**Supplementary Material**

1. **Prediction Modelling Steps using Stepwise Procedures**

We developed a series of model-building steps that are outlined below. All modelling was conducted using multivariable least squares regression, with selection of predictors based on using both the forward and backward procedures for the main analyses. We also conducted a sensitivity analysis using the lasso method.(1) In instances where the forward and backward procedures yielded different predictors, all predictors selected by at least one of the two procedures were included at the next step of modeling. As our initial examination of distributions of variables showed that the prevalence of vitamin D supplement use was higher than expected based on past studies(2,3) we conducted a set of preliminary analyses where interaction terms between each predictor variable and supplement use were included. We observed several statistically significant interactions (results not shown), thus, we performed all our multivariable model building separately for users versus non-users of vitamin D supplements.

**Table S1**: Description of the different steps of the model building strategy for stepwise methods for users

|  |  |  |  |
| --- | --- | --- | --- |
| Step: Aim (threshold for p-value) | Variables forced in | Variables to test | Retained variables |
| 1: Linearity (0.15) | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking + supplement vitamin D intake | age2 + BMI2 + sun sensitivity score2 + dietary vitamin D intake2 + alcohol intake2 + outdoor time with total sun protection2 + outdoor time with partial sun protection2 + outdoor time with no sun protection2 + smoking2 + supplement vitamin D intake2 | BMI2 + supplement vitamin D intake2 |
| 2: Interaction with season (0.15) | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking + supplement vitamin D intake | season:age + season:ethnicity + season:education level + season:BMI + season:menopausal status + season:sun sensitivity score + season:recent sun vacation and sun protection + season:dietary vitamin D intake + season:alcohol intake + season:smoking + season:supplement vitamin D intake | season:menopausal status |
| 3.1: Model Building – Selection among all linear terms (0.15) |  | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking + supplement vitamin D intake | supplement vitamin D intake + menopausal status + recent sun vacation and sun protection + alcohol intake + outdoor time with partial sun protection + ethnicity |
| 3.2: Model Building – Selection among all linear terms, quadratic terms (Step 1) and interactions terms (Step 2) (0.10) |  | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection+ dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking + supplement vitamin D intake + BMI2 + supplement vitamin D intake2 + season:menopausal status | supplement vitamin D intake + menopausal status + recent sun vacation and sun protection + alcohol intake + outdoor time with partial sun protection + ethnicity + supplement vitamin D intake2 + season+BMI + sun sensitivity score + BMI2+ season:menopausal status |
| 3.3: Model Building – Final complex model (0.05) |  | season + ethnicity + BMI + menopausal status+ sun sensitivity score + recent sun vacation and sun protection+ alcohol intake + supplement vitamin D intake + outdoor time with partial sun protection + BMI2 + supplement vitamin D intake2 + menopausal status:season | season + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + alcohol intake + supplement vitamin D intake + outdoor time with partial sun protection + BMI2+ menopausal status:season |
| 3.4: Model Building – Final simple model (0.15) |  | season + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + alcohol intake + supplement vitamin D intake + outdoor time with partial sun protection | menopausal status + recent sun vacation and sun protection+ alcohol intake + supplement vitamin D intake + outdoor time with partial sun protection |

