Identifying smoker subgroups with high versus low smoking cessation attempt probability: A decision tree analysis approach

Hua-Hie Yong[a],*, Chandan Karmakar[b], Ron Borland[c], Shitanshu Kusmakar[b], Matthew Fuller-Tyszkiewicz[a], John Yearwood[b]

[a] *School of Psychology, Deakin University, Geelong, Australia*
[b] *School of Information Technology, Deakin University, Geelong, Australia*
[c] *School of Psychological Sciences, University of Melbourne, Melbourne, Australia*

HIGHLIGHTS

- Smoker subgroup characteristics predictive of making quit attempts can be identified.
- Patterns of smoker subgroup characteristics differ by socio-economic status.
- Decision tree analysis is useful and can complement regression-based approach.

ARTICLE INFO

ABSTRACT

Background: Regression-based research has successfully identified independent predictors of smoking cessation, both its initiation and maintenance. However, it is unclear how these various independent predictors interact with each other and conjointly influence smoking behaviour. As a proof-of-concept, this study used decision tree analysis (DTA) to identify the characteristics of smoker subgroups with high versus low smoking cessation initiation probability based on the conjoint effects of four predictor variables, and determine any variations by socio-economic status (SES).

Methods: Data come from the Australian arm of the ITC project, a longitudinal cohort study of adult smokers followed up approximately annually. Reported wanting to quit smoking, worries about smoking negative health impact, quitting self-efficacy and quit intentions assessed in 2005 were used as predictors and reported quit attempts at the 2006 follow-up survey were used as the outcome for the initial model calibration and validation analyses (n = 1475), and further cross-validated using the 2012–2013 data (n = 787).

Results: DTA revealed that while all four predictor variables conjointly contributed to the identification of subgroups with high versus low smoking cessation initiation probability, quit intention was the most important predictor common across all SES strata. The relative importance of the other predictors showed differences by SES.

Conclusions: Modifiable characteristics of smoker subgroups associated with making a quit attempt and any variations by SES can be successfully identified using a decision tree analysis approach, to provide insights as to who might benefit from targeted intervention, thus, underscoring the value of this approach to complement the conventional regression-based approach.

1. Introduction

Smoking is the most preventable cause of tobacco-related mortality and morbidity (WHO, 2017). While many smokers want to quit, few succeed in their attempts because of tobacco addiction (Borland et al., 2012). Quitting is difficult for many smokers and relapse to smoking

after either aided or unaided cessation is a common outcome (Garcia-Rodriguez et al., 2013; Fiore et al., 2008). To date, understanding of the determinants of smoking cessation both in terms of initiation and its maintenance has been largely derived from work using a regression approach. This research has provided useful insights by identifying independent predictors of quit attempts and quit success among those

---

* Corresponding author at: 1 Gheringhap Street, Geelong, Victoria 3220, Australia.
E-mail address: hua.yong@deakin.edu.au (H.-H. Yong).

who tried (Borland et al., 2010; Hyland et al., 2006;15; Li et al., 2010; Li et al., 2011). It is well established that the predictors of quit attempts are not necessarily the same as those that predict quit success (Vangeli et al., 2011). Consistent with psychological theories of behaviour change (Bandura, 1977; Rogers et al., 1997; Ajzen, 1991), motivational factors such as desire to quit, concerns about the negative health impact of smoking, quitting self-efficacy and quit intentions are strong predictors of smoking cessation initiation but not its maintenance among those who quit (Borland et al., 2010). Smokers' intention to quit is the most proximal predictor of smoking cessation initiation with the effects of the other factors such as health concerns related to smoking mediated through quit intentions (Yong et al., 2014). These factors are likely to interact in complex ways to determine initiation of a quit attempt. However, our understanding of the complex interactions among the various predictors of smoking cessation has been limited by the analytic approach taken. Much of the research in this field has been done using a regression-based approach which does not lend itself easily to looking at interactive effects between predictive variables beyond a two-way interaction. Higher-order interaction effects are difficult to model and interpret using such an approach. In addition, regression analysis tends to be insensitive to variables that permit accurate prediction for a relatively small subgroup of smokers (Piper et al., 2011). For this purpose, a different approach such as decision-tree analysis is better suited and can complement regression to provide additional insights that may otherwise be missed.

A decision tree is a data analytic technique that allows exploration of the presence of potentially complicated interactions within the data by creating binary segmentations of individuals into sub-groups (Breiman et al., 1984; Tuffery, 2011). Membership of a subgroup is derived based on response to a set of measured/predictive variables. Individuals who are initially grouped together on the basis of similar scores on one predictor can be later separated into distinct subgroups on the basis of further predictors. This greater level of resolution is akin to looking at higher order interactions than a simple two-way interaction.

Decision-tree models have been successfully applied in public health and health behaviour research as a method for identifying subgroups based on their decision-making processes (Piper et al., 2011; Lemon et al., 2003). For example, Piper et al (Piper et al., 2011) employed this method to help identify subgroups of smokers who were at risk of smoking relapse whereas Fuller-Tyszkiewicz (Fuller-Tyszkiewicz et al., 2016) used similar method to identify women subgroups who were at risk of excessive weight gain during pregnancy.

