ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

Application of Logistic Regression and Random Forests to Assess the Relevance of Chrononutrition Information for Prediction of Overweight in the INRAN-SCAI 2005-2006 Nutrition Survey

Bartoszek Karolina[(1)], *Palla Luigi*[(1)]

(1) Department of Public Health and Infectious Diseases, University of Rome La Sapienza, Italy

CORRESPONDING AUTHOR: *Palla Luigi, luigi.palla@uniroma1.it*

INTRODUCTION

Obesity is a world wide crisis nowadays it contributes to a number of serious diseases like type 2 diabetes[1] and coronary heart disease. Chrononutrition and in particular timing of eating, has been shown to have an influence on obesity in studies with animal models[2]. Previous human epidemiological studies have already investigated the association between timing and regularity of eating with BMI [3,4]. However to date no study specifically assessed the added value of using data on the timing and regularity of calorie intake for predicting obesity, on top of total calorie intake across the day, when this is information is available in nutritional epidemiological studies like surveys with diet diaries.

OBJECTIVES

The target population of this study are Italian adults aged 18 to 64 years old from a past cross sectional nutrition survey INRAN-SCAI[5]. The aim of the study is to compare the performance, in predicting overweight (BMI>25), of models based on 6 day time intervals of calorie intake, to models based on the total calorie intake across the day and comparing prediction performance of logistic regression and random forests using ROC curves.

METHODS

Data collected during the Italian nationally representative food consumption survey INRAN-SCAI in the years 2005-2006 was used to make the predictive models. The data used for our analysis was collected from 2312 adults aged 18-64, in the form of a diet diary over 3 consecutive days as well as sociodemographic questionnaire. The days were divided into 6 time intervals, which correspond to the times of 3 main meals and the intervals between them (6am-9am; 9am-12pm; 12pm-3pm; 3pm-7pm; 7pm-10pm; 10pm-6am).

Three different types of models were compared in this study: (i) models trained on the mean energy intake and irregularity of the 6 time intervals, (ii) models trained on the 6 time intervals but using repeated measures from the 3 days and (iii) models trained using the mean energy intake and irregularity for the whole day.

Logistic regression models were built using the whole data sample as well as separating the training(70%) and testing(30%) set for the model, with overweight (chosen due to the small number of strict obesity cases) as outcome and including as predictors also several available sociodemographic variables (including physical activity) chosen by a preliminary application of a forward variable selection algorithm. Sensitivity, specificity, negative predictive value, positive predictive value and error rate were calculated for two different cut off points, one being 0.5 and the second being the optimal cutoff point maximizing the specificity and sensitivity simultaneously. Cross validated models (10 fold) were also generated. The ROC curves corresponding to all the models were compared

For all 3 conditions random forest models were also generated. Their ROC curves were compared and metrics such as specificity, sensitivity NPV, PPV and error rate were in turn calculated.

RESULTS

When logistic regression models were trained on 100% of the data sample (including 34.6% overweight subjects), the

DOI: 10.54103/2282-0930/29248

models using time intervals repeated measures or means both performed better than the model using a day mean. The AUC for repeated measures was 0.7664, very similar to AUC for interval means (0.7639), the AUC for the mean of the day model was 0.7520. The difference between them was statistically significant both when tested across the whole ROC curve and when tested for optimal cutoff point and 0.5 cutoff point. At the optimal cutoff point specificity, NPV, PPV and error rate performed better in the models using time intervals, whereas sensitivity was higher in the model using day mean. For the 0.5 cutoff the models with intervals performed better in all values except for specificity, which was roughly the same for all models.

After separating the training set and test set, the differences in AUC between the models were no longer statistically significant. Similar result was obtained for cross validated models, the differences in AUCs were slightly larger than when separating the test and training sets by hand but the confidence intervals were overlapping. All the metrics (sensitivity, specificity, PPV, NPV and error rate) for training and test sets for all 3 models were very close, at both mentioned cutoff points, the only one which stood out was sensitivity in the repeated measures at 0.5 cutoff (40.7% vs. 37.1%, 35.5%).

