national contexts.

Methods

Study design and participants

This study was primarily designed to develop a ML model for substance use prediction among adolescents utilizing three distinct nationwide datasets: KYRBS from South Korea[10], YRBS from the United States[11], and Ungdata from Norway[12]. KYRBS was initially used to train the ML model, followed by the external validation process utilizing the YRBS and Ungdata.

The discovery dataset, KYRBS, constitutes 1,145,178 participants[13] who primarily represents Korean adolescents - a demographic largely comprised of East Asians. Meanwhile, YRBS, which began with 1,690,108 participants, was narrowed down to 54,709 after exclusions due to missing and inconsistent data. Since the YRBS encompasses a diverse racial spectrum of American adolescents, we strategically extracted the Asian subset of YRBS for subsequent validation to observe the gradual difference in culture. Similarly, from the initial count of 793,879 in the Ungdata, only 720,812 participants were selected after data processing. During the data processing phase, certain variables in the extra validation cohort from the YRBS and Ungdata were found to be absent. To manage this issue, values for these missing variables were imputed using the median from the discovery dataset, KYRBS.

In order to evaluate the generalizability of our model across diverse cultural groups, we employed a phased validation approach. Initially, our model was validated using an Asian subset of the YRBS, which shares some cultural similarities with our discovery dataset. This allowed us to assess the model performance in a context with moderate cultural divergence. Subsequently, we expanded our validation to include a dataset from Norway, representing a significantly different cultural background. This phased approach, starting with a dataset that shares some cultural overlap and progressively moving to a completely different cultural context, provided a rigorous test of the model's robustness and versatility across varying cultural backgrounds.[14]

Adolescents aged between 13 and 18 years who completed their respective surveys were included. Our primary outcome, substance use, was derived from the question "Have you ever consumed illicit substances at least once in your lifetime". We distinguished smoking and alcohol from other substances in our analysis, recognizing their unique consumption patterns and sociocultural implications, and health effects.[15] Substances other than smoking and alcohol are distinguished primarily due to concerns regarding their potential for misuse and health risks.[16] This decision was made to ensure that our model captures nuances specific to each substance, thereby enhancing the specificity and relevance of our predictions. Integral covariates under consideration spanned across factors including age, sex, region, BMI, academic achievement, household income, smoking status, alcoholic consumption, stress status, sadness and despair, suicidal thinking, and suicide attempts.[13, 17]

Model development and validation

Utilizing the KYRBS, we developed a predictive model to extrapolate the behavioral patterns of Korean adolescents regarding substance use. Due to rigorous substance regulations and law enforcement measures in South Korea, accessibility and consumption of substances are notably limited.[18] Consequently, the number of instances representing substance usage within the KYRBS was sparse (n=14,548; 1.27%). Given the intricate nature of the data, we employed tree-based models, specifically focusing on the random forest algorithm due to its powerful predictive capabilities. It was further validated by assessing the AUROC score compared to other tested models in evaluating the risk of substance use among adolescents.

The ML model trained with the KYRBS was proceeded with external validation using the YRBS and Ungdata. We utilized the Asian subset of YRBS for our initial external validation. After achieving optimal model performance, we further validated our model using Ungdata. This external validation process facilitated the calibration and evaluation of our model, ensuring consistent

performance across heterogeneous adolescent cohorts.

To further strengthen the validity of our results, hyperparameter tuning was pivotal in our methodology. To ensure resilience and optimizing the performance of our random forest model, we performed hyperparameter tuning using GridSearchCV, focusing on maximizing the AUROC score to determine the best combination of hyperparameters. Various metrics such as AUROC score, accuracy, sensitivity, specificity, and balanced accuracy were utilized to evaluate the performance of the model.[13, 19]

Performance assessment

The tools and techniques we employed for assessment were consistent. Our evaluation metrics comprised AUROC, accuracy, sensitivity, specificity, and balanced accuracy. To provide a visual representation of the model efficacy, we utilized visualization techniques, notably the receiver operating characteristic (ROC) curve.[20]

Software and libraries

All computations, model training, validation, and evaluation process were executed using Python 3.11.4 (Python Software Foundation, Wilmington, DW, USA). Key libraries from our toolbox included Scikit-learn 1.2.2, NumPy 1.24.0, and Pandas 2.1.0 for ML tasks and data wrangling. Visualization was facilitated using Matplotlib 3.7.2 and Seaborn 0.12.2.

Ethical statement

The study protocol was approved by the Institutional Review Board of the Korean Disease Control and Prevention Agency (KDCA), U.S. Centers for Disease Control and Prevention (CDC), Norwegian Social Research institute (NOVA)[21], and Kyung Hee University (KHUH 2022–06–042), and all participants provided written informed consent. This research followed the guidelines

outlined in the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[22]

Results

Demographic characteristics

This research utilized a detailed exploration using nationwide adolescent datasets from South Korea, aiming to design and validate a ML model to predict substance usage tendencies among adolescents. The primary demographic consisted of adolescents aged between 13 and 18 years (Figure 1 and Figure 2).

