Title

A data mining approach to investigate food groups related to incidence of bladder cancer in the BLadder cancer Epidemiology and Nutritional Determinants International Study

Journal

British Journal Of Nutrition, 124(6)

ISSN

0007-1145

Authors

Yu, Evan YW
Wesselius, Anke
Sinhart, Christoph
et al.

Publication Date

2020-09-28

DOI

10.1017/s0007114520001439

Peer reviewed

HHS Public Access

Author manuscript

Br J Nutr. Author manuscript; available in PMC 2022 August 31.

Published in final edited form as:

Br J Nutr. 2020 September 28; 124(6): 611–619. doi:10.1017/S0007114520001439.

A Data Mining Approach to Investigate Food Groups related to Incidence of Bladder Cancer in the BLadder cancer Epidemiology and Nutritional Determinants International Study

Evan Y.W. Yu[1], **Anke Wesselius**[1,Ψ], **Christoph Sinhart**[2], **Alicja Wolk**[3], **Mariana Carla Stern**[4], **Xuejuan Jiang**[4], **Li Tang**[5], **James Marshall**[5], **Eliane Kellen**[6], **Piet van den Brandt**[7], **Chih-Ming Lu**[8], **Hermann Pohlabeln**[9], **Gunnar Steineck**[10], **Mohamed Farouk Allam**[11], **Margaret R. Karagas**[12], **Carlo La Vecchia**[13], **Stefano Porru**[14,15], **Angela Carta**[15,16], **Klaus Golka**[17], **Kenneth C. Johnson**[18], **Simone Benhamou**[19], **Zuo-Feng Zhang**[20], **Cristina Bosetti**[21], **Jack A. Taylor**[22], **Elisabete Weiderpass**[23], **Eric J. Grant**[24], **Emily White**[25], **Jerry Polesel**[26], **Maurice P.A. Zeegers**[27,28]

[1]Department of Complex Genetics and Epidemiology, School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands.

[2]DKE Scientific staff, Data Science & Knowledge Engineering, Faculty of Science and Engineering.

[3]Division of Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institute, Stockholm, Sweden.

[4]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA.

[5]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA.

[6]Leuven University Centre for Cancer Prevention (LUCK), Leuven, Belgium.

[7]Department of Epidemiology, Schools for Oncology and Developmental Biology and Public Health and Primary Care, Maastricht University Medical Centre, Maastricht, The Netherlands.

[8]Department of Urology, Buddhist Dalin Tzu Chi General Hospital, Dalin Township 62247, Chiayi County, Taiwan.

[9]Leibniz Institute for Prevention Research and Epidemiology-BIPS, Bremen, Germany.

[10]Department of Oncology and Pathology, Division of Clinical Cancer Epidemiology, Karolinska Hospital, Stockholm, Sweden.

Ψ**Corresponding Author:** Anke Wesselius, anke.wesselius@maastrichtuniversity.nl Mailing address: Universiteitssingel 40 (Room C5.570), 6229 ER, Maastricht, the Netherlands. Phone number: +31 6 39014333.

Author contributions

Study conception and design: AW and MPZ; Analyses and interpretation of data: EYY and CS; Drafting of the manuscript: EYY; Revised the manuscript: AW and MPZ; Provided the data and revised the manuscript: AW, MCS, XJ, LT, JM, EK, PvdB, CML, HP, GS, MFA, MRK, CLV, SP, AC, KG, KCJ, SB, ZFZ, CB, JAT, EW, EJG, EW, JP; Approved the manuscript: all authors.

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Conflict of interest

All the authors declare no conflict of interest.

[11]Department of Preventive Medicine and Public Health, Faculty of Medicine, University of Cordoba, Cordoba, Spain.

[12]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA.

[13]Department of Clinical Medicine and Community Health, University of Milan, Milan, Italy.

[14]Department of Diagnostics and Public Health, Section of Occupational Health, University of Verona, Italy.

[15]University Research Center “Integrated Models for Prevention and Protection in Environmental and Occupational Health” MISTRAL, University of Verona, Milano Bicocca and Brescia, Italy.

[16]Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, Italy.

[17]Leibniz Research Centre for Working Environment and Human Factors at TU Dortmund, Dortmund, Germany.

[18]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON, Canada.

[19]INSERM U946, Variabilite Genetique et Maladies Humaines, Fondation Jean Dausset/CEPH, Paris, France.

[20]Departments of Epidemiology, UCLA Center for Environmental Genomics, Fielding School of Public Health, University of California, Los Angeles (UCLA), Los Angeles, CA, USA.

[21]Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri-IRCCS, Milan, Italy.

[22]Epidemiology Branch, and Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA.

[23]International Agency for Research on Cancer (IARC), World Health Organization, Lyon, France.

[24]Department of Epidemiology Radiation Effects Research Foundation, Hiroshima, Japan.

[25]Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

[26]Unit of Cancer Epidemiology, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Italy.

[27]CAPHRI School for Public Health and Primary Care, University of Maastricht, Maastricht, The Netherlands.

[28]School of Cancer Sciences, University of Birmingham, Birmingham, UK.

Abstract

At present, the analysis of diet and bladder cancer (BC) is mostly based on the intake of individual foods. The examination of food combinations provides a scope to deal with the complexity and unpredictability of the diet and aims to overcome the limitations of the study of nutrients and foods in isolation. This article aims to demonstrate the usability of supervised data mining methods to extract the food groups related to BC. In order to derive key food groups associated with BC risk, we applied the data mining technique C5.0 with 10-fold cross validation in the BLadder cancer Epidemiology and Nutritional Determinants (BLEND) study, including data from

18 case-control and 1 nested case-cohort study, compromising 8,320 BC cases out of 31,551 participants. Dietary data, on the 11 main food groups of the Eurocode 2 Core classification codebook and relevant non-diet data (*i.e.* sex, age and smoking status) were available. Primarily, five key food groups were extracted; in order of importance: beverages (non-milk); grains and grain products; vegetables and vegetable products; fats, oils and their products; meats and meat products were associated with BC risk. Since these food groups are corresponded with previously proposed BC related dietary factors, data mining seems to be a promising technique in the field of nutritional epidemiology and deserves further examination.

