The University Of Sheffield.

A methodology for estimating annual probabilities of smoking initiation, quitting and relapse from repeat cross-sectional survey data of smoking in England

Version 1.3.0

December 2020

Duncan Gillespie[1], Laura Webster[1], Magdalena Opazo Breton[1], Colin Angus[1] & Alan Brennan[1]

[1]School of Health and Related Research (ScHARR), The University of Sheffield

Address for correspondence:
Dr Duncan Gillespie
Section of Health Economics and Decision Science,
School for Health and Related Research,
The University of Sheffield,
Regent Court, Regent Street, Sheffield, S1 4DA, UK
Email: duncan.gillespie@sheffield.ac.uk

WARNING

This is a working draft version that is subject to review and future versions are likely.

The code that accompanies this report can be found in the `smktrans` R package (https://stapm.gitlab.io/r-packages/smktrans/), and in the other R packages that underlie the Sheffield Tobacco Policy Model as described on the Sheffield Tobacco and Alcohol Policy Modelling website (https://stapm.gitlab.io/).

Summary

This report presents a methodology to estimate these smoking state transition probabilities from repeat cross-sectional smoking survey data. We developed equations for the annual probabilities of smoking initiation, quit and relapse, based on previous methods to infer smoking state transition probabilities from repeat cross-sectional data, and on the mathematical description of the STPM. We describe the data sources that we use to inform the parameters in these equations for the estimation of smoking state transition probabilities. In particular, our methodology advances previous methods to estimate the probabilities of quitting smoking by including adjustments of the estimated probabilities for long-term smoking relapse and the differential mortality of smokers. We also apply adjustments in an effort to account for biased recall of survey respondents and demographic variation in population structure.

Glossary of terms

Table 1: Explanation of terms.

Term	Definition
STPM	The Sheffield Tobacco Policy Model
Projection or forecast	An estimate of a future situation based on past trends and a set of assumptions about how these past trends might continue into the future. Often multiple projections are presented that represent the alternative assumptions that might be made about how past trends might continue.
State transition probability	The state transition probability of a Markov chain gives the probabilities of transitioning from one state to another in a single time unit, e.g. the probability of moving from the 'never' to the 'current' smoking states.

Mathematical notation

Table 2: Overview of mathematical notation.

Symbol	Description
a	age in years (0, 1, 2, 3, ...)
a_{min}	youngest age in the model
a_{max}	oldest age in the model
y	period in calendar years (2001, 2002, 2003, ...)
c	birth cohort in calendar years i.e. the years of birth of the individuals in our model, $cohort = y - a$ (1940, 1941, 1942, ...)
A	population size i.e. the number of individuals who are simulated to be alive in year y
$current$ or $_c$ (subscript)	current regular cigarette smoker
$former$ or $_f$ (subscript)	former regular cigarette smoker
$never$ or $_n$ (subscript)	never regular cigarette smoker
θ	probability density function (used to describe the distribution of the population among discrete states)
$quityears$	years spent as a former smoker i.e. time since quitting (0, 1, 2, 3, ... k years)
$P(ever\text{-}smoker)$	probability that someone has ever smoked regularly
$P(initiate)$	probability of smoking initiation by never smokers i.e. the probability that someone starts to smoke regularly for the first time between ages a and $a+1$
$P(relapse)$	probability of relapse to smoking by former smokers i.e. the probability that, between ages a and $a+1$, someone starts to smoke regularly again after having been a former smoker for a specified number of years ($quityears$)
$P(quit)$	probability of quitting smoking by current smokers i.e. the probability that a regular smoker aged a will have become a former smoker by age $a+1$
$P(survive)$	the probability that an individual survives from age a to age $a+1$
i	individual index
j	population subgroup (10 strata: 2 x sexes, 5 x Index of Multiple Deprivation quintiles)
m	mortality rate
rr	relative risk of disease
h	disease index
l	the cohort survivorship function (the probability that someone born in a particular year will survive to age a years)

Contents

1 Introduction

To reliably model the effects of tobacco control policies and interventions on population trends in smoking rates, it is important to reliably estimate the baseline smoking state transition probabilities of smoking initiation, quit and relapse by age, sex and socio-economic conditions, and to investigate how these transition probabilities change over the life-course of individuals and at the population-level over calendar time.