We collected data from the Korea Youth Risk Behavior Web-based Survey (KYRBS) , the Youth Risk Behavior Survey (YRBS), and the Norwegian nationwide Ungdata surveys (Ungdata), and subsequently standardized the covariates for the ML predictive modeling process. Within the primary cohort from the KYRBS to develop the prediction model, the sex distribution was as follows: male (48.42%) and female (51.58%). For the initial external validation cohorts of an external validation process, the YRBS (Asian subset) features a sex distribution of: male (49.02%) and female (50.98%). In the second validation step, the Ungdata has the following sex distribution: male (49.74%) and female (50.26%) (Table 1).

It is important to highlight that both the primary and the extra-validated cohorts included adolescents from a diverse range of socioeconomic backgrounds, such as household income. Moreover, risk behavioral habits, such as alcoholic consumption and different types of smoking status, were taken into account. The comprehensive scope of these cohorts ensures robust representativeness, which is crucial for the development and subsequent evaluation of the predictive ML model being examined.[13]

Given these demographic details, our ML-based predictive model provides detailed insights into the potential risks associated with substance consumption among adolescents. This study not only seeks statistical significance but also provides a deeper understanding of adolescent behavior, offering valuable insights for potential interventions.

Machine learning model results

As illustrated in Figure 3, extensive model evaluations revealed that the random forest was the optimal model for predicting substance use among adolescents. The primary model, sourced from the KYRBS and assessed with a 95% CI, disclosed that the random forest model notched an area under the receiver operating characteristic curve (AUROC) score of 80.75% (95% CI, 80.69-80.82) with detailed analysis on a sensitivity of 72.91% (95% CI, 72.84-72.98), specificity of 72.91% (95% CI, 72.85-72.98), accuracy of 72.91% (95% CI, 72.84-72.98), and balanced accuracy of 72.91% (95% CI, 72.84-72.98). Other models exhibited the following scores: CatBoost at 80.56% (95% CI, 80.49-80.63), AdaBoost at 80.54% (95% CI, 80.47-80.61), LightGBM at 80.48% (95% CI, 80.41-80.55), and XGBoost at 80.12% (95% CI, 80.05-80.19).

For the initial external validation, the independent YRBS Asian dataset was utilized. The random forest model displayed an AUROC score of 73.20%, followed by a sensitivity of 67.89%, specificity of 68.26%, accuracy of 68.20%, and balanced accuracy of 68.08%. In the subsequent external validation, the random forest model yielded an AUROC score of 75.69%, coupled with a sensitivity of 72.38%, specificity of 72.69%, accuracy of 72.68%, and balanced accuracy of 72.54%.

Across all evaluations, both internal and external, the random forest model consistently exhibited a predominant performance, particularly in terms of the AUROC score, cementing its adoption for the study objective.

Feature importance

Figure 4 illustrates the importance of various features as determined by the random forest model in predicting substance use among adolescents. Specifically, smoking status was identified as the most significant predictor, accounting for 27.8% importance. This was closely followed by body mass index (BMI), with 14.7% and alcoholic consumption at 12.5%. Other notable features include sex (10.65%), suicide attempts (9.20%), and suicidal thinking (7.90%). The remaining factors, in

descending order of importance, were household income, sadness and despair, academic achievement, stress status, age, and region.

Code availability

Based on the results of ML model, we established a web-based application for policy implementation or health system management to support in their decision-making process for cases involving substance use in adolescents (website: https://predictsubstance.streamlit.app/). An example of a web interface and the results are shown in Figure S1. Custom code for the website is available online at https://github.com/centerfordh/predict_substance/.

Discussion

Our study stands out as one of the first comprehensive machine-learning based approaches to predict adolescent substance use on an international scale. One of our critical insights revolves around the influence of cultural diversity on substance use, drawing datasets from South Korea, the United States, and Norway. Moreover, the outcome revealed that the random forest model proves to be commendable. It displayed predictive capabilities with an AUROC score of 80.75% (95% CI, 80.69-80.82) in the discovery dataset. Our model consistently exhibited robust performance across external validation sets, achieving AUROC scores of 73.20% and 75.69% in each respective dataset. Upon closer inspection, we discerned pivotal features influencing adolescent substance use predictions. Smoking status emerged as the predominant predictor for substance use, followed by BMI and alcoholic consumption. To apply our findings to real-world scenarios, we devised a cutting-edge web-based platform. We believe that this tool will serve as an insightful methodology for the public to navigate potential substance-related challenges.

The role of smoking status and alcoholic consumption in broader adolescent substance use cannot be understated. Neurobiological evidence indicates that nicotine, especially when introduced during formative years, can alter the brain reward pathways, making other substances more appealing.[23] Similarly, alcohol can modify neurotransmitter levels during these formative years, making the brain more susceptible to effects from other substances.[24] Smoking and drinking during adolescence often signal risk-taking behaviors[25], leading to further experimentation with other substances.