Keywords

Bladder cancer; Data mining; Food groups; Epidemiological studies

Introduction

Bladder cancer (BC) is the most common malignancy of urinary tract and the seventh cause of mortality for cancer (2.8% of all cancer deaths), with nearly 430,000 new cases and 165,000 deaths per year worldwide [1; 2]. According to Al-Zalabani *et al*, up to 80% of BC can be attributed to lifestyles, including occupation, smoking, exercise and diet [3]. Particularly, it is biologically plausible for dietary factors to influence BC risk considering that beneficial as well as harmful components of a diet are excreted through the urinary tract and in direct contact with the epithelium of the bladder [4]. However, as stated in the report by WCRF/AIRC [5], there is still 'limited' evidence for the role of diet on the BC risk.

Analysis of overall dietary patterns related to BC has been gained a lot of attention during past years [6; 7]. Instead of looking at individual foods or nutrients, analysis of dietary patterns examines the effects of the overall diet, considering the inter-correlations in the consumption of various foods and nutrients. Conceptually, dietary patterns represent a broader picture of food and nutrient consumption, and analysis of dietary patterns may help in better understanding and preventing the development of common cancers.

Several conventional analysis techniques are available for extracting dietary patterns including factor and cluster analyses: investigator-driven methods, such as dietary indices and dietary scores; and data-driven methods, such as principle component analysis. Although these techniques are widely used and might reveal some important information on the relation between dietary patterns and common cancers, they all draw subjective conclusions since they are based on series of priori assumptions, which may differ among researchers. A relatively new approach in the field of nutritional epidemiology is 'data mining'. Data mining is a process that uses a variety of data analysis tools to extract hidden predictive information from large data. This technique is considered to be a powerful technology with great potential to help people focus on the most important information of their data [8]. A previous study in the field of nutritional epidemiology already showed that data mining allowed to define unexpected dietary patterns that might not be recognized using conventional statistical methods [9]. Therefore, in the present study we used this

technique to examine the combinational foods at individual level to extract some food groups related to the BC risk.

Methods

Study Population

The dataset used in the present study is part of the 'BLadder cancer Epidemiology and Nutritional Determinant (BLEND)' study, which aims at assessing the association between diet and the BC risk. Details on the methodology of the BLEND consortium have been described elsewhere [10]. The present study included data of 18 case-control [11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28] and 1 nested case-cohort study (29) providing information on diet and BC, from 12 different countries across the world, including data on 8,320 BC cases and 23,231non-cases within the age range of 18–100 years. Each study ascertained incident bladder cancer, defined to include all urinary bladder neoplasms according to the International Classification of Diseases for Oncology (ICD-O-3 code C67) using population-based cancer registries, health insurance records, or medical records. Each participating study has been approved by the local ethic committee. Informed consent was obtained from all individual participants included in each study. Most of the BC cases were diagnosed and histologically confirmed in 1990s.

Data Collection

All included studies made use of a validated self-administrated food frequency questionnaire (FFQ) or an FFQ administered by a trained interviewer. Homogenization of the dietary data was done by making use of the Eurocode 2 Core classification codebook [30]. This codebook consists of main food groups and their first and second-level subgroups [31]. In order to reduce the variance of individual food items across the world (Supplementary Table 1), foods were attributed into 11 main groups: milk and dairy products (A); eggs and egg products (B); meats and meat products (C); fishes and fish products (D); fats, oils and their products (E); grains and grain products (F); pulses, seeds, kernels, nuts and their products (G); vegetables and vegetable products (H); fruits and fruit products (I); sugars and sugar products (J); beverages (non-milk, K). All food groups were measured as servings of food intake per week and divided into quartile, with Q1-Q4 respectively corresponding to lowest- and highest intake. In addition to information on diet, the BLEND dataset also included data on study characteristics (design, method of dietary assessment, and geographical region), participant demographics [age (continuous), sex (male, female)] and smoking status (never/current/former).

Baseline analysis

Continuous variables were described as mean and standard deviation (SD), and categorical variables as absolute and relative frequencies. Missing values were tested for missing at random (MAR) or missing completely at random (MCAR) [32; 33]. To test for MAR, logistic regression was performed with a missing data indicator created for each variable. No significant relationship between the missingness indicators and the outcome of interest suggests MAR. The assumption that missing data are MCAR were assessed using the Little's MCAR chi-squared test [34; 35].

Data Mining Method

All the 11 main food groups and the non-diet variables (*i.e.* age, sex and smoking status) were selected and entered into data mining procedures.

A classification technique called C5.0 [36], which is a variant of the C4.5 algorithm developed by Ross Quinlan, was used since it can represent solutions as decision trees and as rulesets [37]. It builds a decision tree based on the training/validation sets using the concept of information entropy. The decision tree is built by splitting the data in two parts at the value of one variable that yields the highest normalized information gain. That is, it splits on the value of the chosen variable that separates positive and negative observations (*i.e.* BC status: case and non-case), most efficiently. The pruning severity of the model was set at the default level of 75. This level yielded the lowest complexity (*i.e.* which refers to the minimum number of records in each tree branch to allow a split) with sufficient accuracy. Standard tenfold cross-validation was used in which the entire eligible BLEND dataset was divided into ten approximately equally sized parts. Nine parts were used in turn as training sets and the remaining tenth part was used as the validation set. The validation set (10%) was chosen within the entire dataset according to the distribution of BC status. The participants with missing values were taken into account by using the ratio of the participants with missing values multiplied by the information entropy of the subset of participants without missing values for each variable [38]. The classification C5.0 algorithm was run for the included diet and non-diet variables within the BLEND dataset; meanwhile, variable importance (*i.e.* attribute usage) for C5.0 model was calculated by determining the percentage of training set samples that fall into all the terminal nodes after the split, which defines the variable importance value of each diet and non-diet variables in relation to BC [39; 40; 41; 42]. These importance values range from 0% to 100%, where 0% indicates ‘unimportant’ and 100% indicates ‘extremely important’. Both continuous and categorical variables were included in the models. Node splits in continuous variables can occur at any value and were not predetermined.

Rules were then generated by using the ‘ruleset’ function in C5.0, which transformed the decision tree into specific context associated with BC. A BC status (either case or non-case) was predicted by each rule, and a value between 0% and 100% indicates the confidence of the risk in relation to BC outcome. The overall performance of the C5.0 classifier was evaluated by classification accuracy, true positive rate (TPR), false positive rate (FPR) and receiver operating characteristic (ROC) with the area under the ROC curve (AUC). This is the number of correct classifications of the instances from the validation set divided by the total number of these instances, expressed as a percentage. The greater the classification accuracy, the better is the classifier. A sensitivity analysis was performed by categorizing age in to six groups (years): ≤55, 55–60, 60–65, 65–70, 70–75, >75, based on same data mining procedure.