2 Background

In the 1950s, Doll and Hill published the first evidence that smoking increases mortality (Doll and Hill 1954, 1964), and in 1962, the Royal College of Physician's 'Smoking and health' report promoted a range of smoking prevention measures, including preventing tobacco advertising, increasing prices, making public places smoke free, providing treatment for smokers, educating the public and restricting young people's access to cigarettes (Royal College of Physicians 1962). In 1998, the UK's Smoking White Paper 'Smoking Kills' introduced the first coordinated set of tobacco control policies (Secretary of State for Health 1998). These actions, followed by successive national tobacco control strategies, precipitated the decline of the tobacco epidemic. England is now in the final phase of the tobacco epidemic when, although the rate of smoking in adults was 14.4% in 2018 according to the Annual Population Survey (around 6 million smokers), there is a high frequency of former smokers and the health and mortality effects of smoking are still emerging in cohorts where smoking was once common. The government recently set the ambition for the England to be smoke-free by 2030, meaning that smoking rates in adults should be less than 5%. Given the short time to achieve this ambition, it will be necessary to at least maintain and probably increase the rate at which people who currently smoke are making attempts to quit, increase the success of quit attempts, and continue the declines in smoking initiation. In addition, smoking is now concentrated among people living in more deprived socio-economic conditions (Hiscock et al. 2012; NHS Digital 2019), which indicates that to understand how England might achieve its smoke-free target, it is important to understand how the population dynamics of smoking have differed among people living in different socio-economic conditions, and how these differences might continue in future.

2.1 The use of smoking state transitions probabilities in previous computer models of smoking

Computer models are a useful tool to project what might happen to the population-level rates of smoking in the future. The main components of smoking behaviour are the transitions among never, current and former smoking states, and individuals also transition from these smoking states to death. Alternative future scenarios of smoking behaviour might therefore be defined in terms of different future trends in the smoking state transition probabilities of initiation, quitting and relapse. There are a range of existing computer models that might allow such projections. Feirman et al (2015, 2017), Berg at al (2017) and Singh et al (2020) all carried out reviews of models that have been used to project future trends in smoking behaviour and to estimate the potential effects of new policies and interventions on smoking behaviour, and the associated health and economic outcomes. A large category of these models simulate the progression of smoking initiation, quitting and relapse over the entire life-courses of individuals in one-year time steps, and by simulating successive cohorts of individuals, make projections of the population-level trajectories of smoking rates. Models of this sort use a range of methods and have been developed for several geographic regions (e.g. the United Kingdom (Hunt et al. 2018; Song, Elwell-Sutton, and Naughton 2020), the Netherlands with the RIVM-Chronic Disease Model (Hoogenveen et al. 2008), the European Dynamo-HIA (DYNAmic MOdel for Health Impact Assessment) (Boshuizen et al. 2012), the United States (Tengs, Osgood, and Lin 2001; Mendez, Warner, and Courant 1998; Levy et al. 2016; Tam et al. 2018), New Zealand (Deen et al. 2014) and Australia (Gartner, Barendregt, and Hall 2009)).

However, a major difficulty in developing these models is how to use population survey data to inform the probabilities that individuals will transition between different smoking states, and the choice of methodology