Both smoking and alcoholic consumption align closely with societal expectations and peer pressures. Societal dynamics and peer interactions, which normalize or even glorify smoking and drinking, serve as powerful catalysts pushing adolescents beyond smoking or drinking.[26] In many cultures, both behaviors are viewed as rites of passage that expose adolescents to other available illicit substances.[27]

Psychological factors also play a role. Many adolescents resort to smoking or drinking as coping mechanisms for stress or emotional turmoil.[28, 29] These initial coping behaviors may prompt adolescents to seek stronger stimuli for more intense experiences, potentially resulting in substance addiction or misuse.[30]

Our study further emphasized the relationship between BMI and substance use, as another predictor of adolescent substance use. From a physiological perspective, substances can modulate metabolic rates and appetite. Adolescents engaged in substance use might experience weight changes, due either to the direct effects of the substance or inconsistent dietary habits.[31] Additionally, the role of BMI extends beyond mere physiological changes. Adolescents with “non-standard” BMIs often face societal challenges such as weight-centric bullying and entrenched societal ideals regarding body standards. The prevailing societal standards of an ideal body can lead some adolescents towards substance use, either as a means to conform to societal expectations or as relief from the associated mental distress.[32]

Another noteworthy observation was the understated importance of academic achievement and stress status. Although stress is traditionally considered influential in adolescent behaviors[33, 34], its limited representation in our study may be attributed to data constraints. These variables were absent in our extra validation dataset, and we resorted to imputing these values using the median from our primary training cohort. This modification might have contributed to its reduced importance in our results.

The results of our study offer significant insights into both clinical and political implications. We underscore the vital role of factors such as smoking status, BMI, and alcoholic consumption in predicting substance use among adolescents. These critical determinants enable clinicians to identify and monitor at-risk adolescents more effectively, assisting in their decision-making process.[35] Following further refinement, this model has potential commercial viability[36], especially when combined with a streamlined self-report questionnaire. The existence of multiple models assessing

substance use further attests to the commercial potential of our model[37, 38]. Emphasizing characteristics predictive of substance use is essential, suggesting the need for systems to alert parents about potential risks their children might face. Since parental intervention has proven to be effective in preventing adolescent substance use[39], establishing an early detection system becomes paramount.

Findings from this study must be interpreted in light of several limitations. The external validation datasets revealed an abundance of missing values which could have potentially degraded the overall predictive accuracy of the model. To address this issue, we resorted to interpolation using the median value from our discovery dataset.[40] Additionally, our study used discovery dataset derived from adolescents in South Korea. This biased discovery dataset could unexpectedly reflect the specific racial and cultural features unique to Korean adolescents. While our model underwent external validation from diverse cultural and demographical landscape, we also acknowledge that it may reduce sample diversity and potentially cause overfitting issues.[41] Furthermore, this study did not pinpoint a definitive causal link between the significant risk factors and adolescent substance use. In other words, it remains unclear whether substance use influences other factors or if those factors stimulate substance use. Thus, further comprehensive studies are needed to elucidate this intricate cause-and-effect relationship.

Despite these limitations, this study offers significant contributions. By utilizing extensive datasets from South Korea, the United States, and Norway, our ML model boasts enhanced prediction accuracy, highlighting its global relevance and robustness.[42] Our strategic phased validation approach, beginning with the Asian subset of the YRBS and progressing to the distinct Ungdata from Norway, underlines the model versatility across diverse socio-cultural backgrounds. This phased validation not only ensures consistent model evaluation but also establishes its capability in different cultural contexts.[43] Moreover, the features incorporated into the model are derived from simple questionnaires. The primary advantage of our model-based platform is its

exceptional accessibility, allowing users to gather insights through straightforward surveys. This ease enables swift evaluations and widens its scope of use, equipping both clinicians and individuals with valuable insights conveniently. Our findings provide relative importance of numerous factors. These results can guide decision-making process by identifying key areas for the prevention of substance use among adolescents.

Conclusions

This study introduced a ML model using data from three distinct national cohorts to predict adolescent substance use. Among five unique predictive models, the random forest model consistently revealed a notable performance (AUROC: KYRBS, 80.75% [discovery]; YRBS, 73.20% [extra-validation]; and Ungdata, 75.69% [extra-validation]). Further analysis of feature importance revealed that factors such as smoking status, BMI, and alcoholic consumption contribute to a risk of substance use. The findings of this study indicate the potential of ML-driven predictive models to swiftly predict the likelihood of substance use among adolescents using a simplistic survey. It is anticipated that with further refinement and development, these models could be broadly employed as efficient tools for preventing adolescent substance use.

Funding: This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C1976). The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Conflicts of interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Contributors

Dr DKY had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final version before submission. *Study concept and design*: HL, HK, SK, SW, and DKY; *Acquisition, analysis, or interpretation of data*: HL, HK, SK, SW, and DKY; *Drafting of the manuscript*: HL, HK, SK, SW, and DKY; *Critical revision of the manuscript for important intellectual content*: all authors; *Statistical analysis*: HL, HK, SK, SW, and DKY; *Study supervision*: DKY. DKY supervised the study and is guarantor for this study. Hojae L, HK, and SK contributed equally as co-first authors. DKY, SW, and RK contributed equally as co-corresponding authors. DKY is the senior author. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Acknowledgments

None

Data Sharing Statement

Data are available on reasonable request. Study protocol, statistical code: available from DKY (email: yonkkang@gmail.com). Data set: available from the Korean Disease Control and Prevention Agency (KDCA), U.S. Centers for Disease Control and Prevention (CDC), and Norwegian Social

Research institute (NOVA) through a data use agreement.

References

1. Davidson LL, Grigorenko EL, Boivin MJ, Rapa E, Stein A. A focus on adolescence to reduce neurological, mental health and substance-use disability. Nature. 2015 2015/11/01;527(7578):S161-S6. doi: 10.1038/nature16030.