All data analyses were performed with R software version 3.5.1 (using packages ‘C5.0’ and ‘caret’ developed by Max Kuhn; ‘rpart’ developed by Beth Atkison; ‘ROCR’ developed by Tobias Sing and Oliver Sander).

Results

Baseline analyses of the included data

The characteristics of the BLEND participants are presented in Table 1. In total 31,551 participants are included in the analyses, of which 8,320 (26.37%) were BC cases. The mean age of non-cases (59 years old) was lower than cases (62 years old), and most of the participants were Caucasian (92.27%). Around 66.68% of participants were smokers, with 33.62% of those being current smokers and 33.06% being former smokers.

Significant results of logistic regression for food-group variables indicated that missing dietary data was not MAR (all $P_{MAR} < 0.05$). Little's test also provided evidence against the assumption that missing data were MCAR (all $P_{MCAR} < 0.001$). Rejection of both MAR and MCAR indicates the missing values are missing not at random (MNAR). Therefore, the observations with missing data could not be deleted, and the missing values were marked as blank and not replaced by any value.

Extraction of food groups in relation to BC via the data mining procedure

Figure 1 presents an example of a decision tree with three different variables. The variables are ranked according to how they were used to split the participants from decision nodes to end nodes. A position of 1 (A) corresponds to the variable that in all trees is the first variable used to split; a position of 2 (B) corresponds to the variable that on average is the second variable used to spit, and so on till finally the all participants were split into BC cases and non-cases. 'Sex' is on the first rank split of the tree, which indicates dietary patterns are differentiated in males and females related to BC. Both non-diet variables (age, sex and smoking status) and five food groups (C, E, F, H, K) were identified as having an influence on development of BC. The observed importance values of these variables are (Figure 2): sex (100%); smoking status (74.60%); age (62.80%); beverages (55.81%); grains and grain products (37.98%); vegetables and vegetable products (24.30%); fats, oils and their products (2.95%); meats and meat products (2.71%). Other input variables showed to have an importance value of 0% and were, therefore, considered non-relevant for BC development. The overall classification accuracy is 75.10%, with TPR 0.86 and FPR 0.31 (the ROC curves, with AUCs from 0.690–0.701, for each cross-validation run were performed in Supplementary Figure 1).

Table 2 presents the extracted eight rules resulting into BC outcome after application of the 'ruleset' classifier of C5.0, with a classification accuracy of 74.90%. The results from 'ruleset' show that the variables identified by the 'decision tree' approach are also identified by using the 'ruleset' approach. Here we see that current/former male smokers tended to be BC cases and never male smokers tended to be non-BC cases. However, to be able to split the participants into case or non-case is depending on their dietary habits. Females show relatively simple rules, in which only 'grain and grain products' and 'beverages (non-milk)' were identified to be related to BC.

A sensitivity analysis by transforming age into categorial variable was performed based on C5.0 algorithm, the results show similar with identification of same food groups related to BC (Supplementary Figure 2).

Discussion

To our knowledge, this is among the first studies to apply the data mining approach to extract food groups associated with BC risk based on the complexity of the combinational food intake. By applying C5.0 algorithm, the decision tree and rules derived from this approach showed that sex, smoking status, age and five food groups [C: meats and meat products, E: fats, oils and their products, F: grains and grain products, H: vegetables and vegetable products, K: beverages (non-milk)] are in relation with BC risk in both males and females. Apart from the well-established factors (*e.g.* age, sex and smoking) for BC identified in the data mining procedures, the association of diet, especially specific dietary pattern, with BC risk deserves to be explored due to the limited evidence on this topic, and because it reflects a person's dietary exposure in aggregate rather than in isolation.

Although the use of data mining is relatively new for unravelling diet in relation to the cancer risk, previous studies already examined dietary intake with BC risk using other techniques. In 2008, De Stefani *et al.* [43] found that the dietary patterns labelled as 'sweet beverages' (high loadings of coffee, tea, and added sugar) and 'Western' (high loadings of red meat, fried eggs, potatoes, and red wine) were directly associated with the risk of BC based on factor analysis. In addition, the negative influence of the Western diet was also observed for BC recurrence: BC patients in the highest tertile of adherence to a Western dietary pattern had a 48% higher risk of recurrence of BC compared to patients in the lowest tertile [6]. The Western diet is especially low in fresh fruits and vegetables, but generally high in saturated fats and red and processed meats. Results from the present study are in line with these results, with respect to high of fat being associated with an increased risk for the development of BC and high intake of vegetables and vegetable products being associated with a reduced risk.

Previous studies on single food item or food groups in relation to BC risk, also reported that high intake of vegetables was associated with reduced risk of BC [44; 45; 46; 47]. These studies suggest that the preventive effect could possibly be due to the antioxidant action of vegetables [48; 49] and that each serving of vegetable may result in a 10% risk decline. Although very powerful, results from the present study only identify 'vegetables and vegetable products' as a possible main food group related to BC risk. It remains unclear which specific subgroup is responsible (*e.g.* starchy/non-starchy, processed/fresh, citrus/cruciferous). Detailed analyses of BLEND data may help to elucidate this uncertainty.

Limited evidence is available on the influence of 'grains and grain products' on BC risk. However, our findings are in line with results from a previously conducted case-control [50], suggesting that a high intake of whole grains may reduce the risk of BC. In contrast, a more recent study found that BC risk was negatively influenced by a high intake of refined carbohydrate foods [51]. Thus, future detailed analyses, especially those focusing on whole grains and refined grain products, may be useful. Of note, our results on grain products might have been influenced by the fact that the 'grain and grain products' group of the present study included sweet 'Fine bakery wares' such as 'Sweet biscuits and cookies' which are high in sugar and thereby promote obesity, known to be a risk factor for BC [52].

Only few studies discussed the associations between fat, oil and their products and BC risk and were summarized in a systematic review. This review showed that the total fat intake was positively related to BC risk when combing results from three case-control studies. However, no such association was observed in cohort studies [53]. The present study confirms findings from the case-control studies, in that a positive association was found.

