A Prognostic Tool to Identify Youth at Risk of at Least Weekly Cannabis Use

Marie-Pierre Sylvestre, Simon de Montigny, Laurence Boulanger, Danick Goulet, Isabelle Doré, Jennifer O'Loughlin, Slim Haddad, Richard E. Bélanger, and Scott Leatherdale

Abstract We developed and validated an 8-item prognostic tool to identify youth at risk of initiating frequent (i.e., at least weekly) cannabis use in the next year. The tool, which aims to identify youth who would benefit most from clinician intervention, can be completed by the patient or clinician using a computer or smart phone application prior to or during a clinic visit. Methodological challenges in developing the tool included selecting a parsimonious model from a set of correlated predictors with missing data. We implemented Bach's bolasso algorithm which combines lasso with bootstrap and investigated the performance of the prognostic tool in new data

M.-P. Sylvestre (✉) · S. de Montigny · D. Goulet · J. O'Loughlin
Université de Montréal, 7101 avenue du Parc, Montréal, Canada
e-mail: marie-pierre.sylvestre@umontreal.ca

S. de Montigny
e-mail: simon.de.montigny@umontreal.ca

D. Goulet
e-mail: danick.goulet@umontreal.ca

J. O'Loughlin
e-mail: jennifer.oloughlin@umontreal.ca

L. Boulanger
Centre de recherche du CHUM, 800 St-Denis, Montréal, Canada
e-mail: laurence.boulanger@umontreal.ca

I. Doré
Université de Montréal, 2100, boul. Édouard-Montpetit, Montréal, Canada
e-mail: isabelle.dore@umontreal.ca

S. Haddad · R. E. Bélanger
Université Laval, 1050, avenue de la Médecine, Québec, Canada
e-mail: slim.haddad@fmed.ulaval.ca

R. E. Bélanger
e-mail: richard.belanger@chudequebec.ca

S. Leatherdale
University of Waterloo, 200 University Avenue West, Waterloo, Canada
e-mail: sleatherdale@uwaterloo.ca

A. Shaban-Nejad et al. (eds.), *AI for Disease Surveillance and Pandemic Intelligence*,
Studies in Computational Intelligence 1013,
https://doi.org/10.1007/978-3-030-93080-6_4

collected in a different time period (temporal validation) and in another location (geographic validation). The tool showed adequate discrimination abilities, as reflected by a c-statistic above 0.8, in both validation samples. Most predictors selected into the tool pertained to substance use including use of cigarettes, e-cigarettes, alcohol and energy drinks mixed with alcohol, but not to mental or physical health.

Keywords Cannabis use · Adolescents · Prognostic tool

1 Introduction

Canadian youth have one of the highest prevalence rates of cannabis use in developed countries, with a past 12-month prevalence of 18% in 2018–19 [1]. Average age at first cannabis use is 14 years in Canada [1], and one in every six individuals who try cannabis during adolescence further develop problematic use [2]. Age of onset and frequency of cannabis use during this critical developmental period [4] are strongly associated with adverse cannabis-related health impacts including detrimental effects on the structure and function of the brain [3]. Addressing this issue before an adolescent becomes a frequent cannabis user could increase the likelihood of successful intervention [5]. However, although validated screening tools exist [6], they generally aim to identify problematic cannabis users (i.e., individuals already on the pathological spectrum of use disorder) [7]. No validated tool to date aims to identify adolescents at risk of weekly use (i.e., before pathological use is established), a key milestone in the natural course of problematic cannabis use that is strongly associated with adverse health outcomes.

The objective of this paper is to describe the development and validation of a short, easy-to-administer screening tool to identify youth who are at the greatest risk of initiating weekly cannabis use in the next year. The tool is intended for use in clinical practice since clinicians report lack of time (in addition to lack of knowledge about available tools) as a major barrier to discussing psychoactive substance use with adolescents [8]. The tool can be completed using an online application prior to or during a clinic visit so that clinicians can identify who would benefit most from intervention. A recent study on the perceptions of pediatric primary care providers regarding computer-administered screening tools for substance use suggested high utility, acceptability and feasibility [9].