2. Volkow ND, Wargo EM. Association of Severity of Adolescent Substance Use Disorders and Long-term Outcomes. JAMA Network Open. 2022;5(4):e225656-e. doi: 10.1001/jamanetworkopen.2022.5656.

3. Miething A, Almquist YB. Childhood peer status and circulatory disease in adulthood: a prospective cohort study in Stockholm, Sweden. BMJ Open. 2020;10(9):e036095. doi: 10.1136/bmjopen-2019-036095.

4. Williams EC, Fletcher OV, Frost MC, Harris AHS, Washington DL, Hoggatt KJ. Comparison of Substance Use Disorder Diagnosis Rates From Electronic Health Record Data With Substance Use Disorder Prevalence Rates Reported in Surveys Across Sociodemographic Groups in the Veterans Health Administration. JAMA Network Open. 2022;5(6):e2219651-e. doi: 10.1001/jamanetworkopen.2022.19651.

5. Rodríguez-Cano R, Kypriotakis G, Cortés-García L, Bakken A, von Soest T. Polysubstance use and its correlation with psychosocial and health risk behaviours among more than 95,000 Norwegian adolescents during the COVID-19 pandemic (January to May 2021): a latent profile analysis. The Lancet Regional Health - Europe. 2023 2023/05/01/;28:100603. doi: https://doi.org/10.1016/j.lanepe.2023.100603.

6. Chaffee BW, Cheng J, Couch ET, Hoeft KS, Halpern-Felsher B. Adolescents' Substance Use and Physical Activity Before and During the COVID-19 Pandemic. JAMA Pediatrics. 2021;175(7):715-22. doi: 10.1001/jamapediatrics.2021.0541.

7. Kim J, Lee H, Lee J, Rhee SY, Shin JI, Lee SW, et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning:

a post hoc analysis of a diagnostic trial and validation of an external additional trial. Alzheimers Res Ther. 2023 Jul 22;15(1):127. PMID: 37481573. doi: 10.1186/s13195-023-01268-9.

8. Lam JY, Shimizu C, Tremoulet AH, Bainto E, Roberts SC, Sivilay N, et al. A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the USA: a retrospective model development and validation study. The Lancet Digital Health. 2022 2022/10/01/;4(10):e717-e26. doi: https://doi.org/10.1016/S2589-7500(22)00149-2.

9. Clift AK, Dodwell D, Lord S, Petrou S, Brady M, Collins GS, et al. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. BMJ. 2023;381:e073800. doi: 10.1136/bmj-2022-073800.

10. Woo HG, Park S, Yon H, Lee SW, Koyanagi A, Jacob L, et al. National Trends in Sadness, Suicidality, and COVID-19 Pandemic-Related Risk Factors Among South Korean Adolescents From 2005 to 2021. JAMA Netw Open. 2023 May 1;6(5):e2314838. PMID: 37223902. doi: 10.1001/jamanetworkopen.2023.14838.

11. Boakye E, Erhabor J, Obisesan O, Tasdighi E, Mirbolouk M, Osuji N, et al. Comprehensive review of the national surveys that assess E-cigarette use domains among youth and adults in the United States. The Lancet Regional Health - Americas. 2023 2023/07/01/;23:100528. doi: https://doi.org/10.1016/j.lana.2023.100528.

12. Kaldenbach S, Strand TA, Solvik BS, Holten-Andersen M. Social determinants and changes in energy drink consumption among adolescents in Norway, 2017–2019: a cross-sectional study. BMJ Open. 2021;11(8):e049284. doi: 10.1136/bmjopen-2021-049284.

13. Kwon R, Lee H, Kim MS, Lee J, Yon DK. Machine learning-based prediction of suicidality in adolescents during the COVID-19 pandemic (2020-2021): Derivation and validation in two independent nationwide cohorts. Asian J Psychiatr. 2023 Jul 22;88:103704. PMID: 37541104. doi: 10.1016/j.ajp.2023.103704.

14. Johnston KC, Connors AF, Wagner DP, Haley EC. Predicting Outcome in Ischemic Stroke.

Stroke. 2003;34(1):200-2. doi: doi:10.1161/01.STR.0000047102.61863.E3.

15. Nguipdop-Djomo P, Rodrigues LC, Smith PG, Abubakar I, Mangtani P. Drug misuse, tobacco smoking, alcohol and other social determinants of tuberculosis in UK-born adults in England: a community-based case-control study. Scientific Reports. 2020 2020/03/27;10(1):5639. doi: 10.1038/s41598-020-62667-8.

16. Compton WM, Flannagan KSJ, Silveira ML, Creamer MR, Kimmel HL, Kanel M, et al. Tobacco, Alcohol, Cannabis, and Other Drug Use in the US Before and During the Early Phase of the COVID-19 Pandemic. JAMA Network Open. 2023;6(1):e2254566-e. doi: 10.1001/jamanetworkopen.2022.54566.

17. Kim N, Song JY, Yang H, Kim MJ, Lee K, Shin YH, et al. National trends in suicide-related behaviors among youths between 2005-2020, including COVID-19: a Korean representative survey of one million adolescents. Eur Rev Med Pharmacol Sci. 2023 Feb;27(3):1192-202. PMID: 36808368. doi: 10.26355/eurrev_202302_31226.

18. Huang J. Drug licensing as evidence of evolution, diffusion and catch-up in East Asia. Nature Biotechnology. 2023 2023/02/01;41(2):189-92. doi: 10.1038/s41587-023-01659-1.