A meta-analysis reported that overall meat intake was not related to the risk of BC; however, high red and processed meat intake was reported as a significant risk factor for BC risk, 17% and 10% risk respectively [54]. This increase is probably caused by the N-Nitroso compounds, which have been proposed as possible bladder carcinogens, found in red and processed meats [55]. In the present study, a high intake of 'meats and meat products (C)' was associated with an increased risk of developing BC. Again, future studies investigating specific types of meat could identify the types of meat or meat products that might have beneficial effects.

As an excretory organ, fluid intake might play an important role in the development of BC. A well-established risk factor is arsenic [56], through which people are most likely exposed by drinking water. The influence of other fluid sources on BC risk, however, are lacking evidence or are inconstant. Here we observed that high beverage intake is positively associated with BC risk. Again, it should be noted that only total 'beverage' intake was assessed, including both beverages with a potential protective effect on BC risk (*e.g.* green tea [57]) and beverages with a potential harmful effect on BC risk (*e.g.* alcoholic [58] and sweet non-alcoholic beverages [43]). It, therefore, remains unclear which caused the observed increased BC risk.

Since nutrition and cancer epidemiology is a complex field, the use of advanced analytic tools, such as data mining, is becoming increasingly important for unrevealing diet and health associations. Data mining has demonstrated its potential to complement conventional statistical regressions, particularly for nonlinear phenomena such as our dietary habits [59], and without requiring a priori assumptions on the relationship between diet and health outcomes [60]. In addition, data mining splits data files into training and validation sets, especially using cross-validation method gives relatively accurate predictive estimates. Furthermore, over fitting problem of both decision tree and rules could be minimized by using reduced error pruning technique in C5.0 [36] which is often problematic in conventional statistical techniques with a large number of variables and observations, such as the BLEND dataset. The strength of the present study is the high classification accuracy, which indicates the data mining methodology could adequately handle missing data and complex-investigating measurements. Therefore, the revealed food groups in the present study could be considered foods or pattern in relation to BC development.

A limitation of our study, however, is that the use of data mining in nutritional cancer epidemiology might only be useful in identifying key food items and can therefore only be seen as a hypothesis generator, which needs further detailed investigation in order to establish causation. Furthermore, we should acknowledge it is a complicated technique, which requires special knowledge and expertise and, thus, translating the results from data mining into simple health message is difficult challenge. In addition, the trees and

rules retrieved in the present study only include main food-groups, thereby, conflicting effects on BC risk of food subgroups or specific items was inevitable. Another limitation might have occurred by the designs of the data collection, which may have introduced recall- and/or selection bias, especially in case-control studies. In addition, for most included studies, the exposure variable was assessed by FFQs. Therefore, measurement error and misclassification of study participants in terms of the exposure and outcome are unavoidable: a) the inability of an FFQ to capture many details of dietary intake, such as all kinds and exact amounts of foods consumed, b) the difficulty in quantification of the intake and c) the high dependency on memory, which in turn may have influenced the robustness of dietary patterns extracted via the data mining procedure [(61)]. Lastly, due to the nature of data mining such as C5.0, there are concerns regarding multiple testing and spurious associations, which might cause some of the observed consequences due to chance alone.

Conclusion

In summary, the data mining technique provided an effective approach to identify some food groups related to BC risk in the large epidemiological BLEND study. The main findings from this study support the data mining approach to be a valuable additional methodology in nutrition- and cancer epidemiology, which deserve further examines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We gratefully acknowledge all principal investigators for their willingness to participate in this jointed project. The author E Y.W. Yu gives thanks to the financial support from China Scholarship Council (NO. 201706310135).

Funding

This work was partly funded by the World Cancer Research Fund International (WCRF 2012/590) and European Commission (FP7-PEOPLE-618308).

Hessen Case-control study on bladder cancer was supported by the Bundesanstalt für Arbeitsschutz (No. F 1287). The Kaohsiung study was supported by grant NSC 85-2332-B-037-066 from the National Scientific Council of the Republic of China. The Stockholm Case-control study was supported by grant from the Swedish National Cancer Society and from the Swedish Work Environment Fund. The Roswell Park Memorial Institute Case-control study on bladder cancer was supported by Public Health Service Grants CA11535 and CA16056 from the National Cancer Institute. The New England bladder cancer study was funded in part by grant numbers 5 P42 ES007373 from the National Institute of Environmental Health Sciences, NIH and CA57494 from the National Cancer Institute, NIH. The Italian Case-control study on bladder cancer was conducted within the framework of the CNR (Italian National Research Council) Applied Project ‘Clinical Application of Oncological Research’ (contracts 94.01321.PF39 and 94.01119.PF39), and with the contributions of the Italian Association for Cancer Research, the Italian League against Tumours, Milan, and Mrs. Angela Marchegiano Borgomainerio. The Brescia bladder cancer study was partly supported by the International Agency for Research on Cancer. The French INSERM study was supported by a grant from the Direction Générale de la Santé, Ministère des Affaires Sociales, France. The Molecular Epidemiology of Bladder Cancer and Prostate Cancer was supported in part by grants ES06718 (to Z.-F.Z.), U01 CA96116 (to A.B.), and CA09142 from the NIH National Institute of Environmental Health Sciences, the National Cancer Institute, the Department of Health and Human Services, and by the Ann Fitzpatrick Alper Program in Environmental Genomics at the Jonsson Comprehensive Cancer Center, UCLA. The Women’s Lifestyle and Health Study was funded by a grant from the Swedish Research Council (Grant number 521-2011-2955). The Netherlands Cohort Study on diet and cancer was supported by the Dutch Cancer Society. The RERF atomic bomb survivors Study was supported by The Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan, a public interest foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the US Department of Energy (DOE). The research was also funded in part through DOE award DE-HS0000031

to the National Academy of Sciences. This publication was supported by RERF Research Protocol RP-A5-12. The VITamins and Lifestyle Study (VITAL) was supported by a grant (R01CA74846) from the National Cancer Institute.