is likely to have a strong influence on the population trajectories of smoking rates that are estimated by the model. One approach, employed by the RIVM-Chronic Disease Model, estimates smoking transition probabilities directly from a population survey that asks people about their current smoking status, about their smoking a year ago and, if they were a former smoker, when they had quit smoking (Capannesi et al. 2009). A second approach, employed by the Dynamo-HIA model, estimates 'net' smoking transition probabilities using just a single year of cross-sectional data, such that when the probabilities are applied, rates of smoking remain constant between years (Van de Kassteele et al. 2012). A third approach, employed by some models of smoking in the United States, estimates time-varying probabilities of smoking initiation and quitting from several years of cross-sectional survey data (Anderson et al. 2012; Holford et al. 2014; Tam et al. 2018). This third approach has strengths over the other approaches because it can estimate time trends in smoking state transition probabilities and makes use of widely available annual cross-sectional survey data. However, the method as currently published doesn't produce estimates of the probabilities of long-term relapse to smoking (i.e. relapse beyond one-year since quitting), likely due to the limitations of the cross-sectional data used. Despite the importance of information on relapse to smoking over several years since quitting (Stapleton, Sutherland, and Russell 1998), evidence is scarce (Etter and Stapleton 2006; Hughes, Peters, and Naud 2008), and the best opportunity to estimate the probabilities of smoking relapse is offered by survey data that tracks individuals over several years, e.g. the British Household Panel Survey (BHPS) (Hawkins, Hollingworth, and Campbell 2010). A second limitation of the method is that it doesn't adjust the estimates of smoking quitting to reduce the influence of bias that might result from age-related changes to the proportions of current and former smokers due to the elevated mortality associated with their smoking history (Christopoulou et al. 2011).

2.2 Programme of work

Our development of the Sheffield Tobacco Policy Model (STPM) is part of a wider programme of work that aims to provide decision-support to policymakers in the fields of alcohol and tobacco control, at a range of geographic scales, including national and local authority level. The two main objectives are:

1. To investigate past trends in alcohol and/or smoking behaviour and related health and economic outcomes, and to investigate past policy effects;
2. To investigate the potential future effects on alcohol and/or smoking behaviour and related health and economic outcomes of proposed changes to alcohol and/or tobacco policy or the introduction of new interventions.

3 The aim of this report

This report was written as part of our explanation of the methodology that underlies the STPM for England, an individual-based population microsimulation model. The intended audience for this report are fellow modellers using, adapting or building on the STPM methods, and users of the findings from STPM, who would like to understand the methods used to obtain the results. We describe the data sources and methods used to estimate smoking state transition probabilities from several years of cross-sectional survey data.

4 Methods

4.1 Approach to smoking state transition probability estimation

We build on the methodology developed by Anderson et al. (2012) and Holford et al. (2014) for the United States, which estimates age- and period-specific probabilities of smoking initiation and quitting for ages over 15 years using cross-sectional data for years 1965–2009 (see the detailed methods explanation in the appendix to Holford et al. (2014)). The main way we extend these methods is by adding adjustments to the method

for estimating the probabilities of quitting smoking - to account for the influences on the estimates of quit probabilities of smoking initiation, relapse to smoking, and smoking-related mortality. To do so, we use two sources of external data:

1. On the probabilities of relapse to smoking from Hawkins et al (2010);
2. On the probabilities of death according to smoking status.

The outputs of our estimation method are probabilities of smoking initiation, quitting smoking and relapse to smoking by calendar year (y), age (a), and strata defined by sex and socio-economic conditions (j). The method is designed to use a metric of socio-economic conditions that is age invariant (which means that our estimation method is simplified as we do not need to account for dynamic changes to socio-economic conditions over individual life-courses). The literature on the indirect estimation of transition probabilities from cross-sectional data talks about this method of stratification as grouping individuals into a set of 'pseudopanels', which are created from several years of cross-sectional data by grouping individuals according to similar time-invariant characteristics to form relatively homogenous cohorts (Verbeek and Vella 2005).

4.2 Approach to coding and reproducible research

We implement our analysis in the R environment (R Core Team 2020). To implement our calculations in a reproducible and version controlled way, we developed functions in the `smktrans` R package of code (Gillespie, Webster, and Brennan 2020) to estimate smoking state transition probabilities. Further explanation of our approach to coding and reproducible research can be found on the Sheffield Tobacco and Alcohol Policy Modelling website (https://stapm.gitlab.io/).

4.3 Mathematical framework

The method used to estimate smoking transition probabilities is based on the structural assumptions that underlie our STPM microsimulation model of the population dynamics of smoking. STPM models the movement of individuals among current, former and never smoking states as they age, and between these smoking states and death – in one-year time steps (see the STPM methodology report). At each time step, a new cohort of individuals aged a_{min} years are added to the population, and individuals aged a_{max} years are removed from the population. This structure allows us to model the population-level trends in the rates of smoking over time in calendar years y (e.g. from the year $y = 2001$ in one-year time steps) as a function of the dynamics of smoking over individual life-courses, i.e. the population-level trends are produced by the individual-level smoking transitions within successive birth cohorts. We stratify our modelling of the dynamics of smoking by sex and IMD quintile strata (i.e. 10 separate strata that we denote j).