2 Methods

The current study adheres to TRIPOD statement for the development and validation of prediction models [10].

Data source

Data were drawn from COMPASS, an ongoing prospective study (inception 2012–13) of grade 9–12 students in a convenience sample of Canadian high schools which was designed to investigate how changes in the school environment and in provincial, territorial, and national policies affect youth health behaviours [11]. Students complete in-class self-report questionnaires annually, that assess demographics, health behaviours and school-related characteristics [11]. The current study uses data from the 61 schools sampled in Ontario (ON) in 2016–18 and the 36 schools sampled in Québec (QC) in 2017–18. The University of Waterloo Research Ethics Board, and the Research Ethics Review Board of the Centre intégré universitaire de santé et de services sociaux de la Capitale-Nationale approved the COMPASS study in Ontario and Québec, respectively. Students were recruited in participating schools using active-information passive-consent permission protocols [12]. Parents/guardians received an information letter about the COMPASS study by mail and could opt out of the study by emailing or calling the COMPASS recruitment coordinator. Students could withdraw from the study at any time during the consent or data collection procedures without prejudice [11].

Study variables
Frequency of cannabis use was measured using the question "In the last 12 months, how often did you use marijuana or cannabis (a joint, pot, weed, hash)?" (never; not in the past 12 months; less than once a month; once a month; 2 or 3 times per month; once a week; 2 or 3 times per week; 4–6 times per week; every day). Participants were categorized as at least weekly cannabis users (yes, no) if they used cannabis at least weekly. A total of 45 potential predictor variables were selected based on the literature on risk factors for weekly or more frequent cannabis use, and included sociodemographic characteristics, indicators of substance use (i.e., cannabis, alcohol, tobacco and nicotine), personality traits, mental health, school connectedness, bullying/victimization, academic achievement, and health behaviours (e.g., physical activity, nutrition, sleep). Most potential predictor variables were coded as binary indicators to facilitate administration of the prognostic tool, with the value of '1' indicating the presence of the factor.

Data preprocessing
We created three analytical samples, each with two waves of data collection. The training sample was drawn from 13,759 participants who completed questionnaires in 2017 and 2018 in Ontario. A subset of 9174 participants had complete data on the 45 predictors and were used to train the prognostic tool. We assumed data were missing completely at random, but we also considered multiple imputation under the assumption that data were missing at random, as described in sensitivity analyses. A temporal validation sample including 13,652 participants who completed questionnaires in Ontario in 2016 and 2017 was used to assess variation of the performance of the tool in the same location (Ontario) but in a different time period. The 2017–18 Ontario sample was selected for training the tool because it contained mental health variables that were not available in the 2016-17 Ontario sample. Finally, a geographical validation sample included 9435 participants who completed questionnaires in 2017–18 in Québec. This sample was used to test transferability of the model to other locations (i.e., in the same time period during which the prognostic tool was developed). A total of 6199 participants provided data for both the training and temporal

validation. However, the sample used for geographical validation was independent of the training sample, and thus represents an external validation. Predictors were measured in the first wave of each sample, and the event (i.e., cannabis use at least weekly) was assessed a year later, in the second wave. Adolescents who had already used cannabis at least weekly in the first wave of each sample were excluded.