Abbreviations

BLEND	BLadder cancer Epidemiology and Nutritional Determinants
BC	Bladder Cancer
WCRF	World Cancer Research Fund
IARC	International Agency for Research on Cancer
FFQ	Food Frequency Questionnaires
SD	standard deviation (SD)
MAR	missing at random
MCAR	missing completely at random
MNAR	missing not at random
TPR	true positive rate
FPR	false positive rate
ROC	receiver operating curve
AUC	area under the curve

References

1. Ferlay J, Soerjomataram I, Dikshit R et al. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer 136, E359–386. [PubMed: 25220842]
2. Siegel RL, Miller KD, Jemal A (CA Cancer J Clin 2017) Cancer Statistics, 2017. 67, 7–30.
3. Al-Zalabani AH, Stewart KF, Wesselius A et al. (2016) Modifiable risk factors for the prevention of bladder cancer: a systematic review of meta-analyses. Eur J EPDEMIOL 31, 811–851.
4. Piyathilake C (2016) Dietary factors associated with bladder cancer. Investigative and clinical urology 57 Suppl 1, S14–25. [PubMed: 27326403]
5. Wiseman M World Cancer Research Fund/American Institute for Cancer Research. Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. Continuous Update Project Expert Report 2018.
6. Westhoff E, Wu X, Kiemeney LA et al. (Int J Cancer 2018) Dietary patterns and risk of recurrence and progression in non-muscle-invasive bladder cancer. 142, 1797–1804.
7. Witlox WJA, van Osch FHM, Brinkman M et al. (EUR J NUTR 2019) An inverse association between the Mediterranean diet and bladder cancer risk: a pooled analysis of 13 cohort studies. European journal of nutrition, 1–10.
8. Han J, Kamber M (2001) Data mining concept and technology. Publishing House of Mechanism Industry Amsterdam Elsevier Inc 3rd, 70–72.
9. Hearty AP, Gibney MJ (Am J Clin Nutr 2008) Analysis of meal patterns with the use of supervised data mining techniques--artificial neural networks and decision trees. 88, 1632–1642.
10. Goossens ME, Isa F, Brinkman M et al. (2016) International pooled study on diet and bladder cancer: the bladder cancer, epidemiology and nutritional determinants (BLEND) study: design and

baseline characteristics. Archives of public health = Archives belges de sante publique 74, 30. [PubMed: 27386115]

11. Bernstein L, Ross R (1991) Cancer in Los Angeles County. Los Angeles, CA: University of Southern California.
12. Tang L, Zirpoli GR, Guru K et al. (2008) Consumption of raw cruciferous vegetables is inversely associated with bladder cancer risk. Cancer Epidemiology and Prevention Biomarkers 17, 938–944.
13. Kellen E, Zeegers M, Lousbergh D et al. (2005) A Belgian case control study on bladder cancer: rationale and design. Archives of public health = Archives belges de sante publique 63, 17–34.
14. Wakai K, Takashi M, Okamura K et al. (2000) Foods and nutrients in relation to bladder cancer risk: a case-control study in Aichi Prefecture, Central Japan. Nutrition and cancer 38, 13–22. [PubMed: 11341038]
15. Lu C-M, Lan S-J, Lee Y-H et al. (1999) Tea consumption: fluid intake and bladder cancer risk in Southern Taiwan. Urology 54, 823–828. [PubMed: 10565741]
16. Pohlabeln H, Jöckel K-H, Bolm-Audorff U (1999) Non-occupational risk factors for cancer of the lower urinary tract in Germany. European journal of epidemiology 15, 411–419. [PubMed: 10442466]
17. Steineck G, Hagman U, Gerhardsson M et al. (1990) Vitamin A supplements, fried foods, fat and urothelial cancer. A case-referent study in Stockholm in 1985–87. International journal of cancer 45, 1006–1011. [PubMed: 2351481]
18. METTLIN C, GRAHAM (1979) Dietary risk factors in human bladder cancer. American journal of epidemiology 110, 255–263. [PubMed: 582494]
19. Baena AV, Allam MF, Del Castillo AS et al. (2006) Urinary bladder cancer risk factors in men: a Spanish case–control study. European journal of cancer prevention 15, 498–503. [PubMed: 17106329]
20. Brinkman MT, Karagas MR, Zens MS et al. (2010) Minerals and vitamins and the risk of bladder cancer: results from the New Hampshire Study. Cancer Causes & Control 21, 609–619. [PubMed: 20043202]
21. La Vecchia C, Negri E, Decarli A et al. (1995) Attributable risks for bladder cancer in northern Italy.
22. Shen M, Hung RJ, Brennan P et al. (2003) Polymorphisms of the DNA repair genes XRCC1, XRCC3, XPD, interaction with environmental exposures, and bladder cancer risk in a case-control study in northern Italy. Cancer Epidemiology and Prevention Biomarkers 12, 1234–1240.
23. Johnson K, Mao Y, Argo J et al. (1998) The National Enhanced Cancer Surveillance System: a case-control approach to environment-related cancer surveillance in Canada. Environmetrics: The official journal of the International Environmetrics Society 9, 495–504.
24. Ovsiannikov D, Selinski S, Lehmann M-L et al. (2012) Polymorphic enzymes, urinary bladder cancer risk, and structural change in the local industry. Journal of Toxicology and Environmental Health, Part A 75, 557–565. [PubMed: 22686316]
25. Clavel J, Cordier S (1991) Coffee consumption and bladder cancer risk. International journal of cancer 47, 207–212. [PubMed: 1988365]
26. Hemelt M, Hu Z, Zhong Z et al. (2010) Fluid intake and the risk of bladder cancer: Results from the South and East China case-control study on bladder cancer. International journal of cancer 127, 638–645. [PubMed: 19957334]
27. Cao W, Cai L, Rao JY et al. (2005) Tobacco smoking, GSTP1 polymorphism, and bladder carcinoma. Cancer: Interdisciplinary International Journal of the American Cancer Society 104, 2400–2408.
28. Taylor JA, Umbach DM, Stephens E et al. (1998) The role of N-acetylation polymorphisms in smoking-associated bladder cancer: evidence of a gene-gene-exposure three-way interaction. Cancer research 58, 3603–3610. [PubMed: 9721868]
29. van den Brandt PA, Goldbohm RA, van ‘t Veer P et al. (1990) A large-scale prospective cohort study on diet and cancer in The Netherlands. Journal of clinical epidemiology 43, 285–295. [PubMed: 2313318]

