Supplementary

Formulas of the risk prediction models along with the mathematical and statistical aspects

A comparative simulation study of (I) Tammemagi's $PLCO_{m2012}$, (II) LLP, and (III) Bach's lung cancer risk models' was performed. Data required for risk estimation by different models was collected by the core research team and two independent groups of dieticians and cardiologists. The variables of less than 5% of missing data were subjected to missing data imputation; the k-nearest neighbour algorithm was used for that purpose.

Tammemagi's $PLCO_{m2012}$ individual risk was calculated as (24):

$$risk = \frac{\exp(\beta \cdot x)}{1+\exp(\beta \cdot x)}$$

where β denotes the vector of risk factor weighting coefficients (the log of odds ratios), while x means the vector of the individual estimates of risk factors. The following factors were considered in the study: (I) ethnicity; (II) age; (III) education; (IV) body mass index; (V) family history of lung cancer; (VI) personal history of cancer; (VII) COPD diagnosis; (VIII) Smoker status, smoke duration and number of cigarettes per day; and (IX) quit duration (if any).

The construction of the LLP model required personal data on (I) presence of pneumonia, (II) asbestos exposure, (III) personal history of cancer, (IV) family history of cancer, (V) smoking duration, and (VI) age. The risk estimate was obtained with the use of the following formulae (25):

$$risk = \frac{1}{1+\exp\left(-(\beta \cdot x)\right)}$$

where, as above, β denotes the vector of risk factor weighting coefficients and x means the vector of the risk factors. Similarly to Tammemagi's model, logs of odds ratio served as the weighting coefficients.

The last of the analysed models, proposed by Bach *et al.* (26) expresses risk as:

$$risk = \sum\nolimits_{\downarrow}(i=0)^{\uparrow}(T-1)\left[\left(1-S_{\downarrow}0^{\uparrow}\left(\exp\left(\beta_{\downarrow}0\cdot x\cdot(\alpha+i)\right)\right)\cdot S_{\downarrow}1^{\uparrow}\left(\exp\left(\beta_{\downarrow}0\cdot x\cdot(\alpha+i)\right)\Pi_{\downarrow}(j<i)\left[\left[(S]\right]_{\downarrow}0^{\uparrow}\left(\exp\left(\beta_{\downarrow}0\cdot x\cdot(\alpha+j)\right)\right)\cdot S_{\downarrow}1^{\uparrow}\left(\exp\left(\beta_{\downarrow}0\cdot x\cdot(\alpha+j)\right)\right)\right]\right)\right)\right]$$

where S_0 is the baseline lung-cancer free survival beyond 1-year (equal to 0.996229), S_1 is the baseline overall survival beyond 1-year (equal to 0.9917663), β_0 is the relative risk factors for lung cancer, β_1 represents the relative risk factors for competing mortality at age α, and T stands for time of the prediction window. In contrary to the Tammemagi's and LLP models, the β vector includes the logs of relative risk estimates.

Education variable production

In our database a type of education was specified by two variables—a type of work (given answers were "physical" or "mental") and current, professional status (given answers were "active" or "retired").

Initial three education classifiers ("less than high school grad"—1; "high school grad"—2; "post-high school training"—3) in Polish labour market are ascribable to a "physical" type of work, whereas educational categories from class 4 to 6 ("some college"—4; "college grad"—5; "postgraduate"—6) are related to a "mental", i.e., conceptual job. The latter entails some amount of analysis, abstractive reasoning, and decision making while performing one's duties. Instead, "physical" work is usually repetitive action of mechanical effort aimed at fulfilling some relatively simple tasks. Therefore, in the category type of work, we have dichotomised Tammemagi's model into two almost equal subsets. It is evident while calculating lung cancer development probabilities with the model. The relationship pertains to former and current smokers alike. Each level of rising education from 1 to 6, elicits the decline of lung cancer risk by 0.2% to 0.3%. Since Tammemagi's model does not comprise the variable active or retired worker, this category, specified in our classification as non-existent, has no implication on its predictive efficacy. Thus, given the socio-economic status in Poland, authors considered that answer "physical" (manual) refers to "high school grads" (levels 1–3) and "mental" (conceptual) to "college grads" (levels 4–6).

References

24. Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. N Engl J Med 2013;368:728-36.
25. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 2008;98:270-6.
26. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95:470-8.