The random forest models also did not show a significant difference between the AUC corresponding to the 3 conditions. The sensitivity for the random forest models was generally low (38% for mean of the day, 32.9% for means of intervals, 36.1% for repeated measures). NPV was much higher for the mean of the day model at 86.3%, compared to around 70% for others while the rest of the metrics, also except for sensitivity, performed better in the models using time intervals. Comparison of sensitivity/specificity/ROC for random forests and 0.5 cut-off of logistic models are shown in Figure 1.

CONCLUSIONS

The study provides an insight into how chrononutritional data based on repeated diet diaries summarized by calories consumed in 6 time intervals (as defined in previous work [3,4]) compares in terms of overweight prediction with using just mean total calorie intake in the day, while also comparing the performance of a machine learning technique like random forest with a classical method like logistic regression. The area under ROC was significantly higher when using time intervals (repeated or averaged) compared to whole day only when they were trained on 100% of the data sample, and this result became nonsignificant after a 30% test set was randomly taken from the sample, both within the training and test sample.

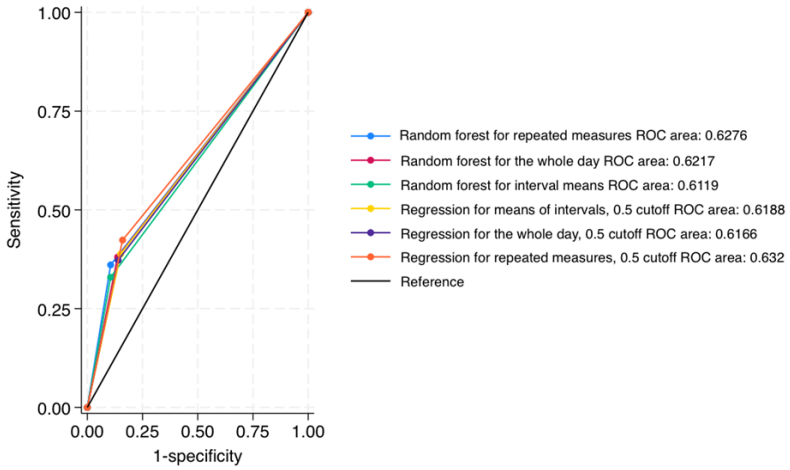
Chrononutrition information in this case allowed for a significantly better prediction of overweight only when testing on the same data as the model was trained on and may thus be attributable to overfitting. Hence the results, despite being based on a nationally representative sample, don't seem to be generalizable for public health purposes. On the other hand, population heterogeneity may also be hindering prediction. Another limitation is the cross-sectional nature of the INRAN-SCAI nutrition survey. In conclusion, timing/regularity of eating may still capture useful information to study overweight/obesity but a properly powered prospective study is warranted to truly assess its potential different impact/relevance also for subgroups of the population.

REFERENCES

1. Klein S, Gastaldelli A, Yki-Järvinen H, Scherer PE. Why does obesity cause diabetes?. Cell Metab. 2022;34(1):11-20
2. Hawley JA, Sassone-Corsi P, Zierath JR. Chrono-nutrition for the Prevention and Treatment of Obesity and Type 2 diabetes: from Mice to Men. Diabetologia. 2020;63. doi:https://doi.org/10.1007/s00125-020-05238-w
3. Palla L, Lopez Sanchez L, Characterising Diurnal and Irregularity Eating Patterns and Their Relationship with Obesity in the Italian Population in the INRAN-SCAI 2005–2006 Nutrition Survey Proceedings 2023, 91(1), 346; https://doi.org/10.3390/proceedings2023091346
4. Di Traglia L, Vestri A, Palla L ,Predicting Obesity via Deep Learning: The Role of Chrononutrition, European Journal of Public Health, 34, Supplement_3, https://doi.org/10.1093/eurpub/ckae144.894
5. Leclercq C, Arcella D, Piccinelli R, et al. The Italian National Food Consumption Survey INRAN-SCAI 2005-06: main results in terms of food consumption. Public Health Nutr. 2009;12(12):2504-2532.

DOI: 10.54103/2282-0930/29248

Figure 1. The area under ROC obtained from random forest and logistic regression models across 3 different set of nutritional intake predictors: the repeated measures across time intervals; the means and irregularities of time intervals; the mean and irregularity of the whole day



DOI: 10.54103/2282-0930/29248