Supplementary

Appendix 1

The melanoma skin cancer staging analysis plan

Step 1: analytic dataset preparation

(I) Prepare the primary analytic dataset by merging three data sources: Alberta Tomorrow Project Health and Lifestyle Questionnaire (HLQ), Canada Diet History Questionnaire-I (CDHQ-I), Past Year Total Physical Activity Questionnaire (PYTPAQ).

(II) Link the above merged data to Alberta Provincial Administration Cancer data and retain the data that contains the records from both sources.

(III) Restrict the data to melanoma skin cancer site and those with exact cancer stage information.

(IV) Check the variable distributions. Exclude the variables that have large percentage of missing, also impute the missing value of continuous variables through mean value imputation and collapse the missing category to the reference category level for the categorical variables.

Step 2: model building process

(I) Perform the descriptive statistics against the different cancer stage levels, using Mood's test of medians for continuous factors, and Chi-square or Fisher's exact test for the categorical factors.

(II) Conduct bivariable covariate screening to select the candidate covariate list of the model building with the significance level of P=0.10.

(III) Build the final model through backward variable elimination process with the significance level of P=0.05.

(IV) Recheck the final model by adding any other covariates suggested by the literature and the Random Forest model through a R package: partykit (v1.2-5).

Step 3: model checking

(I) Check if any significant interaction term exists by adding the interaction term of two potential factors with the criteria of the significance level of P=0.05.

(II) Test the validity of the linear relationship functional term.

(III) Perform the regression diagnostics and check for the influential outliers through the residual plot and the DFBETAS statistic for each case.

(IV) Re-visit the goodness-of-fit model overall fitting.

List of variables considered

(I) Based on literature search with corresponding variable available from Alberta's Tomorrow Project database: education level, socioeconomic status (total family income), employment, age, obesity, family history of melanoma, smoking status, sunburn history, marital status, gender, race, urban *vs.* rural place of residence, physical activity including occupational physical activity and time spent outdoors for recreational activities.

(II) Based on additional lifestyle variables available from Alberta's Tomorrow Project database (https://myatpresearch.ca/data-dictionaries/): health-related (e.g., self-rated health, family history of melanoma skin cancer and any cancer), psychosocial factors (e.g., stress, social support), and health practices (e.g., cancer screening participation, physical activity in recreational, occupational, and household domains), dietary intake (e.g., specific foods such as fruit and vegetable consumption and nutrients), body measurements (e.g., waist and hip circumference)

https://dx.doi.org/10.21037/ace-23-7