**Table S2**: Description of the different steps of the model building strategy for stepwise methods for non-users

|  |  |  |  |
| --- | --- | --- | --- |
| Step: Aim (threshold for p-value) | Variables forced in | Variables to test | Retained variables |
| 1: Linearity (0.15) | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking | age2+ BMI2 + sun sensitivity score2 + dietary vitamin D intake2 + alcohol intake2 + outdoor time with total sun protection2 + outdoor time with partial sun protection2 + outdoor time with no sun protection2+ smoking2 | BMI2 + sun sensitivity score2 + outdoor time with no sun protection2 |
| 2: Interaction with season (0.15) | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking | season:age + season:ethnicity + season:education level + season:BMI + season:menopausal status + season:sun sensitivity score + season:recent sun vacation and sun protection+ season:dietary vitamin D intake + season:alcohol intake+ season:smoking | season:education level + season:dietary vitamin D intake |
| 3.1: Model Building – Selection among all linear terms (0.15) |  | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake+ outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking | age + season + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection |
| 3.2: Model Building – Selection among all linear terms, quadratic terms (Step 1) and interactions terms (Step 2) (0.10) |  | age + season + ethnicity + education level + BMI + menopausal status + sun sensitivity score + recent sun vacation and sun protection + dietary vitamin D intake + alcohol intake + outdoor time with total sun protection + outdoor time with partial sun protection + outdoor time with no sun protection + smoking + BMI2 + sun sensitivity score2 + outdoor time with no sun protection2 +  season:education level + season:dietary vitamin D intake | season + education level + recent sun vacation and sun protection + BMI + BMI2 + sun sensitivity score + sun sensitivity score2 + season:education level |
| 3.3: Model Building – Final complex model (0.05) |  | age + season + BMI + menopausal status+ sun sensitivity score + recent sun vacation and sun protection + BMI2 + sun sensitivity score2 + education level + season:education level | season + recent sun vacation and sun protection + BMI + BMI2 + sun sensitivity score + sun sensitivity score2 |
| 3.4: Model Building – Final simple model (0.15) |  | season + BMI + sun sensitivity score + recent sun vacation and sun protection | season + recent sun vacation and sun protection + BMI + sun sensitivity score |

1. **Prediction Modelling using Lasso**

**Introduction**

In addition to forward and backward procedures, the prediction modelling was conducted using the lasso method using the *glmnet* R package. In contrast to the forward and backward procedures, all covariates, that is, all linear terms (.i.e categorical and continuous linear terms), quadratic terms and interaction terms were simultaneous candidates in the lasso method. To select the best fitting model, a lasso cross-validation was carried out to see for which coefficients the minimal mean square error is reached. The final model is defined by the covariates for which the coefficient was not assigned to zero after shrinkage when the minimal mean square error is achieved.

**Method**

To avoid unsteady results due to the small sample size, the lasso cross-validation was carried out 2000 times. The covariates retained more than 80% of the time then defined the model. If no covariate was retained more than 80% of the time, the model was defined by the most retained covariate. The coefficients of these covariates were equal to the mean of the coefficients when the covariates were retained among these 2000 times.

**Result**

The best predictors identified by the lasso procedure for users of vitamin D supplements were the dose of vitamin D supplement use and menopausal status, which were the only covariates that were retained between 70% and 90% of the time. They are both positively associated with 25(OH)D (Table 2).

The best predictors of 25(OH)D concentration, based on the lasso results for non-users of vitamin D supplements, were BMI, for which a non-linear effect was identified, and recent vacation with sun protection (Table 3). These two covariates were the only ones retained between 10% and 35% of the time. No covariate was retained more than 35% of the time.

1. **Application of previously published prediction models**

The prediction models from our forward and backward procedures were compared to the performance of two models from the literature.(4,5) Sahota, *et al.*, was selected for comparison as their study represented exposures experienced in a Canadian population that would be relevant to our study population.(4) Bertrand *et al*., was selected as this study was the largest and most comprehensive modelling endeavor for 25(OH)D.(5) Briefly, among 217 and 203 women in Toronto, Canada, Sahota *et al*., examined the relationship between vitamin D-related questions collected from a telephone interview with serum 25(OH)D, among those participating in summer and winter months, respectively.(4) In their models, 25(OH)D levels was best predicted, in the summer, by outdoor time (i.e. number of days when more than 0.5 hour was spent outdoors per week), sun protection behaviours (limb coverage) and milk intake (glass/week) in the summer while adjusting for age, BMI and skin colour. In the winter, the best predictors of 25(OH)D levels included sunlamp use, milk intake and vitamin D-containing supplement intake while adjusting for age, BMI and skin colour.