While linear-based models like logistic regression allow the simultaneous explanation of outcome using multiple explanatory variables, their goal is to understand how these variables are independently related to the outcome variable of interest in a linear fashion based on a single prediction equation that works for the whole sample. When a dataset has a large number of explanatory variables that interact in a complicated manner, building a global linear model is difficult. For example, in order to model interactions in logistic regression analysis, one would need to know *a priori* which set of variables interact, appropriate cut-points to use for converting a continuous variable into a categorical form, how to correctly interpret a higher-order interaction effect, and an easy way to communicate a complex set of results. By contrast, decision-tree models allow exploration of complex interactions in the data easily by creating segmentations or sub-groups of data based on complex interactions of the relevant predictor variables, and the intuitive nature of the output lends itself easily to meaningful interpretation.

Identifying subgroups of smokers with high versus low probability of smoking cessation is extremely valuable as this would allow the deployment of limited resources to help those most in need of intervention to motivate them to quit and also increase their chances of success. Insights gained from those with high probability of smoking cessation can be used to inform interventions targeting those with low

probability of smoking cessation.

Evidence to date suggests that smokers from disadvantaged backgrounds may have greater difficulty in quitting because of the more favourable smoking environment (Hitchman et al., 2014) and greater level of nicotine dependence (Siahpush et al., 2006;15). Despite this, some individuals do succeed and it is important to understand the critical success factors/conditions for this subgroup to help inform the design of more effective interventions.

This study makes use of decision-tree models to help identify smoker subgroups and the conditions under which they are more or less likely to stop smoking. As a proof of concept, this study only focused on smoking cessation initiation as an outcome of interest and will explore smoking cessation maintenance in future studies.

The specific aims of this proof-of-concept study are to employ a decision-tree model to help: (WHO, 2017) identify modifiable characteristics of smoker subgroups with high versus low probability of smoking cessation initiation; and (Borland et al., 2012) determine whether and how these characteristics might differ between low, medium and high socio-economic status subgroups.

2. Methods

2.1. Sample, study design and procedures

This study is a secondary analysis of survey data from the Australian arm of the International Tobacco Control Four-Country project, a longitudinal cohort study of adult smokers surveyed that began in 2002 with approximately annual follow-up. Data from current smokers who participated in the 2005 survey and provided outcome data at the 2006 survey were used for initial calibration and validation work (n = 1475) and the models developed were further cross-validated using the 2012–2013 data (n = 787). We chose data from 2005 to 2006 surveys for developing our models because one of the survey questions of interest was only asked from 2005 survey onwards. We chose 2012–2013 survey data for cross-validation to determine the performance of our models over time, particularly in the context of the presence of vaping products from around 2008 onwards (Ayers et al., 2011) and their availability for smoking cessation. We anticipated minimal impact on our model performance since use of these products is low in Australia as retail sale, import and personal possession/use of nicotine vaping products is illegal without a permit or a doctor's prescription (Yong et al., 2017; Yong et al., 2015). The study has received ethical approval from the relevant ethics committee at the Cancer Council Victoria and Deakin University, Australia (#2018-346).

2.2. Measures

Respondents were asked questions about their attitudes and behaviour related to smoking and quitting including socio-demographics such as age, gender, annual household income divided into three strata (low < $30,000, medium from $30,000 to $59,999, and high $60,000 and over), and education (low = primary or some high school, medium = completed high school, technical or trade school, and high = at least some university). Socio-economic status (SES) index was derived based on reported income and education into low, medium and high as follows: if high on both variables, coded as high and low on both coded as low with all other combinations coded as medium, or where income was not disclosed, education alone was used for deriving the index.

2.2.1. Outcome variable

The main outcome of interest was making any quit attempt since the last survey date using the question: "*Have you made any attempts to stop smoking since we last talked with you?*" with Yes/No as response options.