4.4 Population and data

Our analysis relates to people aged from 11 to 89 years in England, with the lower age limit of 11 chosen because much of smoking initiation occurs when children begin secondary education at age 11 in England. Table 3 summarises our data inputs and processes.

4.4.1 Data on smoking

We used the Health Survey for England (HSE) 2001–2018 as our primary data source for developing this methodology (Mindell et al. 2012). The HSE data is processed using functions in the `hseclean` R package (Gillespie et al. 2020). Missing values in these variables are multiply imputed using the R package *mice* (Buuren and Groothuis-Oudshoorn 2010). From 2016 onwards, the version of the HSE available from the UK Data Service only reported respondent age in categories (predominantly 5 year intervals). For these years of data, we convert the categories to single years by randomly assigning individuals an age within each category.

The HSE data comes in the form of self-reports of whether a survey respondent is a current, former or never smoker, and with recalled information from current and former smokers on when they started and stopped smoking. We define smoking states as follows:

- A *never smoker* is someone who has never smoked or has only tried a cigarette once or twice in their lifetime.
- A *current smoker* is someone who smokes cigarettes either regularly or occasionally.
- A *former smoker* is someone who used to smoke cigarettes either regularly or occasionally. For former smokers we also track the number of years since quitting as discrete states.

For current and former smokers, the HSE contains data on the ages that individuals reported starting and stopping being smokers (providing a simplified and potentially biased view of what might be a complicated life history of smoking, e.g. someone might have smoked to different levels or started and stopped smoking multiple times between these ages).

We measure socio-economic conditions using quintiles of the English Index of Multiple Deprivation (IMD), a composite area-level measure based on 37 indicators reflecting income, employment, health and disability, education and skills, housing, services, accessibility, crime and living/physical environment. The IMD is calculated for small geographic areas in England of approximately 1,500 people. We divided IMD scores into quintiles, the first being the least deprived, and the fifth the most deprived.

4.4.2 Smoking relapse

The HSE does not contain information sufficient to estimate the probabilities of long-term relapse to smoking, e.g. it does not ask current smokers about previous failed quitting attempts. We overcome this by linking previously published estimates from the BHPS (Hawkins, Hollingworth, and Campbell 2010) into the HSE data, matching on education (degree or not), employment (employed or not), relationship status (married, cohabiting or neither), mental health (has condition or not) and income.

4.4.3 Mortality and associations between smoking and disease

We consider 52 ICD-10 defined categories of adult diseases related to smoking and the corresponding relative risks of these diseases in current vs. never smokers, and in former smokers according to the time since they quit (Webster et al. 2018). We calculated mortality rates from counts of death and population sizes obtained from the UK's Office for National Statistics (ONS), by calendar year (2001–2018), single years of age, and strata defined by sex and IMD quintiles.

4.4.4 Population data and cohort survivorship

Future population sizes are informed by the ONS's primary population projection for England. We also use information on historical mortality rates (prior to 2001) to inform the cohort survivorship functions, and we obtain these data from the Human Mortality Database (Barbieri et al. 2015).

Table 3: Overview of model inputs, processes and outputs.