Algorithm
Selection of predictors was undertaken in the training sample using the bolasso algorithm proposed by Bach [13], which combines the variable selection algorithm of lasso (i.e., least absolute shrinkage and selection operator) with bootstrap aggregating (i.e., bagging), to improve the stability and accuracy of the prediction. Use of bootstrap also addresses the potential issue of strongly correlated predictors, which may lead to inconsistent estimators with lasso [13]. As with lasso, model selection is performed by penalizing coefficients of less influential variables to exactly zero. However, with bolasso, a variable enters the final model only if selected in most bootstrapped copies. Our implementation of the bolasso algorithm uses 100 bootstrap copies and required variables to be selected in 99 of the 100 bootstrap copies to be included in the final model. In each of the 100 bootstrap copies, the hyperparameter controlling the level of shrinkage was selected by minimizing the AUC statistics using ten-fold cross-validation. The final value for the hyperparameter was obtained by averaging over the 100 resulting values. Coefficients of the variables selected by the bolasso algorithm were estimated using logistic regression. A decision rule to identify adolescents at risk of initiation cannabis use at least weekly in the next year was derived using a utility-based approach that emphasized sensitivity over specificity. This is warranted when the intervention (e.g., counselling) is not invasive, such that intervening with low-risk adolescents is a less important problem than not intervening with those at-risk. Specifically, we selected the lowest threshold that maximized specificity under the constraint that a sensitivity ≥ 0.8 in the training sample.

Assessment of predictive ability
Discrimination was measured using the c-statistic. Calibration plots were used to assess level of agreement between observed and predicted cannabis use at least weekly. Model accuracy was measured using the Brier score, a measure that combines components of both discrimination and calibration. Finally, Spiegelhalter's z-test was used to test the calibration component of the Brier's score, with rejection of the null hypothesis suggesting poor calibration. The performance of the prognostic tool was assessed in the temporal validation sample. In addition, the prognostic tool was compared to a more parsimonious model reflecting data that clinicians could extract from charts and/or ask during a routine visit (i.e., age, sex, questions on cannabis use selected by the bolasso algorithm).

Model update and recalibration
Coefficients for the prognostic tool were re-estimated in the combined training and temporal validation samples, resulting in an updated model and decision rule. Then, the coefficient corresponding to the intercept in the updated prognostic tool was

refitted to Quebec to reflect the difference in the prevalence of cannabis use at least weekly compared to Ontario. This recalibration was done by calculating the difference between the predicted and observed prevalence of cannabis use at least weekly in the geographical validation sample. Recalibration is preferred over developing new tools, since new models waste data, are prone to over-optimism and can contribute to too many non-validated models, limiting uptake of such tools [16].

Sensitivity analyses
A total of six sensitivity analyses were performed to investigate the robustness of our findings. First, we investigated the stability of the selection of predictors under an alternative coding of the binary indicators which minimized loss of information by choosing the cut-point value for dichotomization that led to a distribution that was the closest possible to a 50–50 split. Second, we considered inclusion of two-way interaction terms between predictors in the bolasso algorithm to investigate the appropriateness of a linear representation of the predictors. Third, we investigated the stability of the model by considering a more lenient threshold of 95 for the number of the 100 bootstrap copies that selected a predictor. Fourth, we assessed the impact of the decision rule on sensitivity and specificity by implementing the more conventional Youden index to identify the threshold that optimized both sensitivity and specificity in the training sample. Fifth, we assessed the benefit of sex-specific versions of the prognostic tool by conducting analyses in sex-stratified subsamples of the training sample and by testing the performance of the proposed model separately in each sex. Sixth, in a sensitivity analysis that assumed a missing at random process, we used multiple imputation by chained equations [14] to impute missing values in 10 imputation sets using the 45 prognostic factors in addition to the outcome variable. Missing values were imputed by province and by year, using the entire sample and the original coding for each variable. The bolasso algorithm was adapted as follows: (i) the shrinkage hyperparameter was selected in each of the 10 imputed datasets; and (ii) a variable was selected in the prognostic tool if it was selected in all imputed datasets. Rubin's rule was applied to summarize the estimated coefficients and standard errors obtained from fitting a logistic regression on the selected variables in each imputed dataset.

The data analysis was performed using R version 3.6.3 with the glmnet, rms, plyr, dplyr, tidyr, summarytools, pROC, MICE and mitools packages.

3 Results

The 1-year cumulative incidence of cannabis use at least weekly was 6.3% in the training sample, but lower in the geographical validation sample (3.0%), which included a larger proportion of younger participants. Once standardized to the age distribution in the training sample, the 1-year cumulative incidence of cannabis use at least weekly was 4.2% in the geographical validation sample. The sample-specific distributions of selected predictors are presented in Table 1.