2.2.1.1. Predictor variables. Baseline predictors of interest (all shown to
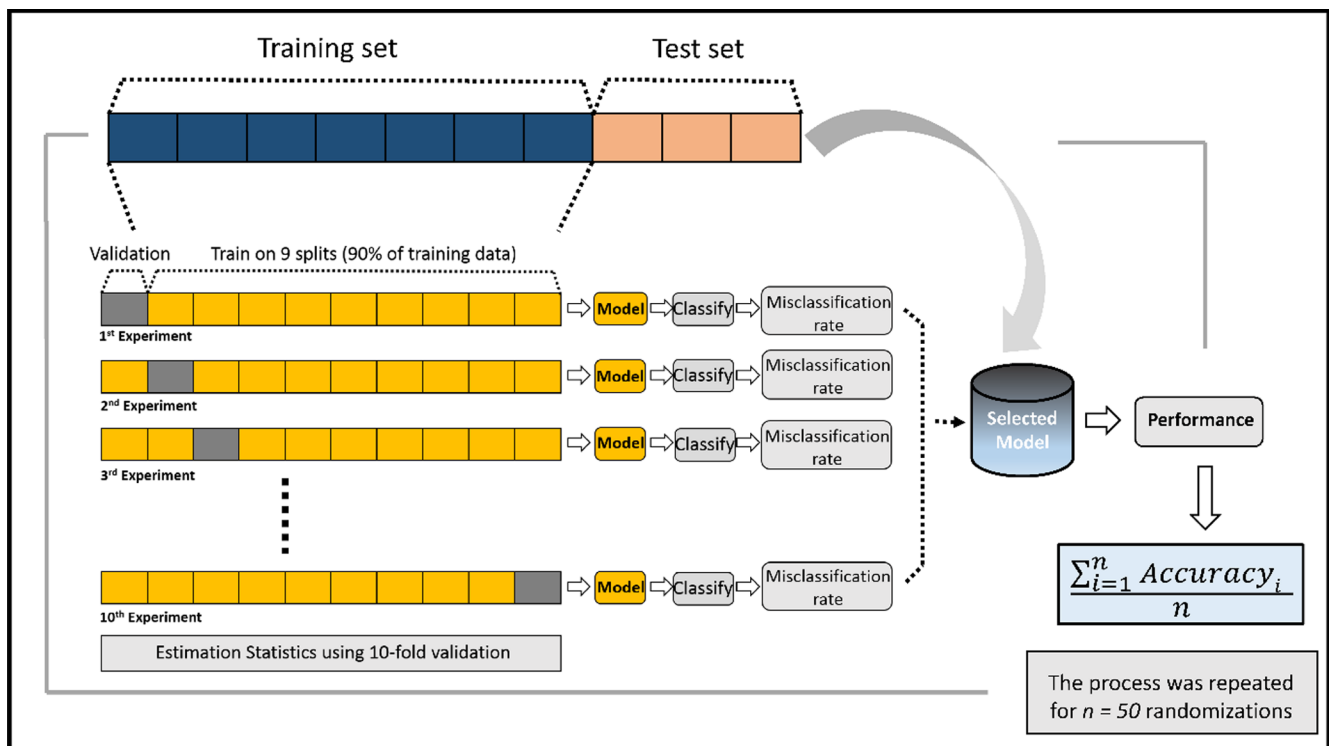
Fig. 1. The model selection and evaluation procedure for the 2005–2006 survey data.

have predictive utility in past research (Borland et al., 2010; Hyland et al., 2006;15; Li et al., 2010) were as follows:

Concerns about the negative impact of smoking [WO, worry] were derived based on averaging of the responses to the following two questions: "*How worried are you, if at all, that smoking will damage your health in the future?*" and "*How worried are you, if at all, that smoking will lower your quality of life in the future?*" with response options: *Not at all/A little/Moderately/Very worried.*

Expressed desire to quit [WA, want to quit] was assessed using the question: "How much do you want to quit smoking?" with response options: Not at all/A little/Somewhat/A lot.

Quit intentions [PL, plans to quit] were assessed and derived as a 5-point ordinal variable using two questions: "Are you planning to quit smoking…within the next month; within the next 6 months; sometime in the future beyond 6 months; or are you not planning to quit?" Those who planned to quit in the next month were asked whether they had set a firm date to quit as an additional category.

Confidence to quit [SE, quitting self-efficacy] was assessed using the question: "If you decided to give up smoking completely in the next 6 months, how sure are you that you would succeed?" with response options: Not at all/Slightly/Moderately/Very/Extremely sure.

2.3. Data classification analysis

All statistical and classification analyses were conducted using Matlab 2017. Effect of the four predictor variables on quit attempts was evaluated using decision tree classifiers (DTCs) designed for a supervised framework. In supervised frameworks, a set of predictor variables are used to predict the category of the outcome or target variable. DTCs employ a top-down greedy search algorithm that partitions the data by comparing the chi-square statistics in relation to the outcome (Breiman et al., 1984).

For the classification task, the sample as a whole (the parent node) was split into two subgroups (child nodes) by using a split predictor

which best discriminated people into two classes. Gini diversity index was computed to derive the Gini improvement measure for use as a split criterion, and the "curvature test" was used to select the best split predictor at each node as it accounts for prediction interactions and allows identification of important predictors in the presence of irrelevant predictors (Loh and Shih, 1997). The chosen predictors represent optimal discrimination between the classes in that subgroup.

The decision trees were trained using the Breiman et al. (Breiman et al., 1984) algorithm where, after the initial growing of the tree, a cross-validation procedure was followed to prevent overfitting. This involved randomly splitting the dataset into a training set (70%) and a test set (30%). The training set was used to select the best decision tree model using a 10-fold cross-validation procedure (see Fig. 1) where 90% of the training set (9–folds) were used for training a model (i.e., the learning set) and the remaining 10% (1–fold) used for validation (i.e., the validation set). The learning set was used to grow a large tree representing pure terminal nodes. A post-pruning strategy ensured that the effects of all predictor variables were considered. After pruning, the validation set in every fold was used to find the misclassification rate of the pruned tree model. From the 10 trained tree models, the final tree model selected was the one with the lowest probability of misclassification. This model was used on the test set, an independent sample, to evaluate its performance accuracy. To further improve the generalisability of the model, the whole process of model selection and evaluation was repeated using 50 randomisations and the results were averaged to obtain the final prediction accuracy of the model.

To evaluate whether the decision-making processes differed by SES, we conducted the classification task stratified by low, medium and high SES strata, and compared the decision tree outputs of these three SES strata.

As an additional validation process, the tree models developed using the 2005–2006 data were tested on the 2012–2013 data to assess their prediction performance.

Table 1
Sample characteristics.