19. Kim J, Kim SC, Kang D, Yon DK, Kim JG. Classification of Alzheimer's disease stage using machine learning for left and right oxygenation difference signals in the prefrontal cortex: a patient-level, single-group, diagnostic interventional trial. Eur Rev Med Pharmacol Sci. 2022 Nov;26(21):7734-41. PMID: 36394721. doi: 10.26355/eurrev_202211_30122.

20. Bolo K, Aroca GA, Pardeshi AA, Chiang M, Burkemper B, Xie X, et al. Automated expert-level scleral spur detection and quantitative biometric analysis on the ANTERION anterior segment OCT system. British Journal of Ophthalmology. 2023:bjo-2022-322328. doi: 10.1136/bjo-2022-322328.

21. Leonhardt M, Granrud MD, Bonsaksen T, Lien L. Associations between Mental Health, Lifestyle Factors and Worries about Climate Change in Norwegian Adolescents. Int J Environ Res

Public Health. 2022 Oct 7;19(19). PMID: 36232127. doi: 10.3390/ijerph191912826.

22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ : British Medical Journal. 2015;350:g7594. doi: 10.1136/bmj.g7594.

23. Le Foll B, Piper ME, Fowler CD, Tonstad S, Bierut L, Lu L, et al. Tobacco and nicotine use. Nature Reviews Disease Primers. 2022 2022/03/24;8(1):19. doi: 10.1038/s41572-022-00346-w.

24. Yip SW, Lichenstein SD, Liang Q, Chaarani B, Dager A, Pearlson G, et al. Brain Networks and Adolescent Alcohol Use. JAMA Psychiatry. 2023. doi: 10.1001/jamapsychiatry.2023.2949.

25. Kruckow S, Santini ZI, Hjarnaa L, Becker U, Andersen O, Tolstrup JS. Associations between alcohol intake and hospital contacts due to alcohol and unintentional injuries in 71,025 Danish adolescents – a prospective cohort study. eClinicalMedicine. 2023 2023/09/06/:102187. doi: https://doi.org/10.1016/j.eclinm.2023.102187.

26. Zhao Y, Di X, Li S, Zeng X, Wang X, Nan Y, et al. Prevalence, frequency, intensity, and location of cigarette use among adolescents in China from 2013–14 to 2019: Findings from two repeated cross-sectional studies. The Lancet Regional Health - Western Pacific. 2022 2022/10/01/;27:100549. doi: https://doi.org/10.1016/j.lanwpc.2022.100549.

27. Maeng SJ, Lee DJ, Kang JH. First Drinking Experiences during Adolescence in South Korea: A Qualitative Study Focusing on the Internal and External Factors. Int J Environ Res Public Health. 2021 Aug 3;18(15). PMID: 34360493. doi: 10.3390/ijerph18158200.

28. Meienberg A, Mayr M, Vischer A, Zellweger MJ, Burkard T. Smoking prevention in adolescents: a cross-sectional and qualitative evaluation of a newly implemented prevention program in Switzerland. BMJ Open. 2021;11(12):e048319. doi: 10.1136/bmjopen-2020-048319.

29. Skylstad V, Babirye JN, Kiguli J, Skar A-MS, Kühl M-J, Nalugya JS, et al. Are we overlooking alcohol use by younger children? BMJ Paediatrics Open. 2022;6(1):e001242. doi: 10.1136/bmjpo-2021-001242.

30. Cheron J, Kerchove d'Exaerde Ad. Drug addiction: from bench to bedside. Translational Psychiatry. 2021 2021/08/12;11(1):424. doi: 10.1038/s41398-021-01542-0.

31. Treasure J, Duarte TA, Schmidt U. Eating disorders. The Lancet. 2020 2020/03/14/;395(10227):899-911. doi: https://doi.org/10.1016/S0140-6736(20)30059-3.

32. Bornioli A, Lewis-Smith H, Slater A, Bray I. Body dissatisfaction predicts the onset of depression among adolescent females and males: a prospective study. Journal of Epidemiology and Community Health. 2021;75(4):343-8. doi: 10.1136/jech-2019-213033.

33. Sylvestre M-P, Dinkou GDT, Naja M, Riglea T, Pelekanakis A, Bélanger M, et al. A longitudinal study of change in substance use from before to during the COVID-19 pandemic in young adults. The Lancet Regional Health - Americas. 2022 2022/04/01/;8:100168. doi: https://doi.org/10.1016/j.lana.2021.100168.

34. Liu M, Koh KA, Hwang SW, Wadhera RK. Mental Health and Substance Use Among Homeless Adolescents in the US. JAMA. 2022;327(18):1820-2. doi: 10.1001/jama.2022.4422.

35. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364:l886. doi: 10.1136/bmj.l886.

36. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health. 2021 2021/03/01/;3(3):e195-e203. doi: https://doi.org/10.1016/S2589-7500(20)30292-2.

37. Sharma V, Kulkarni V, Jess E, Gilani F, Eurich D, Simpson SH, et al. Development and Validation of a Machine Learning Model to Estimate Risk of Adverse Outcomes Within 30 Days of Opioid Dispensation. JAMA Network Open. 2022;5(12):e2248559-e. doi: 10.1001/jamanetworkopen.2022.48559.