Table 1 Selected descriptive statistics for the training sample (Ontario 2017–18), temporal validation sample (Ontario 2016–17) and geographic validation sample (Québec 2017–18), COMPASS study

	Training sample Ontario 2017-18 n = 11,792	Temporal validation sample Ontario 2016-17 n = 11,743	Geographic validation sample Québec 2017-18 n = 8347
1-year cumulative incidence of regular cannabis use	6.3	5.4	3.0
Sociodemographics			
Age (years), %			
≤12	0.0	0.0	12.0
13	1.3	1.4	24.1
14	32.0	30.7	24.6
15	34.6	34.9	24.5
16	25.5	25.6	13.4
17	6.1	6.8	1.3
≥18	0.5	0.5	0.2
Male, %	46.5	46.6	43.6
Substance use			
Used cannabis in the past 12 months, %	11.5	26.0	7.0
Ever tried cigarettes, %	11.3	11.5	13.6
Smoked a cigarette in the past 30 days, %	3.2	7.4	3.6
Ever tried e-cigarettes, %	24.8	17.2	31.9
Used e-cigarettes in the past 30 days, %	12.0	15.9	15.1
Ever tried alcohol, %	66.6	77.0	69.7
Used alcohol weekly, %	3.9	8.6	4.7
Drink high-energy drinks weekly, %	9.0	11.3	8.4
Mixed alcohol with an energy drink in past 12 months, %	5.7	5.5	7.4

Table 2 Estimated model coefficients for the variables selected by bolasso with and without multiple imputation, COMPASS 2016–2018

Variable	Coefficient (SE)	
	Complete cases	Imputed data
Age (years)	−0.21 (0.03)	−0.12 (0.04)
Male	0.44 (0.06)	0.46 (0.08)
Time since first use of cannabis (years)	0.61 (0.04)	0.49 (0.05)
Ever tried e-cigarettes	0.95 (0.07)	1.03 (0.09)
Perceived easiness of obtaining cannabis (yes)	0.87 (0.07)	0.94 (0.55)
Ever tried cigarettes	0.81 (0.07)	0.74 (0.09)
Failed last math and/or last English/French class(es)	0.50 (0.07)	0.55 (0.10)
Mixed alcohol with an energy drink in the last 12 months	0.49 (0.09)	0.36 (0.10)
Drink high-energy drinks weekly	NA	0.38 (0.10)

The prognostic tool derived using the training sample included 8 predictors of initiating cannabis at least weekly. In addition to sex and age, it included one predictor pertaining to school performance (i.e., failing their last math and/or English/French classe(s)[1]). The remaining five predictors captured substance use, including years since first use of cannabis, perceived easiness of obtaining cannabis, ever smoking cigarettes, ever using e-cigarettes, and mixing alcohol with energy drinks. The estimated coefficients for each selected predictor are shown in the left column of Table 2. The larger estimated coefficients suggested that the predictions were more heavily affected by ever-trying e-cigarettes, followed by perceptions that it was easy to obtain cannabis.

Performance statistics for each sample are shown in Table 3. The overall accuracy of prediction was satisfactory as suggested by the low Brier score in the training and validation samples. The tool also showed adequate discrimination abilities, as reflected by the c-statistic above 0.8 in all samples. The prognostic tool performed better than the reduced model which included age, sex and the two cannabis-related variables (i.e., likelihood ratio test p-value <0.001, and fraction of new predictive information $= 0.32$), suggesting that inclusion of predictors associated with other substance use and school performance increased the predictive ability of the tool considerably.

Calibration plots for the combined validation sample and the geographic validation sample are shown in Figs. 1 and 2. The plots suggested that the prognostic tool was well-calibrated for participants with predicted probabilities of up to 0.4 in the

[1] The language class variable refers to an English class in the Ontario sample and to a French class in the Québec sample.