Variables	Low SES	Moderate SES	High SES	Full sample
Initial calibration and validation sample from 2005 to 2006 surveys				
N	763	419	293	1475
N (after filtering missing data)	738	413	291	1442
N (70% training set)	517	290	204	1011
N (30% test set)	221	123	87	431
N used for building tree (learning set/90% of training set)	466	261	184	911
Validation sample from 2012 to 2013 surveys				
N	351	225	211	787
N (after filtering missing data & used as test set)	318	214	194	726
Socio-demographics				
Age group in years - (%)				
2005–2006 surveys	9.17	9.8	8.5	9.2
18-24	27.7	37.0	33.5	31.5
25-39	38.0	40.6	43.7	39.9
40-54	25.2	12.7	14.3	19.5
55 +				
2012–2013 surveys	3.7	1.3	2.8	2.8
18-24	11.7	25.8	18.5	17.5
25-39	36.2	48.0	48.8	42.9
40-54	48.4	24.9	29.9	36.8
55 +				
Sex - % male				
2005-2006 surveys	40.5	51.3	46.8	44.8
2012-2013 surveys	40.2	47.6	54.0	45.9
Predictors				
Quit self-efficacy [SE] -				
% at least slightly sure	64.6	69.9	75.4	68.3
2005-2006 surveys	50.4	57.1	60.3	55.0
2012-2013 surveys				
Plan to quit [PL] - %				
Yes	72.8	74.0	76.7	73.9
2005-2006 surveys	69.3	76.9	81.3	74.8
2012-2013 surveys				
Want to quit [WA] - %				
Yes	84.5	86.5	88.3	85.9
2005-2006 surveys	81.4	86.4	91.4	85.7
2012-2013 surveys				
Worries re smoking impact [WO] - Mean (SD)	2.55 (0.99) 2.45 (0.95)	2.60 (0.87) 2.58 (0.94)	2.66 (0.97) 2.67 (0.95)	2.59 (0.96) 2.55 (0.95)
2005–2006 surveys				
2012-2013 surveys				
Outcome				
% Made at least one quit attempt since last survey date	39.5	44.4	48.8	42.7
2005-2006 surveys	50.1	54.7	60.7	54.3
2012-2013 surveys				

Note: SES, socioeconomic status; SD, standard deviation.

3. Results

3.1. Sample characteristics

Table 1 presents characteristics of the overall sample and by SES (low, medium and high) for the 2005–2006 survey data and the 2012–2013 survey data. About 43% of the 2006 respondents and 54% of the 2013 respondents reported making at least one quit attempt since the last survey wave with the proportion doing so significantly higher among the high SES strata as compared to the low strata (i.e., 49% vs 40% and 61% vs 50%, respectively, p < .001 for both).

3.2. Decision tree analysis

The classification task using 2005–2006 data revealed that all four predictor variables were able to predict group membership (quit attempt status) better than chance levels with overall prediction accuracy of 64% (62%, 66% and 65% for low, medium and high SES strata, respectively). The optimal decision tree for each SES stratum is presented in Fig. 2. For all three SES strata, plan to quit was the first predictor selected for discriminating the sample into subgroups at the first node, but the predictor selected for splitting at the second and subsequent nodes differed for the three SES strata, resulting in different patterns of the branches within each decision tree. Of particular note is that the tree for each stratum contains different numbers of terminal nodes. When cross-validated on the 2012–2013 data, the decision tree models developed on the 2005–2006 data performed slightly better with prediction accuracy of 0.68, 0.69 and 0.71 for low, medium and high SES strata, respectively. Other model performance indices are presented in Table A1.

3.2.1. Low SES stratum

Low SES smokers were more likely to make quit attempts if they planned to quit within the next month or had a firm date to quit (Fig. 2A Group 1). They were also more likely to make attempts (i.e., Group 2 although no cases were found in the 2012–2013 cross-validation sample) if they reported wanting to quit at least somewhat and were extremely sure of quitting. Among the subgroup who were not "extremely sure" of quitting successfully, more made quit attempts (i.e., Group 3) if they had high concerns about the health effects of smoking (i.e., WO score of 2.75 or more). Among those with less than 2.75 score on WO, more made quit attempts (i.e., Group 4) if they were moderately or very sure of quitting successfully. It is notable that not all who expressed a commitment to quit at least within the next six months ended up belonging to the quit attempt group. Among those with plans to quit (within 6 months), two groups (Groups 5 and 6) were notably less likely to make attempts. However, Group 5 was unstable and became more likely to make attempts in the cross-validation. Low SES smokers who had no plan to quit or planned to do so beyond 6 months were generally less likely to make a quit attempt regardless of the pattern of interaction with other predictor variables with one exception. Group 7 had a higher quit attempt rate than the average rate of the total low SES sample (62% vs 40%) and this subgroup was characterised by low or no plan to quit, but a strong desire to quit and a high level of health concern about the negative effects of smoking. Those who also reported wanting to quit a little or not at all, were particularly unlikely to have made attempts (Group 11).