30. Poortvliet E, Klensin J, Kohlmeier L (Eur J Clin Nutr 1992) Rationale document for the Eurocode 2 food coding system (version 91/2). 46, S9–S24.
31. Hastie T, Tibshirani R, Friedman J (2009). In The elements of statistical learning, vol. 2nd, pp. 485–585. New York: Springer.
32. Rubin DB (1976) Inference and missing data. Biometrika 63, 581–592.
33. Van Ness PH, Murphy TE, Araujo KL et al. (J CLIN EPIDEMIOL 2007) The use of missingness screens in clinical epidemiologic research has implications for regression modeling. 60, 1239–1245.
34. Little RJ (J AM STAT ASSOC 1988) A test of missing completely at random for multivariate data with missing values. 83, 1198–1202.
35. Li C (2013) Little's test of missing completely at random. The Stata Journal 13, 795–809.
36. Pandya R, Pandya J (2015) C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications 117, 18–21.
37. Quinlan JR (2014) C4. 5: programs for machine learning: Elsevier.
38. Quinlan JR (1989) Unknown attribute values in induction. Proceedings of the sixth international workshop on Machine learning, 164–168.
39. Karaolis M, Moutiris JA, Pattichis CS (2008) Assessment of the risk of coronary heart event based on data mining. BioInformatics and BioEngineering, 2008 BIBE 2008 8th IEEE International Conference on, 1–5.
40. Lazarou C, Karaolis M, Matalas A-L et al. (COMPUT METH PROG BIO 2012) Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. 108, 706–714.
41. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. vol. 1: Springer series in statistics New York.
42. Louppe G, Wehenkel L, Sutera A et al. (2013) Understanding variable importances in forests of randomized trees. Advances in neural information processing systems, 431–439.
43. De Stefani E, Boffetta P, Ronco AL et al. (2008) Dietary patterns and risk of bladder cancer: a factor analysis in Uruguay. Cancer Causes Control 19, 1243–1249. [PubMed: 18592382]
44. Xu C, Zeng XT, Liu TZ et al. (2015) Fruits and vegetables intake and risk of bladder cancer: a PRISMA-compliant systematic review and dose-response meta-analysis of prospective cohort studies. Medicine (Baltimore) 94, e759. [PubMed: 25929912]
45. Vieira AR, Vingeliene S, Chan DS et al. (CANCER MED-US 2015) Fruits, vegetables, and bladder cancer risk: a systematic review and meta-analysis. 4, 136–146.
46. Liu H, Wang XC, Hu GH et al. (EUR J CANCER PREV 2015) Fruit and vegetable consumption and risk of bladder cancer: an updated meta-analysis of observational studies. 24, 508–516.
47. Yao B, Yan Y, Ye X et al. (CCC 2014) Intake of fruit and vegetables and risk of bladder cancer: a dose-response meta-analysis of observational studies. 25, 1645–1658.
48. Boeing H, Bechthold A, Bub A et al. (EUR J NUTR 2012) Critical review: vegetables and fruit in the prevention of chronic diseases. 51, 637–663.
49. Riboli E, Norat T (AM J CLIN NUTR 2003) Epidemiologic evidence of the protective effect of fruit and vegetables on cancer risk. 78, 559S–569S.
50. Chatenoud L, Tavani A, La Vecchia C et al. (1998) Whole grain food intake and cancer risk. International journal of cancer 77, 24–28. [PubMed: 9639389]
51. Augustin LSA, Taborelli M, Montella M et al. (BRIT J NUTR 2017) Associations of dietary carbohydrates, glycaemic index and glycaemic load with risk of bladder cancer: a case-control study. 118, 722–729.
52. Sun JW, Zhao LG, Yang Y et al. (2015) Obesity and risk of bladder cancer: a dose-response meta-analysis of 15 cohort studies. PloS one 10, e0119313. [PubMed: 25803438]
53. La Vecchia C, Negri E (CCC 1996) Nutrition and bladder cancer. 7, 95–100.
54. Wang C, Jiang H (MED ONCOL 2012) Meat intake and risk of bladder cancer: a meta-analysis. 29, 848–855.
55. Catsburg CE, Gago-Dominguez M, Yuan JM et al. (Int J Cancer 2014) Dietary sources of N-nitroso compounds and bladder cancer risk: findings from the Los Angeles bladder cancer study. 134, 125–135.

56. Baris D, Waddell R, Beane Freeman LE et al. (JNCI-J NATL CANCER I 2016) Elevated Bladder Cancer in Northern New England: The Role of Drinking Water and Arsenic. 108.

57. Miyata Y, Matsuo T, Araki K et al. (2018) Anticancer Effects of Green Tea and the Underlying Molecular Mechanisms in Bladder Cancer. Medicines (Basel) 5.

58. Vartolomei MD, Iwata T, Roth B et al. (2019) Impact of alcohol consumption on the risk of developing bladder cancer: a systematic review and meta-analysis. World J Urol 37, 2313–2324. [PubMed: 31172281]

59. Huys R, Jirsa VK (2010) Nonlinear dynamics in human behavior. vol. 328: Springer.

60. Crutzen R, Giabbanelli P (SUBST USE MISUSE 2013) Using Classifiers to Identify Binge Drinkers Based on Drinking Motives. 49, 110–115.

61. Rodrigo CP, Aranceta J, Salvador G et al. (2015) Food frequency questionnaires. Nutricion hospitalaria 31, 49–56. [PubMed: 25719771]

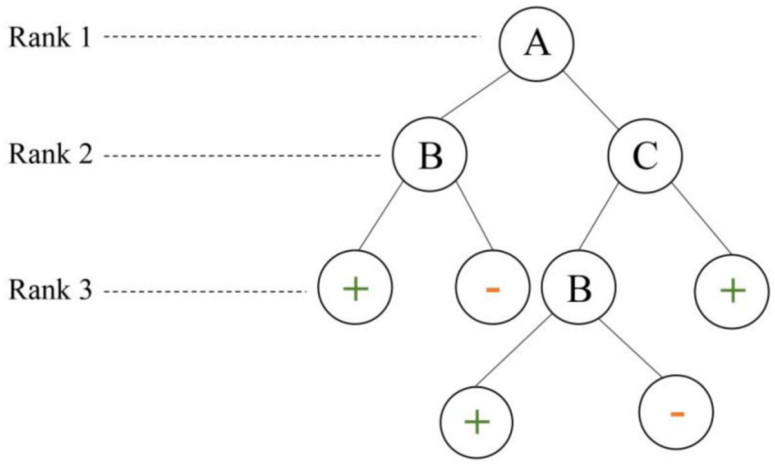

Figure 1.
Example of a decision tree. There are three individual variables, A, B and C, on which the tree splits. Variable A has an average ranking of 1 because it is the root node and appears only once. Variable B has an average ranking of 2·5, since it appears twice, once on the second and once on the third rank. Variable C has an average ranking of 2, since it is present only once and the tree splits on it after it split on A.