Table 3 Performance statistics for the prognostic tool as selected from the bolasso algorithm with a 99% selection threshold, COMPASS 2016–18

Training sample—Ontario 2017–18	
Brier score (Spiegelhalter's z-test p-value)	0.052 (0.336)
c-statistic	0.829
Sensitivity (threshold = 0.046)	0.809
Specificity	0.679
Temporal validation sample—Ontario 2016–17	
Brier score (Spiegelhalter's z-test p-value)	0.046 (0.115)
c-statistic	0.809
Sensitivity (threshold = 0.046)	0.745
Specificity	0.732
Combined validation sample—Ontario 2016–18	
Brier score (Spiegelhalter's z-test p-value)	0.049 (0.168)
c-statistic	0.823
Sensitivity (updated threshold = 0.043)	0.801
Specificity	0.690
Geographic validation sample—Québec 2017–18	
Brier score (Spiegelhalter's z-test p-value)	0.027 (0.524)
c-statistic	0.823
Sensitivity (updated threshold = 0.021)	0.803
Specificity	0.705

combined validation data set, but tended to overestimate the prediction of cannabis use at least weekly in participants with higher predicted probabilities. Sensitivity and specificity of the tool in the combined validation dataset corresponded to 0.801 and 0.690, respectively. The tool was then recalibrated to reflect the lower prevalence of cannabis use at least weekly in the geographic validation set, resulting in a sensitivity and specificity of 0.803 and 0.705.

Sensitivity analyses

Changing the bootstrap threshold for selecting predictors or using Youden's rule to determine the decision rule did not have significant impact on the model or its performance. Use of multiple imputation to address missing values led to selection of the same predictors as in the complete case sample, with one additional predictor (Drink high-energy drinks weekly). We elected to omit this additional predictor from the main tool for the sake of parsimony, but also because its contribution to the prediction was relatively small in comparison to the eight original predictors, as evidenced by its smaller estimated coefficient. In addition, consuming energy drinks was already considered in the tool in another format that focused on its combined use with alcohol (i.e. had alcohol mixed with energy drink (last 12 months)). Similarly, using a different cut-point for dichotomizing the predictors did lead to significant