3.2.2. Medium SES stratum

Smokers from medium SES backgrounds were also more likely to make a quit attempt if they planned to quit at least within the next six months. This was particularly so if they were concerned about the health effects of smoking (see Fig. 2B Group 1). There was a potentially interesting interaction as a function of self-efficacy between those moderately and those less concerned about health (i.e., WO score of less than 3.75). Among those with higher self-efficacy, low levels of worry (Group 3) was associated with higher rates of attempts than moderate worry (Group 2), while for those with lower self-efficacy, moderate worry (Group 4) was associated with a much higher attempt rate than low worry (Group 5). This interaction, however, was not evident in the 2012–2013 data. Of note is that even among smokers with no plan to quit or plan to quit beyond 6 months, there were several subgroups (i.e., Groups 6, 7 and 8) being identified with a higher probability of making a quit attempt than that of the total medium SES sample, and two groups (Groups 9 and 10) with low probability, one group (Group 10) was similar as for the low SES stratum.
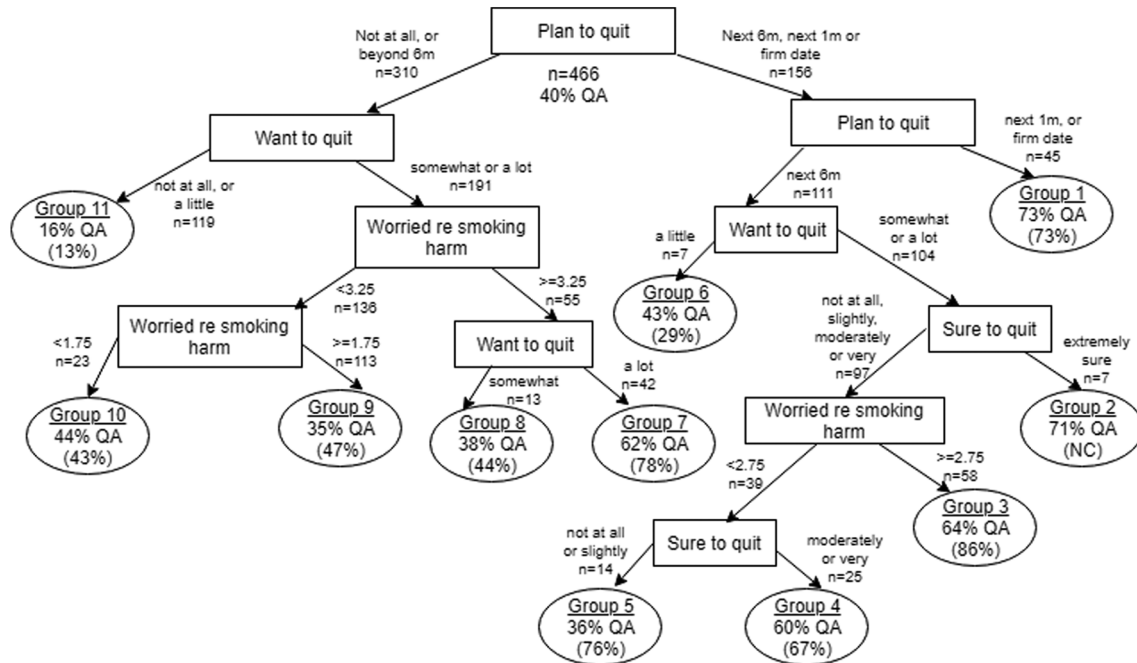
Low SES sample decision-tree diagram

Fig. 2. Tree diagram using four predictor variables, for low, medium and high SES groups (2005–2006 survey data). Note: QA, quit attempt; NC, no cases; percentages making QA in parentheses for the terminal groups are based on 2012–2013 survey data.

3.2.3. High SES stratum

For smokers from high SES stratum, planning was again the main predictor. Those who wanted to quit a lot and had planned to quit in the next month or had a firm quit date were generally more likely to make a quit attempt (see Fig. 2C Group 1). Among the rest, there was considerable diversity. Group 4 who reported only somewhat or less wanting to quit surprisingly had a very high quit rate. Of particular interest are those who want to quit a lot and plan only in the next 6 months, worry about harms made a huge difference. Those with lower worry (Group 3) had a high quit rate, while those with high worry, surprisingly had a very low rate (Group 2, which was not replicated in the 2012–2013 sample) making them distinct to all others who had plans to quit. Among those with a little or no intention to quit, quit

attempt rates for all subgroups (Groups 5–8) were lower than that of the total high SES sample (24–44% vs 49%) although the rate for Group 6 was not replicated. The no plan, no desire (want) group with low probability of making attempts found in the other two SES strata did not appear in this high SES stratum.