38. Lo-Ciganic W-H, Donohue JM, Yang Q, Huang JL, Chang C-Y, Weiss JC, et al. Developing

and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. The Lancet Digital Health. 2022 2022/06/01/;4(6):e455-e65. doi: https://doi.org/10.1016/S2589-7500(22)00062-0.

39. Kuntsche S, Kuntsche E. Parent-based interventions for preventing or reducing adolescent substance use - A systematic literature review. Clin Psychol Rev. 2016 Apr;45:89-101. PMID: 27111301. doi: 10.1016/j.cpr.2016.02.004.

40. Pokorney SD, Holmes DN, Thomas L, Fonarow GC, Kowey PR, Reiffel JA, et al. Association Between Warfarin Control Metrics and Atrial Fibrillation Outcomes in the Outcomes Registry for Better Informed Treatment of Atrial Fibrillation. JAMA Cardiology. 2019;4(8):756-64. doi: 10.1001/jamacardio.2019.1960.

41. Blackburn AM, Vestergren S, Blackburn AM, Vestergren S, Tran TP, Stöckli S, et al. COVIDiSTRESS diverse dataset on psychological and behavioural outcomes one year into the COVID-19 pandemic. Scientific Data. 2022 2022/06/21;9(1):331. doi: 10.1038/s41597-022-01383-6.

42. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2019;365:l4379. doi: 10.1136/bmj.l4379.

43. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA. 2020;324(12):1212-3. doi: 10.1001/jama.2020.12067.

Table 1. Demographic characteristics of KYRBS from South Korea (2005-2022), YRBS from United States (1998-2022), and Ungdata from Norway (2014-2021).

Nation	South Korea	United States	Norway
Characteristics	KYRBS	YRBS	Ungdata
Number, n	1,145,178	54,709	720,812
Region, n (%)			
Urban	529,880 (46.27)	NA	NA
Rural	615,298 (53.73)	NA	NA
Age, years, n (%)			
13	196,311 (17.14)	433 (0.79)	147,978 (20.53)
14	197,221 (17.22)	8,718 (15.94)	145,170 (20.14)
15	197,628 (17.26)	13,665 (24.98)	147,337 (20.44)
16	190,535 (16.64)	14,025 (25.64)	131,573 (18.25)
17	189,638 (16.56)	12,460 (22.78)	92,929 (12.89)
18	173,845 (15.18)	5,408 (9.89)	55,825 (7.74)
Sex, n (%)			
Male	554,445 (48.42)	26,817 (49.02)	358,509 (49.74)
Female	590,733 (51.58)	27,892 (50.98)	362,303 (50.26)
BMI[a], n (%)			
Unknown	90,365 (7.89)	NA	NA
Underweight	85,341 (7.45)	9,604 (17.55)	NA
Normal	801,249 (69.97)	28,321 (51.77)	NA
Overweight	86,541 (7.56)	7,095 (12.97)	NA
Obese	81,682 (7.13)	9,689 (17.71)	NA
Academic achievement, n (%)			
Low (0-19 percentile)	123,284 (10.77)	NA	NA
Lower-middle (20-39 percentile)	268,430 (23.44)	NA	NA
Middle (40-59 percentile)	325,005 (28.38)	NA	NA
Upper-middle (60-79 percentile)	287,599 (25.11)	NA	NA
High (80-100 percentile)	140,860 (12.3)	NA	NA
Household income, n (%)			
Low (0-19 percentile)	48,080 (4.2)	NA	315,700 (43.8)
Lower-middle (20-39 percentile)	169,372 (14.79)	NA	242,279 (33.61)
Middle (40-59 percentile)	536,547 (46.85)	NA	125,524 (17.41)
Upper-middle (60-79 percentile)	298,734 (26.09)	NA	28,798 (4)
High (80-100 percentile)	92,445 (8.07)	NA	8,511 (1.18)
Smoking status, n (%)			
Non-smoker	901,709 (78.74)	48,339 (88.36)	581,652 (80.69)

Smoker	243,469 (21.26)	6,370 (11.64)	139,160 (19.31)
Alcoholic consumption, n (%)			
Non-drinker	1,049,989 (91.69)	46,049 (84.17)	323,931 (44.94)
More than 1 time	95,189 (8.31)	8,660 (15.83)	396,881 (55.06)
Stress status[b], n (%)			
Low	32,848 (2.87)	NA	NA
Mild	167,701 (14.64)	NA	NA
Moderate	473,556 (41.35)	NA	NA
High	336,723 (29.4)	NA	NA
Severe	134,350 (11.73)	NA	NA
Sadness and despair in the past year, n (%)			
No	783,884 (68.45)	40,475 (73.98)	523,929 (72.69)
Yes	361,294 (31.55)	14,234 (26.02)	196,883 (27.31)
Suicidal thinking in the past year, n (%)			
No	953,344 (83.25)	45,158 (82.54)	NA
Yes	191,834 (16.75)	9,551 (17.46)	NA
Suicide attempts in the past year, n (%)			
No	1,101,092 (96.15)	51,721 (94.54)	NA
Yes	44,086 (3.85)	2,988 (5.46)	NA
Substance usage, n (%)			
No	1,130,630 (98.73)	45,443 (83.06)	705,694 (97.9)
Yes	14,548 (1.27)	9,266 (16.94)	15,118 (2.1)

Abbreviations: KYRBS, Korea Youth Risk Behavior Web-based Survey; Ungdata, Norwegian nationwide Ungdata surveys; YRBS, Youth Risk Behavior Survey.