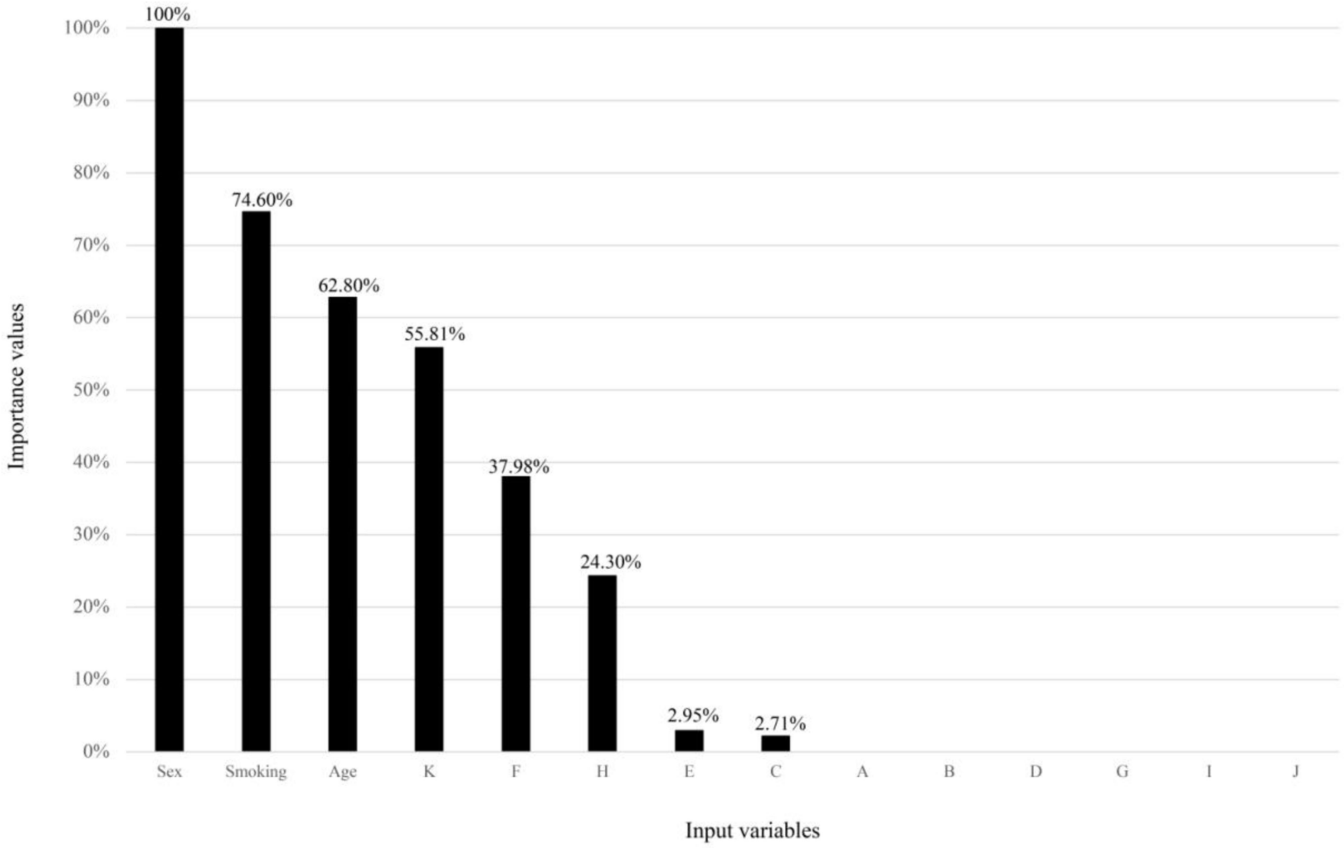

Figure 2.

Importance values of input variables after C5.0 in the BLEND data set. A: milk and dairy products; B: eggs and egg products; C: meats and meat products; D: fishes and fish products; E: fats, oils and their products; F: grains and grain products; G: pulses, seeds, kernels, nuts and their products; H: vegetables and vegetable products; I: fruits and fruit products; J: sugar and sugar products; K: beverages (non-milk). The importance values range from 0 to 100 %, where 0 % indicates ‘unimportant’ and 100 % indicates ‘extremely important’.

Table 1

The Baseline Characteristics and Food Group Information from the BLEND Dataset[†]