4. Discussion

4.1. Main findings

This study serves as a proof-of-concept and is the first to attempt to employ a decision-tree analytic approach to identify smoker subgroups with high versus low probability of smoking cessation initiation. The
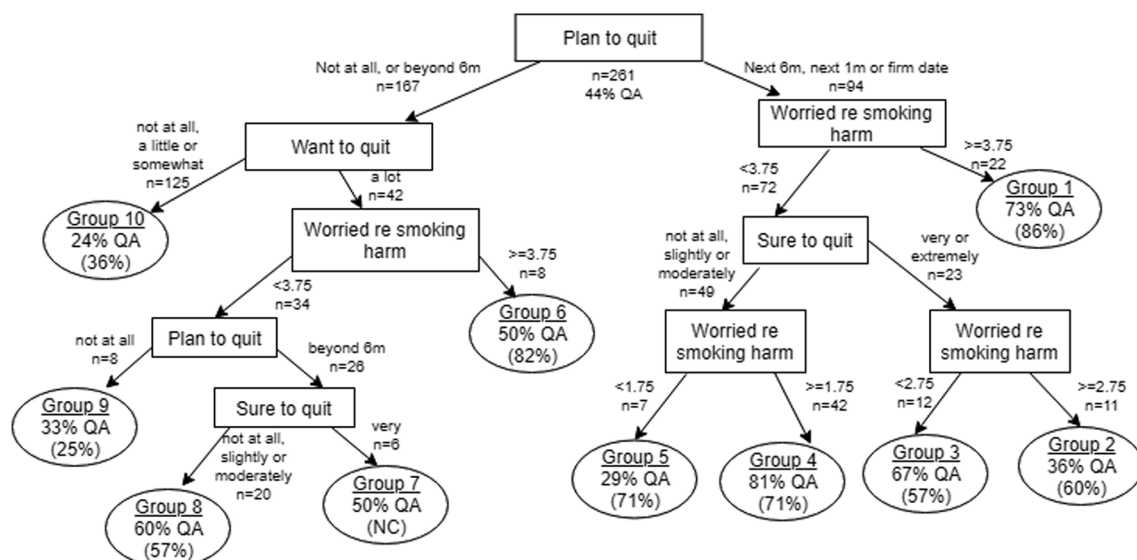
Medium SES sample decision-tree diagram

Fig. 2. (*continued*)

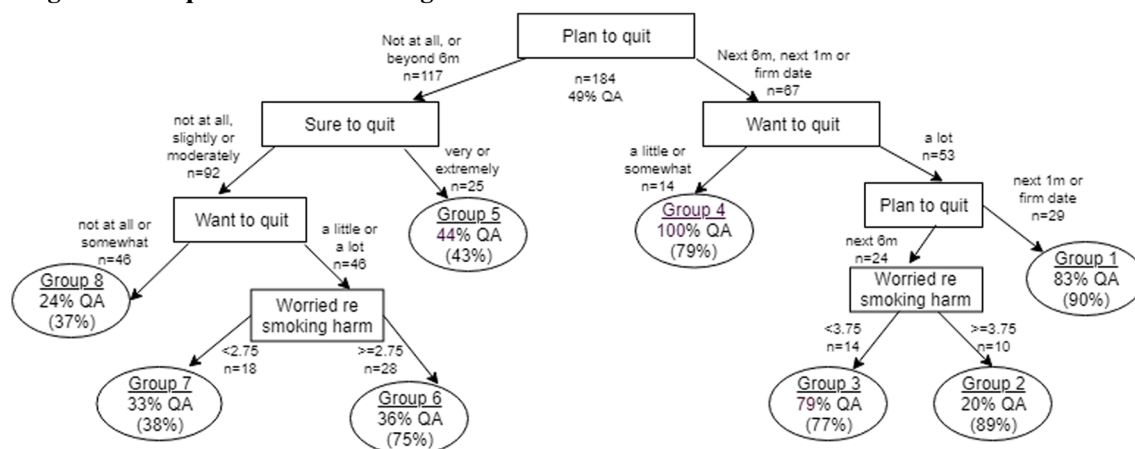**HIgh SES sample decision-tree diagram**

Fig. 2. (*continued*)

Table A1
Model performance indices of decision-tree models trained on 2005–2006 data and tested on the 30% hold-out (validation) data from the 2005–2006 surveys and on new (cross-validation) data from the 2012–2013 surveys.

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	AUC
2005–2006 data								
Low SES	49	42	90	40	0.55	0.68	0.62	0.61
N = 221								
Medium SES	32	20	49	22	0.59	0.71	0.66	0.65
N = 123								
High SES	25	13	32	17	0.60	0.71	0.65	0.65
N = 87								
2012–2013 data								
Low SES	87	29	129	73	0.54	0.82	0.68	0.68
N = 318								
Medium SES	75	25	72	42	0.64	0.74	0.69	0.69
N = 214								
High SES	88	21	50	35	0.72	0.70	0.71	0.71
N = 194								

Note: SES = socio-economic status; TP = true positive; FP = false positive; TN = true negative; FN = false negative; AUC = area under the receiver operating characteristic curve.

results indicated that the four theorized determinants of smoking cessation (quit intentions, wanting to quit, quitting self-efficacy and negative health concern of smoking) can be used for segmenting smokers into subgroups with high versus low probability of smoking cessation initiation and that the segmentation process differs by socio-economic status.

Consistent with past research using a regression-based approach (Borland et al., 2010; Hyland et al., 2006;15; Vangeli et al., 2011), the present study revealed that self-reported quit intentions, wanting to quit, quitting self-efficacy and negative health concern about smoking were all predictive of making a quit attempt with quit intentions (Plan) the most important. However, present findings suggest that these factors do not work in isolation but can interact with each other in meaningful ways to conjointly determine the likelihood of making quit attempts. The complex (higher-order) interactions between variables found here have not been explored in past regression-based studies because of the complexity of conducting such analysis and the difficulty in interpretation. Thus, like many of the previous studies in public health and health behaviour research that have applied decision-tree methodology (Piper et al., 2011; Fuller-Tyszkiewicz et al., 2016; Coughlin et al., 2018; Frisman et al., 2008), current results demonstrated utility of decision-tree models for understanding the complex interactive effects of a set of known predictors of smoking cessation attempts. The simplicity and intuitive nature of the decision-tree

outputs lend themselves well for meaningful interpretation of the decision-making processes leading to the desirable outcome of interest (Lemon et al., 2003).