[a] BMI, body mass index; BMI was divided into four groups according to the 2017 Korean National Growth Charts: underweight (0-4 percentile), normal (5-84 percentile), overweight (85-94 percentile), and obese (95-100 percentile).

[b] Stress was defined by receipt of mental health counseling owing to stress.

Figure 1. Study population. Abbreviations: KYRBS, Korea Youth Risk Behavior Web-based Survey; Ungdata, Norwegian nationwide Ungdata surveys; YRBS, Youth Risk Behavior Survey of United States adolescent.

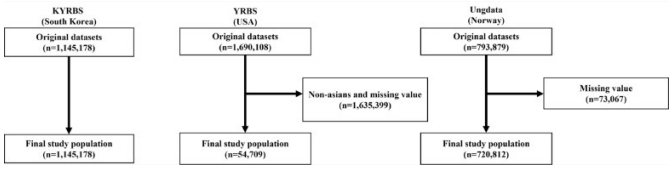


Figure 2. Model architecture. The original KYRBS was partitioned into original dataset for model development, with performance assessed using AUROC score. Selected high-performing models were further validated. The external validations were generated using YRBS and Ungdata. Abbreviations: AUROC, Area Under the Receiver Operating Characteristic Curve; KYRBS, Korea Youth Risk Behavior Web-based Survey; Ungdata, Norwegian nationwide Ungdata surveys; YRBS, Youth Risk Behavior Survey.

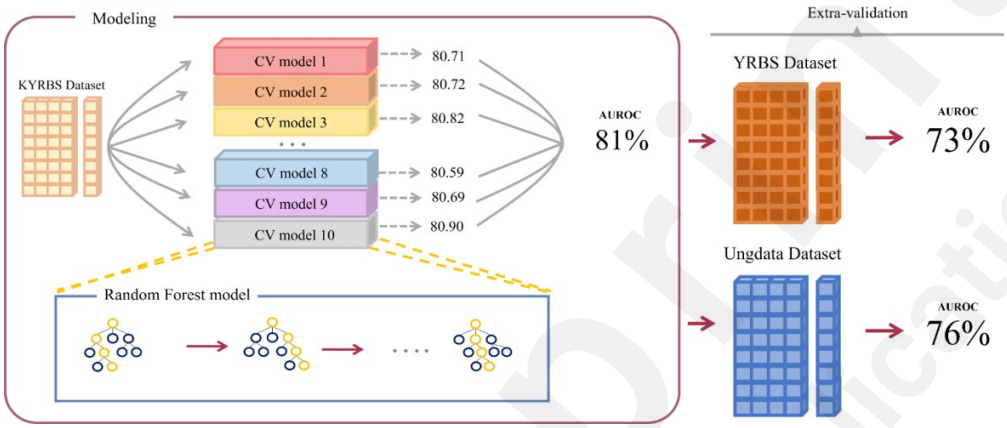


Figure 3. The assessment of five different machine learning algorithms using AUROC score and ROC curve for initial model construction with the KYRBS, and external validation with the YRBS and Ungdata. Abbreviations: AdaBoost, Adaptive Boosting; AUROC, Area Under the Receiver Operating Characteristic Curve; CI, confidence interval; KYRBS, Korea Youth Risk Behavior Web-based Survey; LightGBM, Light Gradient Boosting Model; ROC, Receiver Operating Characteristic; Ungdata, Norwegian nationwide Ungdata surveys; XGBoost, Extreme Gradient Boosting model; YRBS, Youth Risk Behavior Survey.

Nation	Dataset	Model	AUROC	Sensitivity	Specificity	Accuracy	Balanced Accuracy
South Korea	KYRBS	**Random forest**	**80.75(80.69, 80.82)**	**72.91(72.84, 72.98)**	**72.91(72.85, 72.98)**	**72.91(72.84, 72.98)**	**72.91(72.84, 72.98)**
		CatBoost	80.56(80.49, 80.63)	72.69(72.63, 72.76)	72.69(72.63, 72.75)	72.69(72.63, 72.75)	72.69(72.63, 72.75)
		AdaBoost	80.54(80.47, 80.61)	72.77(72.71, 72.83)	72.77(72.71, 72.83)	72.77(72.71, 72.83)	72.77(72.71, 72.83)
		LightGBM	80.48(80.41, 80.55)	72.74(72.66, 72.82)	72.74(72.67, 72.82)	72.74(72.67, 72.82)	72.74(72.67, 72.82)
		XGBoost	80.12(80.05, 80.19)	72.49(72.42, 72.55)	72.49(72.42, 72.56)	72.49(72.42, 72.55)	72.49(72.42, 72.55)
USA	YRBS	**Random forest**	**73.20**	**67.89**	**68.26**	**68.20**	**68.08**
		CatBoost	72.31	66.07	67.01	66.85	66.54
		AdaBoost	71.33	65.39	64.93	65.01	65.16
		LightGBM	72.08	65.99	68.74	68.27	67.37
		XGBoost	70.47	66.42	69.11	68.26	67.76
Norway	Ungdata	**Random forest**	**75.69**	**72.38**	**72.69**	**72.68**	**72.54**
		CatBoost	71.88	66.92	66.83	66.83	66.87
		AdaBoost	70.33	67.10	66.59	66.60	66.84
		LightGBM	71.90	66.03	66.04	66.04	66.04
		XGBoost	70.52	66.42	69.11	68.26	67.76