In Bertrand *et al*., determinants of 25(OH)D was investigated in three nationwide cohorts: the Nurses’ Health Study (NHS; n=2079), NHSII (n=1497) and the Health Professionals Follow-up Study (HPFS; n=911).(5) Variables included in the models by Bertrand *et al*. included age, race, UV-B radiation flux at residence, dietary and supplementary vitamin D intakes, BMI, physical activity, alcohol intake (not in HPFS), post-menopausal hormone use (not in HPFS), and season of blood draw. As the HPFS was conducted among men only, no comparisons were made with the results of the HPFS.

**Methods**

Using the regression coefficients provided in each study, we calculated the predicted 25(OH)D levels among our study participants. We than calculated the correlation coefficient between predicted versus actual 25(OH) measured in our study population.

The following variables were defined our study population to apply the Sahota model: age (years; continuous), skin color (very fair, fair, light olive, dark olive, brown or black), BMI (kg/m2; continuous), milk (<1; 1-5; >5-10; 10+), vitamin D supplement use (none; ≤400 IU (international units), equivalent to moderate; > 400 IU, equivalent to high). The summer analysis also included the following items: days outside per week (≤ 3,5 hours per week outside; equivalent to < 0,5 hours for 7 days outside; ≥ 3,5 hours per week outside; equivalent to 0,5 hours for 7 days outside), limb coverage (total coverage; partial coverage; no coverage). The winter analysis included sunlamp use (yes; no) as an additional item. Among participants of our study that participated during the summer months (April to September), the summer model was used to estimate 25(OH)D concentrations; among participants that participated during the winter months (October to March), the winter model was used.

For the models from Bertrand *et al*., we did not use the variable for average annual UVB flux based on the state of residence as it did not pertain to our study population in Montreal where latitude does not vary meaningfully. The following variables were defined in our study population: age (years; continuous), race (white, black, asian, hispanic, other), dietary vitamin D (IU/day ; < 100; 100-199; 200-299; 300-399; ≥400), supplementary vitamin D (IU/day; 0; 1-199; 200-399; ≥400), body mass index (kg/m2; <19; 19-21.9; 22-24.9; 25-29.9; 30-34.9; ≥35), quintile of physical activity (5 levels whereby level 1 represents a low physical activity and level 5 represents a high physical activity), alcohol intake (0, > 0 - < 5, 5 - < 10, ≥ 10), season of blood draw (fall, summer, spring winter). Postmenopausal hormone use was defined differently in NHS (premenopausal; postmenopausal - never user; postmenopausal - past user; postmenopausal - current user; postmenopausal - unknown user) and NHSII (never user; past user; recent past user; current user; unknown user).

**Results**

The mean difference between the predicted versus observed 25(OH)D values was close to --50.5 and -54.7 nmol/L for the NHS and NHSII models by Bertrand *et. al,* and +32 nmol/L for the models in Sahota *et. al.* The Pearson correlation coefficients between the predicted and observed 25(OH)D values, were 0.37, 0.39 and 0.14 for the Sahota model, the first Bertrand model (based on NHS) and the second Bertrand model (based on NHSII), respectively.

**REFERENCES**

1. Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodol* **73**, 273-282.

2. Kuhn T, Kaaks R, Teucher B *et al.* (2014) Dietary, lifestyle, and genetic determinants of vitamin D status: a cross-sectional analysis from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Germany study. *Eur J Nutr* **53**, 731-741.

3. van der Meer IM, Boeke AJ, Lips P *et al.* (2008) Fatty fish and supplements are the greatest modifiable contributors to the serum 25-hydroxyvitamin D concentration in a multiethnic population. *Clin Endocrinol (Oxf)* **68**, 466-472.

4. Sahota H, Barnett H, Lesosky M *et al.* (2008) Association of vitamin D related information from a telephone interview with 25-hydroxyvitamin D. *Cancer Epidemiol Biomarkers Prev* **17**, 232-238.

5. Bertrand KA, Giovannucci E, Liu Y *et al.* (2012) Determinants of plasma 25-hydroxyvitamin D and development of prediction models in three US cohorts. *Br J Nutr* **108**, 1889-1896.