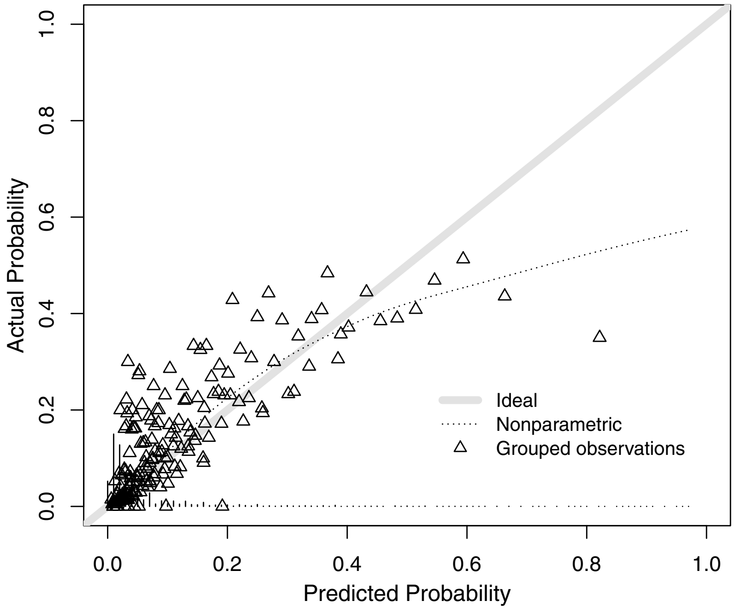

Fig. 1 Calibration plot for combined sets, COMPASS 2016–18

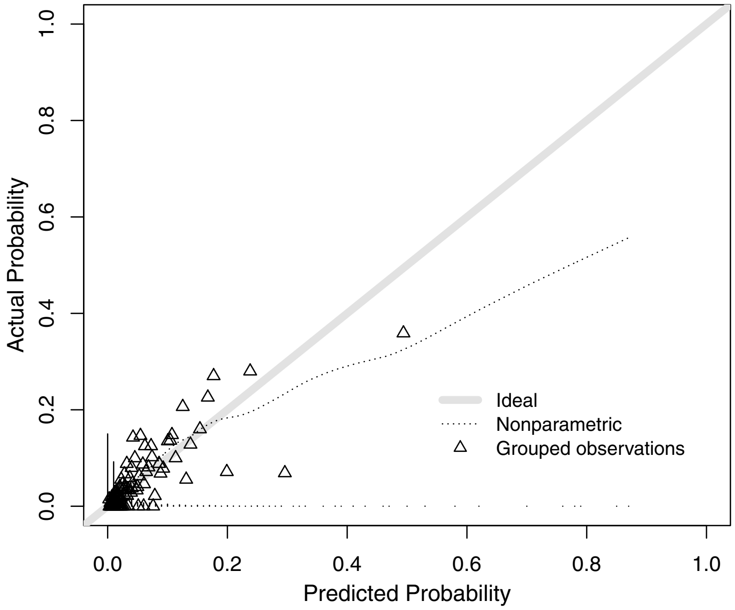

Fig. 2 Calibration plot for geographic sets, COMPASS 2016–18

improvements in the performance of the tool because the variables with the highest estimated coefficients were selected (e.g., age, sex, years since first use of cannabis, perceived easiness of obtaining cannabis, ever smoking cigarettes, ever using e-cigarettes). Inclusion of two-way interaction terms between the predictors into the bolasso algorithm did not improve the performance of the tool, as suggested by the c-statistics that were systematically lower than those of the tool across the samples. Finally, the predictors selected, and their associated coefficients were similar by sex suggesting that a single model for both sexes combined was adequate.

4 Discussion

We developed and validated a simple but effective prognostic tool to identify youth at risk of initiation cannabis use at least weekly in the next year. Most predictors selected into the tool pertained to substance use including cigarette use, e-cigarette use, mixing alcohol and energy drinks, but not to mental and physical health. The geographical validation suggests that the performance of the tool is robust to settings with different distributions of predictors, including age. The proportion of females who initiated cannabis use at least weekly was smaller than that of males. This may have hampered our ability to identify female-specific predictors of frequent cannabis use and may have driven the coefficients estimated in the model away from those estimated in females only. However, our sex-specific modelling suggests that the main predictors of cannabis use at least weekly are very similar in males and females, and that performance of the prognostic tool was satisfactory in both sexes (data not shown).

If incorporated into routine screening in a variety of clinical settings, our simple and effective prognostic tool could help identify adolescents who most need intervention for cannabis use. Advantages of using standardized and validated prognostic tools in clinical settings to identify young people in need of intervention, have been underscored by diverse stakeholders [3]. Clinicians are generally considered by adolescents to be reliable sources of information on substance use and its associated risk [15] and office settings provide a safe and confidential environment to engage with adolescents on substance use [5]. Close to 90% of Canadian adolescents have access to a health care provider [17] and intervening before an adolescent becomes an at-risk cannabis user may increase the probability of successful intervention [5]. Two possibly helpful complements to use of our prognostic tool include: (i) after completion, patients could be shown age-specific educational content on the risk associated with cannabis use using life story vignettes and scientific information [9]; and (ii) health practitioners could be provided with talking points for brief counselling based on the most recent recommendations from pediatric associations [3].

Acknowledgements The COMPASS study was supported by a bridge grant from the Canadian Institutes of Health Research (CIHR) Institute of Nutrition, Metabolism and Diabetes (OOP-110788; awarded to SL) and an operating grant from the CIHR Institute of Population and Public Health (MOP-114875; awarded to SL). The COMPASS-Quebec project benefits from funding from the Ministère de la Santé et des Services sociaux of the province of Québec, and the Direction régionale de santé publique du CIUSSS de la Capitale-Nationale. MPS is supported by a salary award from *Fonds de recherche du Québec*. JOL holds a Canada Research Chair in the Early Determinants of Adult Chronic Disease.

