Standalone Smoking Vaping Calibration

Tim Wilson

July 15, 2024

The goal of smoking and vaping model calibration is to project uptake and cessation rates. The underlying model is a compartmental flow model on the following states:

- `n` - Someone who has never smoked or vaped.
- `s` - Current smoker.
- `v` - Current vaper.
- `sv` - Someone who smokes and vapes.
- `rs` - Someone who doesn't smoke or vape, who most recently quit smoking.
- `rv` - Someone who doesn't smoke or vape, who most recently quit vaping.
- `vrs` - Current vaper who previously quit smoking.
- `dead` - Someone who has died. Differential mortality rates for smokers require us to track this state so that the prevalence estimates among the living make sense.

See `calibrationExample.png` for a diagram of the model. The plots on states are initial prevalence, while the plots on the tails of arrows are the rate from that state. The state `vrs` is omitted because flows out of `vrs` are identical to those out of `v`.

Projections for the aggregate smoking and vaping rates are input to the model. We produce these in an earlier step. Since we already have aggregate projections, the calibration of the compartmental model has two purposes:

- Break down aggregate prevalences, to extract realistic estimates of dual users (`sv`), and people who quit into the future. Quit states are important due to lingering health impacts.
- Provide flow rates between the states. These let us parameterise intervention models in terms of changes in uptake and cessation rates for specific cohorts, then watch the effect sweep through the population.

Unfortunately the true goal of calibration is a bit fuzzy, and we have only been able to encode a few of them so far. The final check is how 'realistic' a set of outputs look, based on the expertise of people who spent decades studying smoking. This is the source of some of the weirder constraints.

1 Underlying Simulation Model

The model tracks 110 ages and two sexes, for a total of 220 cohorts. A compartmental model state exists for each cohort, where the sum of the states is always 1. The state of each cohort is initialised from `initialPrevlance.csv`, and the model then 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 state of the model for a cohort of age n at the start of year t is equal to the state for the cohort with age $n-1$ at the end of year $t-1$. The age is incremented in first year of the model.

Cohorts are removed when they reach age 110, and new cohorts are added with age 0. The newly inserted cohorts are initialised with $\texttt{n} = 1$.

1.2 Evaluate Flows

Flows between states are evaluated using a rough approximation, because the flow rates tend to be low. The quality of the optimisation depends on the number of time steps used. The number of steps per year defaults to 4.

Flow rate are determined by reading the base rate from `flow_rates.csv` and applying an annual percentage change (APC), relative to the base year, from `flow_apc.csv`. Note that these files are indexed by current age, not by year of birth, so a birth cohort will find itself sweeping across different base flow rates.

The flow from state x to state y is denoted `x_y` in the flow files. The flow graph is a directed graph with an incomplete number of arcs. For example, flows `s_sv`, `sv_s` and `n_s` each exist independently, but not `s_n`.

Let S be the states 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. For $x \in S$ and year t, define

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

then the mass that flows from 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, we convert the total flow *rate*, $f_{x,t}$, into the amount of flow in $1/s$ years, then proportionally split the flow between recipients.

The mass, $m_{x,t}$ of a state is then updated as follows

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

This process is repeated s times per year, with the final update to $m_{x,t}$ being used as the state output from the model for the cohort in year t.

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 section.

2.0.1 agecategory

Some input files have agecategory rather than age as part of their index. These files are interpreted by expanding each row to cover all ages up to the following agecatgory. The final age category expands up to age 109 inclusive.

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 are computed from model output at the end of each year as follows.

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

Contributions to the score are weighted by `scoreWeight.csv`. The weights are not a particularly important part of the objective function. Supporting them is good, but I am not wedded to these particular weights.

In summary, the objective function scores the output of a model as follows.

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

The weights are squared for technical reasons to do with how the full framework processes cohort comparisons. This detail is also unimportant.

2.2 Parameter Space

The parameter space consists of the 180 ($6 \times 2 \times 15$) combinations of the following factors.

- The flows `n_s`, `s_rs`, `n_v`, `v_rv`, `s_vrs` and `v_s`.
- The base rate and APC of the flows are separate parameters.
- The age 'pins' 0, 14, 15, 16, 18, 21, 24, 28, 33, 40, 50, 60, 70, 90 and 109.

The `<x>_dead` flows are static, so they should appear as they are in the original flow file. The other parameters are derived from the directly from the flows in the parameter space

The values for the remaining ages are set by linear interpolation between the pins. 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.

The full list of flows can be found in the column names of `flow_rates.csv`. Many of these flows are derived from the six that are directly varied in the parameter space. The file `flowInteractions.csv` defines these relationships, with column `arc1 from arc2` meaning that the values of `arc1` are determined by `arc2`.

- Derived columns of `flow_rates.csv` use the data in `flowInteractions.csv`, eg, the value of `v_sv` is 1.42 times the value of `n_s` for 19 year olds.
- `sv_rs` is also a derived column. Its value at each row is the minimum of `s_rs` and `v_rv`.
- Derived columns of `flow_apc.csv` always copy the source column. In other words, the multiplier is always 1.

These interactions are in place to cut down the size of the parameter space, and to encode what literature exists on increased smoking and vaping update rates for people who already vape or smoke, respectively.

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 Section 2.0.1. In practise the model has fewer than 180 parameters since many of them have and upper and lower bound of zero.

The parameter space has an additional constraint to cut down on extreme or spiky behaviour. This constraint requires that the value among certain collections of pins be uni-modal, ie that they have a unique local maximum. The sets of pins are as follows.

- `[16, 18, 21, 24, 28, 33, 40]`
- `[40, 50, 60, 70, 90, 109]`

Note that the pin at age 15 does not appear. It is unconstrained so that earlier smoking can be captured by a sudden spike of uptake.