

# Standalone Smoking Vaping Calibration

Tim Wilson

June 27, 2024

## 1 Underlying Simulation Model

The underlying model for smoking and vaping is a compartmental flow model. The model tracks 220 cohort, where a *cohort* is a combination of age and sex. The model is initialised with `initialPrevalance.csv` runs from year 2021 to 2039 inclusive. Within each year, the model increments ages and then evaluates the flows in the model.

### 1.1 Increment Ages

The age of each cohort is incremented at the start of the year. In other words, the prevalence among a cohort of age  $n$  at the start of year  $t$  is equal to the prevalence of the cohort with age  $n - 1$  at the end of year  $t - 1$ .

Cohorts of age  $\geq 110$  are removed, and cohorts of age 0 are inserted each year. The newly inserted cohorts are initialised with the never smoker state ( $n$ ) set to 1.

### 1.2 Evaluate Flows

Flows are evaluated using a rough approximation, because the rates tend to be low. The quality of the optimisation depends on the number of time steps used.

The flow rate are determined by reading the base rate from `flow_rate.csv` and applying an annual percentage change (APC), relative to the base year, from `flow_apc.csv`. The flow from state  $x$  to state  $y$  is denoted `x_y` in the file. Let  $S$  be the states (compartments) of the model, then for  $x, y \in S$ , the flow rate in year  $t$  is

$$r_{x,y,t} := b_{x,y}(1 + a_{x,y})^{t-2021}$$

where  $b_{x,y}$  is base rate and  $a_{x,y}$  is the APC.

The flows are applied over a number of time steps,  $s$ , which defaults to 4. Within each time step, the total flow out of a state,  $x \in S$ , is

$$f_{x,t} := \sum_{s \in S} r_{x,y,t}$$

and the mass that flows from state  $x$  to  $y \in S$  is

$$p_{x,y,t} := \left(1 - e^{-\frac{f_{x,t}}{s}}\right) \frac{r_{x,y,t}}{f_{x,t}}.$$

Essentially, the flow *rate* out of  $x$  is being converted to a proportion, then split proportionally by the contribution each individual flow makes to the rate.

The mass,  $m_x$  of a state is then updated as follows

$$\begin{aligned} \text{out}_x &:= m_x \sum_{s \in S} p_{x,s} \\ \text{in}_x &:= \sum_{s \in S} m_s p_{s,x} \\ m_x &\xrightarrow{\text{update}} m_x - \text{out}_x + \text{in}_x \end{aligned}$$

This process is repeated  $s$  times per year.

## 2 Calibration

The model is calibrated by modifying a limited set of base flow rates and APCs. The male and female cohorts can be calibrated separately, so sex is ignored for the following.

### 2.1 Objective Function

The target of the model is the sum of the square of the difference between the prevalence values calculated for each year, and the values in `prevalenceTargets.csv`. The target file sets aggregate smoking and vaping targets, rather than targets for particular states in the compartmental model. The aggregates map to the states as follows, and must be computed from model output to compare to the targets.

$$\begin{aligned} \text{smoking} &= (\mathbf{s} + \mathbf{sv}) / (1 - \mathbf{dead}) \\ \text{vaping} &= (\mathbf{v} + \mathbf{vrs} + \mathbf{sv}) / (1 - \mathbf{dead}) \end{aligned}$$

Contributions to model score is weighted by `scoreWeight.csv`, where the agecategory part of the index is expanded to ages  $n$  to  $m - 1$ , where  $n$  is the current category and  $m$  is the next category. The last category has  $m = 110$ .

In summary, the objective function scores a point in parameters space as follows.

$$\text{score} := \sum_{a \in [0 \dots 110]} \sum_{t \in [2021 \dots 2040]} \sum_{B \in \{\text{smoking, vaping}\}} \left( w_{a,t,B} (\text{target } B_{a,t} - \text{model } B_{a,t}) \right)^2$$

The weights are squared for technical reasons to do with how the full framework processes cohort comparisons.

### 2.2 Parameter Space

The parameter space is a limited set of flows, with pins at particular ages. The parameter space only directly contains the flows `n_s`, `s_rs`, `n_v`, `v_rv`, `s_vrs` and `v_s`. All other flows are either static (as in the case for flows to `dead`), or derived from other flows.

The parameter has pins at ages 0, 14, 15, 16, 18, 21, 24, 28, 33, 40, 50, 60, 70, 90 and 109. Values for other ages are linearly interpolated between these pins. There are pins at ages. The 14, 15, 16 pins allow for sudden sharp uptake rates at age 15, to capture pre-15 prevalence, and the pins are more frequent for younger ages as the calibration target shows sharper changes in behaviour at these ages.

In total, the model has 180 ( $6 \times 15 \times 2$ ) parameters, corresponding to combinations of flows, age pins, and split by base rate and APC. The range of each parameter is defined by the `flow_apc_upper`, `flow_apc_lower`, `flow_rate_upper` and `flow_rate_lower` files. These latter two files are indexed by agecategory, which can be expanded in ages as explained in the section above. In practise the model has fewer than 180 parameters since many of them have an upper and lower bound of zero.