The decision-tree models suggest that there are meaningful differences by socio-economic status. While the model for all three SES groups selected planning to quit as the first, and most important, predictor, the pattern of predictor selection for subsequent segmentations differed between the low, medium and high SES strata. These results suggest that the decision-making processes of subgroups defined by SES are not necessarily the same with the likelihood of smoking cessation initiation dependent on the presence and interaction of different patterns of cognitive and affective factors. For example, we found no major role for self-efficacy in the high SES stratum. This finding is consistent with evidence indicating that smokers from low SES group face more barriers to quitting than their higher SES counterparts as they tend to be less confident in being able to quit smoking, have lower interest in quitting, and are more addicted to smoking (Siahpush et al., 2006;15; Yong et al., 2013).

The results of the decision-tree models also revealed that the relationship between the predictors and the outcome is not necessarily linear. For example, quit attempts can occur even among a subgroup of smokers with no expressed intention of quitting and, similarly, quit attempts can be absent even among the subgroup of smokers who had expressed high intention to quit. Smokers with no stated intention to quit smoking may still, under certain circumstances, have a greater likelihood of initiating a quit attempt. For example, a low SES smoker with no intention to quit their smoking habits may be more likely to make a quit attempt if they have a strong desire to quit and are highly worried about the negative health impact of their smoking. By contrast, a low SES smoker who has made a commitment to quit within the next six months may not follow through with it if they have low desire to quit. In fact, these two effects were even stronger in the cross-validation sample confirming their robustness. The finding of cases counter to what one might expect underscores the value of using a decision-tree approach and provides support for its use to provide additional insights that are often missed by the use of the conventional linear-based approach.

Also notable is the predictive performance of our models appears robust when tested on a new sample which had a much higher quit attempt rate, possibly due to the availability and trial of new vaping products as a quitting method in recent years (Yong et al., 2015). Nevertheless, our decision-tree models yielded only a modest overall prediction accuracy. Thus, caution should be exercised when interpreting findings of subgroups with small sample size as their effects may be unstable and difficult to replicate. For example, for the high SES stratum, the low quit attempt rate of Group 2 failed to be replicated and

became high in the cross-validation sample (20% vs 89%). For the low SES stratum, no cases were identified for Group 2 in the new sample. The reason for these discrepant findings is unclear but may reflect either an artefact of the small sample size, noise in survey responses or a combination of the two. The influence of vaping products to assist with quitting in the 2012–2013 study period could not be ruled out either although use of these products was likely low given that vaping is illegal in Australia (Yong et al., 2015). Future replication studies are needed to explore this further and also to determine whether using a larger sample size could help to improve performance.

The findings from the present study can be useful for the purpose of translation into practice. From the decision tree, subgroups with low and high probability of smoking cessation attempts can be easily identified based on the decision-making criteria, which may aide decision-making and/or the development of relevant intervention. Data from the decision tree can be used to derive empirically-based algorithms for an app that can help to identify persons at risk (i.e., those with a low tendency to quit) who may benefit from targeted intervention.

4.2. Limitations

This study has several noteworthy limitations. First, it is not possible to conduct hypothesis testing as decision tree output is empirically-driven rather than theory-driven; unlike conventional regression analysis whereby interaction effects can be tested for statistical significance, the interactions represented in decision trees cannot be directly tested. Second, quit attempt was assessed as any attempt since last survey date which could have contributed to misclassification as brief attempts especially those lasting less than 24 h might have been under-reported due to memory bias and/or not being considered as serious attempts. Third, the performance of our models beyond 2013 is unclear and warrants further study. Fourth, smoking cessation maintenance as an outcome was not examined.

4.3. Conclusions

Decision-tree analysis reveals that modifiable characteristics

(markers) of smoker subgroups with high versus low probability of smoking cessation attempts can be identified empirically based on the complex interaction between different theorized determinants of smoking cessation initiation and that the specific characteristics/markers may differ by socio-economic status of smokers. These findings underscore the added value of using this approach for understanding the determinants of smoking cessation initiation in addition to conventional regression-based approach.

5. Contributors

HHY, CK, RB and JY contributed to the study's concept and design. CK and SK conducted the statistical analysis. HHY wrote the first draft of the manuscript and all authors contributed to and have approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

True positives (TP): These are cases in which the model predicted yes (they have attempted to quit smoking) and actual case is also yes.
True negatives (TN): The model predicted no, and they have not attempted to quit smoking.
False positives (FP): The model predicted yes, but they have not actually attempted to quit smoking.
False negatives (FN): The model predicted no, but they actually have attempted to quit smoking.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.addbeh.2019.106258.

References

WHO, 2017. WHO report on the global tobacco epidemic.
Borland, R., Partos, T. R., Yong, H. H., Cummings, K. M., & Hyland, A. (2012). How much unsuccessful quitting activity is going on among adult smokers? Data from the International Tobacco Control Four Country cohort survey. *Addiction, 107*(3), 673–682.
Garcia-Rodriguez, O., Secades-Villa, R., Florez-Salamanca, L., Okuda, M., Liu, S. M., &

Blanco, C. (2013). Probability and predictors of relapse to smoking: Results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). *Drug and Alcohol Dependence, 132*(3), 479–485.
Fiore, M. C., Jaen, C. R., Baker, T. B., Bailey, W. C., Benowitz, N., Curry, S. J., et al. (2008). *Treating tobacco use and dependence: 2008 update.* Rockville, MD: U.S. Department of Health and Human Services, U.S. Public Health Service.
Borland, R., Yong, H. H., Balmford, J., Cooper, J., Cummings, K. M., O'Connor, R. J., et al. (2010). Motivational factors predict quit attempts but not maintenance of smoking cessation: Findings from the International Tobacco Control Four country project.