Variables	Cases (*N*=8,320)	Non-cases (*N*=23,231)	Missing Percentage
Sex (%)			0.00%
Male	6,601 (33.95)	12,841 (66.05)	
Female	1,719 (14.20)	10,390 (85.80)	
Smoking (%)			0.00%
Never	1,588 (15.11)	8,925 (84.89)	
Current	3,285 (30.97)	7,321 (69.03)	
Former	3,447 (33.04)	6,985 (66.96)	
Age (± SD)	61.80 (± 10.61)	58.52 (± 12.54)	0.00%
≤55 (%)	1,880 (20.46)	7,310 (79.54)	
55–60 (%)	1,511 (26.67)	4,514 (73.33)	
60–65 (%)	1,708 (28.22)	4,345 (71.78)	
65–70 (%)	1,531 (27.97)	3,943 (72.03)	
70–75 (%)	1,068 (31.23)	2,352 (68.77)	
>75 (%)	622 (35.56)	1,127 (64.44)	
Main Food Groups [mean servings/week (± SD)]			
Milk and Milk Products (± SD)	13.61 (± 18.39)	14.58 (± 23.47)	4.47%
Q1 (%): 0–5 servings/week	1,887 (22.14)	6,632 (79.86)	
Q2 (%): 5–9 servings/week	1,322 (19.93)	5,310 (80.07)	
Q3 (%): 9–18 servings/week	1,626 (21.81)	5,828 (78.19)	
Q4 (%): >18 servings/week	1,500 (19.92)	6,031 (80.08)	
Eggs and Egg Products (± SD)	2.65 (± 2.89)	2.54 (± 2.63)	11.78%
Q1 (%): 0–1 servings/week	2,117 (22.87)	7,141 (77.13)	
Q2 (%): 1–2 servings/week	1,003 (18.89)	4,306 (81.11)	
Q3 (%): 2–3 servings/week	1,246 (17.55)	5,852 (82.45)	
Q4 (%): >3 servings/week	1,275 (26.05)	4,894 (73.95)	
Meat and Meat Products (± SD)	7.75 (± 5.54)	7.35 (± 4.47)	7.62%
Q1 (%): 0–5 servings/week	1,931 (24.41)	5,981 (75.59)	
Q2 (%): 5–8 servings/week	1,810 (22.73)	6,154 (77.27)	
Q3 (%): 8–11 servings/week	1,387 (21.32)	6,505 (78.68)	
Q4 (%): >11 servings/week	1,298 (16.53)	6,555 (83.47)	
Fish and Fish Products (± SD)	1.94 (± 2.08)	1.39 (± 1.73)	5.72%
Q1 (%): 0–0.5 servings/week	918 (24.41)	7,415 (75.59)	
Q2 (%): 0.5–1 servings/week	1,163 (12.02)	8,515 (87.98)	
Q3 (%): 1–2 servings/week	1,387 (17.45)	4,287 (82.55)	
Q4 (%): >2 servings/week	1,298 (19.10)	5,293 (80.90)	
Fats and Oils (± SD)	8.61 (± 7.98)	9.99 (± 8.77)	21.44%
Q1 (%): 0–4 servings/week	1,291 (20.53)	5,641 (79.47)	
Q2 (%): 4–7 servings/week	1,561 (23.79)	5,760 (76.21)	
Q3 (%): 7–10 servings/week	780 (13.49)	5,386 (86.51)	

Variables	Cases (*N*=8,320)	Non-cases (*N*=23,231)	Missing Percentage
Q4 (%): >10 servings/week	1,152 (18.73)	5,654 (81.27)	
Grains and Grain Products (± SD)	16.17 (± 17.33)	15.40 (± 15.67)	5.36%
Q1 (%): 0–7 servings/week	2,061 (25.80)	5,928 (74.20)	
Q2 (%): 7–13 servings/week	1,503 (21.63)	5,446 (78.37)	
Q3 (%): 13–21 servings/week	1,494 (20.01)	5,973 (79.99)	
Q4 (%): >21 servings/week	1,570 (21.06)	5,886 (78.94)	
Pulses, seeds, kernels and nuts (± SD)	2.68 (± 3.88)	2.98 (± 4.44)	31.41%
Q1 (%): 0–0.75 servings/week	766 (13.89)	4,747 (86.11)	
Q2 (%): 0.75–1.5 servings/week	671 (12.06)	4,895 (87.94)	
Q3 (%): 1.5–3 servings/week	631 (12.24)	4,523 (87.76)	
Q4 (%): >3 servings/week	632 (11.69)	4,776 (88.31)	
Vegetables and Vegetable Products (± SD)	29.48 (± 47.94)	26.53 (± 34.52)	4.47%
Q1 (%): 0–12 servings/week	1,992 (26.29)	5,585 (73.71)	
Q2 (%): 12–17 servings/week	1,629 (21.74)	5,864 (78.26)	
Q3 (%): 17–29 servings/week	1,422 (18.86)	6,118 (81.14)	
Q4 (%): >29 servings/week	1,590 (21.12)	5,940 (78.88)	
Fruits and Fruit Products (± SD)	9.18 (± 9.97)	10.74 (± 12.59)	8.16%
Q1 (%): 0–3 servings/week	2,056 (26.46)	5,715 (73.54)	
Q2 (%): 3–6 servings/week	1,100 (14.79)	6,338 (85.21)	
Q3 (%): 6–14 servings/week	2,019 (28.55)	5,052 (71.45)	
Q4 (%): >14 servings/week	1,281 (18.69)	5,574 (81.31)	
Sugar and Sugar Products (± SD)	10.99 (± 14.67)	7.07 (± 10.65)	30.52%
Q1 (%): 0–1 servings/week	667 (10.40)	5,745 (89.60)	
Q2 (%): 1–4 servings/week	438 (9.37)	4,236 (90.63)	
Q3 (%): 4–10 servings/week	641 (11.95)	4,725 (88.05)	
Q4 (%): >10servings/week	896 (16.38)	4,573 (83.62)	
Beverages (non-milk) (± SD)	56.84 (± 17.63)	45.53 (± 13.47)	4.20%
Q1 (%): 0–28 servings/week	2,399 (23.90)	7,083 (76.10)	
Q2 (%): 28–42 servings/week	1,491 (23.86)	4,579 (76.14)	
Q3 (%): 42–62 servings/week	1,696 (24.15)	5,328 (75.85)	
Q4 (%): >62 servings/week	2,570 (34.40)	4,901 (65.60)	

[†]Age was coded as the original continuous values and 6 categorical values, food intakes were coded as quartile-order categorical values, and the other variables were coded as categorical dummy values.

Q1-Q4: lowest intake to highest intake (servings/week).

Abbreviation: BLEND= BLadder cancer Epidemiology and Nutritional Determinant; SD= Standard Deviation.

Table 2

Classification Rules Derived from C5.0 'Ruleset' in the BLEND Dataset [§]

Sex	Rules	Age	Smoking status	C	E	F	H	K	Case (%)	Non-case (%)
Male	1		Current		Q1	Q3–Q4	Q3–Q4	Q1–Q2	24%	76%
	2	40–63	Former			Q1		Q4	67%	33%
	3			Q3–Q4		Q1	Q1–Q2		87%	13%
	4		Never						23%	77%
	5		Current/Former						63%	37%
Female	6					Q2–Q4			13%	87%
	7		Former			Q1		Q1–Q2	80%	20%
	8	>63	Former			Q1		Q3–Q4	84%	16%

[§]Age: years old; C-K: servings/week.

C: meats and meat products; E: fats, oils and their products; F: grains and grain products; H: vegetables and vegetable products; K: beverage (non-milk).

Q1–Q4: lowest intake to highest intake (servings/week).