References

1. Health Canada: Summary of results for 2018 to 2019. In: Canadian Student Tobacco, Alcohol and Drugs Survey (2019) https://www.canada.ca/en/health-canada/services/canadian-student-tobacco-alcohol-drugs-survey.html. Accessed on 18 Sept 2020
2. Volkow, N.D., Baler, R.D., Compton, W.M., Weiss, S.R.: Adverse health effects of marijuana use. N. Engl. J. Med. **370**(23), 2219–2227 (2014)
3. Bélanger, R.E., Grant, C.N.: Counselling adolescents and parents about cannabis: a primer for health professionals. Paediatr. Child Health **25**, S34–S40 (2020)
4. Hasin, D.S.: Epidemiology of Cannabis use and associated problems. Neuropsychopharmacology **43**(1), 195–212 (2018)
5. Laporte, C., Vaillant-Roussel, H., Pereira, B., Blanc, O., Eschalier, B., Kinouani, S., Brousse, G., Llorca, P.M., Vorilhon, P.: Cannabis and young users—a brief intervention to reduce their consumption (CANABIC): a cluster randomized controlled trial in primary care. Ann. Fam. Med. **15**(2), 131–139 (2017)
6. US Preventive Services Task Force: Screening for unhealthy drug use: US preventive services task force recommendation statement. JAMA **323**(22), 2301–2309 (2020)
7. Casajuana, C., López-Pelayo, H., Balcells, M.M., Miquel, L., Colom, J., Gual, A.: Definitions of risky and problematic cannabis use: a systematic review. Subst. Use Misuse **51**(13), 1760–1770 (2016)
8. Van Hook, S., Harris, S.K., Brooks, T., Carey, P., Kossack, R., Kulig, J., Knight, J.R.: The "Six T's": barriers to screening teens for substance abuse in primary care. J. Adolesc. Health **40**(5), 456–461 (2007)
9. Gibson, E.B., Knight, J.R., Levinson, J.A., Sherritt, L., Harris, S.K.: Pediatric primary care provider perspectives on a computer-facilitated screening and brief intervention system for adolescent substance use. J. Adolesc. Health **29**(3), 170–176 (2020)
10. Moons, K.G., Karel, G.M., Altman, D.G., Reitsma, J.B., Ioannidis, J.P.A., Macaskill, P., Steyerberg, E.W., Vickers, A.J., Ransohoff, D.F., Collins, G.S.: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med. **162**(1), W1–W73 (2015)
11. Leatherdale, S.T., Brown, K.S., Carson, V., Childs, R.A., Dubin, J.A., Elliott, S.J., Faulkner, G., Hammond, D., Manske, S., Sabiston, C.M., Laxer, R.E., Bredin, C., Thompson-Haile, A.: The COMPASS study: a longitudinal hierarchical research platform for evaluating natural experiments related to changes in school-level programs, policies and built environment resources. BMC Public Health **14**(1), 331 (2014)
12. Hollmann, C.M., McNamara, J.R.: Considerations in the use of active and passive parental consent procedures. J. Psychol. **133**(2), 141–156 (1999)
13. Bach, F.R.: Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th International Conference on Machine Learning. Association for Computing Machinery, New York, NY, United States (2008)
14. van Buuren, S., Groothuis-Oudshoorn, K.: Mice: multivariate imputation by chained equations in R. J. Stat. Softw. **45**(3), 1–67 (2011)

15. Ackard, D.M., Neumark-Sztainer, D.: Health care information sources for adolescents: age and gender differences on use, concerns, and needs. J. Adolesc. Health **29**(3), 170–176 (2001)
16. Moons, K.G., Kengne, A.P., Grobbee, D.E., Royston, P., Vergouwe, Y., Altman, D.G., Woodward, M.: Risk prediction models: II. External validation, model updating, and impact assessment. Heart **98**, 691–698 (2012)
17. Canadian Institute for Health Information. Primary Health Care in Canada—A Chartbook of Selected Indicator Results (2016)