Nicotine & Tobacco Research, 12(Suppl), S4–S11.

Hyland, A., Borland, R., Li, Q., Yong, H. H., McNeill, A., Fong, G. T., et al. (2006). Individual-level predictors of cessation behaviours among participants in the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control, 15*(Suppl 3), iii83–iii94.

Li, L., Borland, R., Yong, H. H., Fong, G. T., Bansal-Travers, M., Quah, A. C., et al. (2010). Predictors of smoking cessation among adult smokers in Malaysia and Thailand: Findings from the International Tobacco Control Southeast Asia Survey. *Nicotine & Tobacco Research, 12*(Suppl), S34–S44.

Li, L., Feng, G., Jiang, Y., Yong, H. H., Borland, R., & Fong, G. T. (2011). Prospective predictors of quitting behaviours among adult smokers in six cities in China: Findings from the International Tobacco Control (ITC) China Survey. *Addiction. 106*(7), 1335–1345.

Vangeli, E., Stapleton, J., Smit, E. S., Borland, R., & West, R. (2011). Predictors of attempts to stop smoking and their success in adult general population samples: A systematic review. *Addiction, 106*(12), 2110–2121.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.

Rogers, R. W., & Prentice-Dunn, S. (1997). Protection motivation theory. In D. S. Gochman (Ed.). *Handbook of health behavior research 1: Personal and social determinants* (pp. 113–132). New York, NY, US: Plenum Press.

Ajzen, I. (1991). The theory of planned behavior. *Organization Behavior and Human Decision, 50*(2), 179–211.

Yong, H. H., Borland, R., Thrasher, J. F., Thompson, M. E., Nagelhout, G. E., Fong, G. T., et al. (2014). Mediational pathways of the impact of cigarette warning labels on quit attempts. *Health Psychology, 33*(11), 1410–1420.

Piper, M. E., Loh, W. Y., Smith, S. S., Japuntich, S. J., & Baker, T. B. (2011). Using decision tree analysis to identify risk factors for relapse to smoking. *Substance Use & Misuse, 46*(4), 492–510.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (2nd ed.). Pacific Grove, CA: Wadsworth.

Tuffery S. Data mining and statistics for decision making. Giudici P, Givens GH, Mallick BK, editors. UK: Wiley; 2011.

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine, 26*(3), 172–181.

Fuller-Tyszkiewicz, M., Skouteris, H., Hill, B., Teede, H., & McPhie, S. (2016). Classification tree analysis of postal questionnaire data to identify risk of excessive gestational weight gain. *Midwifery. 32*, 38–44.

Hitchman, S. C., Fong, G. T., Zanna, M. P., Thrasher, J. F., Chung-Hall, J., & Siahpush, M. (2014). Socioeconomic status and smokers' number of smoking friends: Findings from the International Tobacco Control (ITC) Four Country Survey. *Drug and Alcohol Dependence, 143*, 158–166.

Siahpush, M., McNeill, A., Borland, R., & Fong, G. T. (2006). Socioeconomic variations in nicotine dependence, self-efficacy, and intention to quit across four countries: Findings from the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control, 15*(Suppl 3), iii71–iii75.

Ayers, J. W., Ribisl, K. M., & Brownstein, J. S. (2011). Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance. *American Journal of Preventive Medicine, 40*(4), 448–453.

Yong, H. H., Hitchman, S. C., Cummings, K. M., Borland, R., Gravely, S. M. L., McNeill, A., et al. (2017). Does the regulatory environment for E-cigarettes influence the effectiveness of E-cigarettes for smoking cessation? Longitudinal findings from the ITC four country survey. *Nicotine & Tobacco Research, 19*(11), 1268–1276.

Yong, H. H., Borland, R., Balmford, J., McNeill, A., Hitchman, S., Driezen, P., et al. (2015). Trends in E-cigarette awareness, trial, and use under the different regulatory environments of Australia and the United Kingdom. *Nicotine & Tobacco Research, 17*(10), 1203–1211.

Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica, 7*, 815–840.

Coughlin, L. N., Tegge, A. N., Sheffer, C. E., & Bickel, W. K. (2018). A machine-learning approach to predicting smoking cessation treatment outcomes. *Nicotine & Tobacco Research*.

Frisman, L., Prendergast, M., Lin, H. J., Rodis, E., & Greenwell, L. (2008). Applying classification and regression tree analysis to identify prisoners with high HIV risk behaviors. *Journal of Psychoactive Drugs, 40*(4), 447–458.

Yong, H. H., Siahpush, M., Borland, R., Li, L., O'Connor, R. J., Yang, J., et al. (2013). Urban chinese smokers from lower socioeconomic backgrounds face more barriers to quitting: Results from the international tobacco control-China survey. *Nicotine & Tobacco Research, 15*(6), 1044–1051.