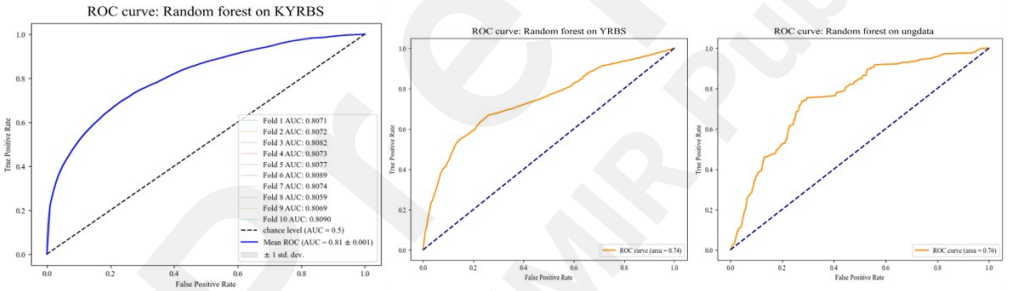
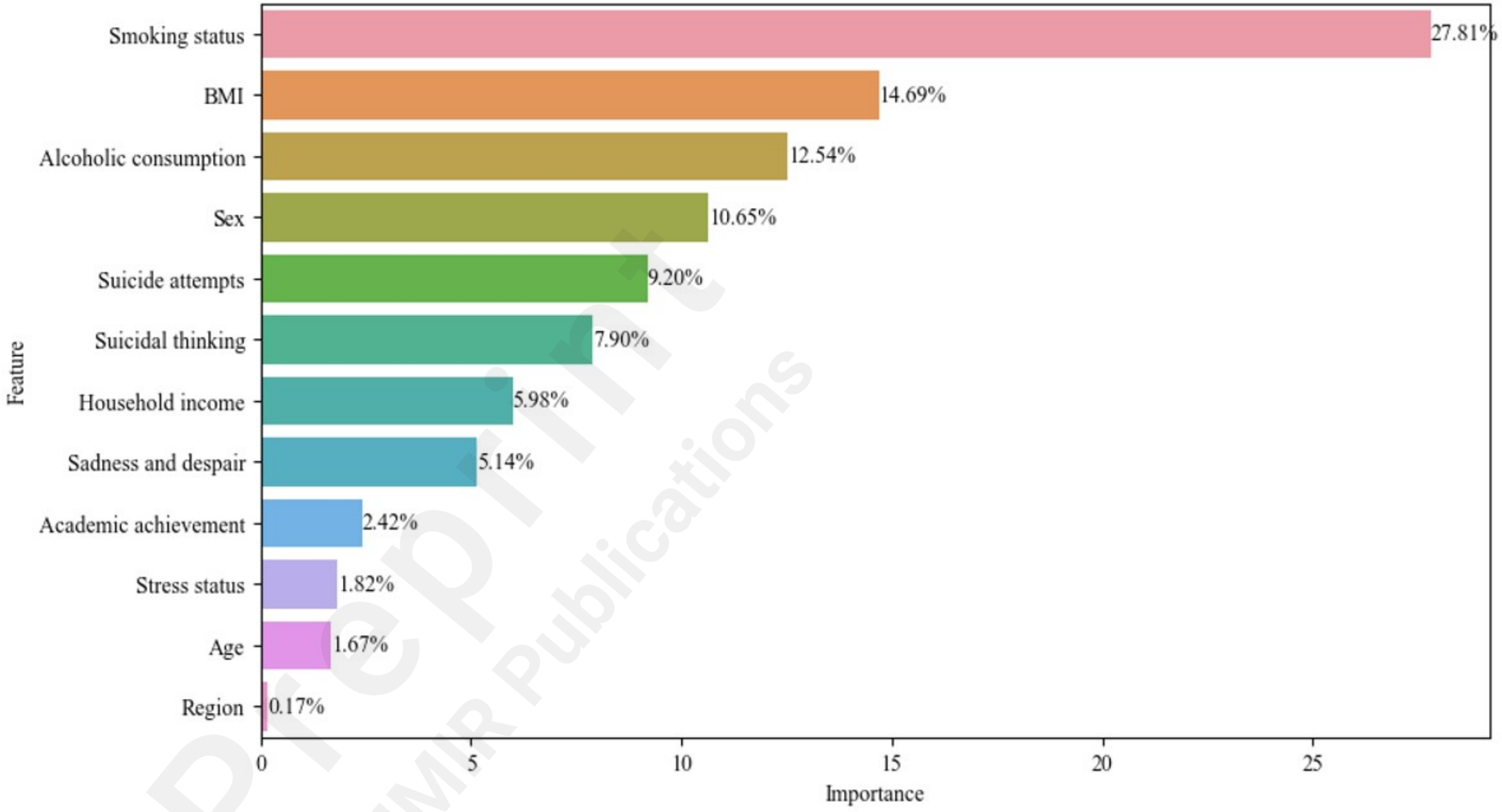


Figure 4. Feature importance of the random forest model. Abbreviations: BMI, body mass index.



Supplementary Files