Inputs	Source	Processes & Outputs
POPULATION DATA: Mid-year population sizes split by age, sex and IMD quintile for years 2001-2018.	Office for National Statistics	Used to calculate rates of mortality, which are used in the method that estimates the probabilities of quitting.
DATA ON SMOKING: Prevalence of smoker status (never smoker, former smoker and smoker) split by age, sex and IMD quintile for years 2001-2018. Also, data on the number of years since quitting for former smokers.	Health Survey for England (Mindell et al. 2012)	Used to estimate smoking initiation probabilities, which are then an input to the method that estimates probabilities of quitting. Also used to estimate the probabilities of survival by smoking status, which are an input to the method that estimates probabilities of quitting.
MORTALITY DATA: The number of deaths from 52 smoking related causes split by age, sex and IMD quintile for years 2001-2018.	Office for National Statistics	Used to estimate the probabilities of survival by smoking status, which are an input to the method that estimates probabilities of quitting. Also used in the estimation of the numbers of current, former and never smokers.
ASSOCIATIONS BETWEEN SMOKING AND DISEASE: Published estimates of the relative risks of disease in current and former smokers, considering the time since quitting.	Reviews of published literature on risks of smoking for adult disease (Webster et al. 2018)	Used to estimate the probabilities of survival by smoking status, which are an input to the method that estimates probabilities of quitting.
SMOKING RELAPSE: Published estimates of the probability of relapse to smoking up to 10 years since quitting, stratified by a range of social and demographic factors.	Hawkins *et al.* (2010), who analysed the British Household Panel Survey	Used to estimate smoking relapse probabilities. The estimated relapse probabilities are then an input to the method that estimates probabilities of quitting.
COHORT SURVIVORSHIP: Period death rates from 1922-2018, stratified by age and sex.	Human Mortality Database Human Mortality Database (Barbieri et al. 2015)	Used to estimate the numbers of current, former and never smokers, which are an input to the method that estimates probabilities of quitting.

4.5 Estimating smoking state transition probabilities

4.5.1 Smoking initiation

Smoking initiation probabilities are estimated using data on the ages at which individuals reported to have started smoking (data that are likely to be biased for a range of reasons, e.g. biased recall and selective mortality (Kenkel, Lillard, and Mathios 2004; Christopoulou et al. 2011)). The estimates are the probabilities $P(\text{initiate}|a,c,j)$ that an individual who is a never smoker at age a will subsequently be observed as a current smoker at age $a+1$. To estimate these probabilities from the HSE data, we first use the data on the ages that people started to smoke to calculate the cumulative probability $P(\text{ever-smoker}|a,c,j)$ that someone had initiated smoking by age a years (this gives us the age-pattern of increase in the proportion of people who have ever smoked by a certain age).

$$P(\text{ever-smoker}|a,c,j) = 1 - \prod_{a=0}^{a=35} [1 - P(\text{initiate}|a,c,j)]. \quad (1)$$

The main problem with the estimates from (1) is that they are subject to several biases in the survey data, and this causes a mismatch between the data in the HSE obtained by asking people of a certain age if they have ever smoked and the estimates from (1). A way to correct for this mismatch is to apply an adjustment to the estimates (1) (as done by Holford et al. (2014)). Our adjustment uses as its reference the proportion of people aged 30 years who report ever having smoked - we denote the reference values $P^*(\text{ever-smoker}|a,c,j)$.

To estimate the reference values, we fit a linear model to the cohort trend in the proportion of people who have ever smoked in the age band 25–34 years, which smooths the trends and allows extrapolation to future cohorts. We then calculate and apply an adjustment factor $k(a = 30, c, j)$ to the estimate so that, at the reference age of 30 years, the estimated and the reference value equal each other

$$k(a = 30, c, j) = \frac{P^*(\text{ever-smoker}|a, c, j)}{P(\text{ever-smoker}|a, c, j)}. \tag{2}$$

The adjusted proportions of people who have ever smoked are therefore $P(\text{ever-smoker}|a, c, j) \times k(a = 30, c, j)$, which we convert to probabilities of initiating smoking according to (1).

4.5.2 Relapse to smoking

We define relapse to smoking as the probability $P(\text{relapse}|a, y, \text{quityears}, j)$ that an individual aged a, who has been a former smoker for a number of years denoted by "quityears", transitions to being a current smoker at age $a + 1$ years. Because it is not possible to estimate probabilities of smoking relapse reliably from the HSE data, we instead use an external source of estimates of relapse to smoking from Hawkins et al. (2010), who analysed individual longitudinal data from the British Household Panel Survey. Hawkins' regression analysis adjusts for several covariates, which we use to link the estimates to the HSE dataset, including time since quit (1, 2, 3, 4, 5, 6+ years), age, sex, education (degree or not), employment (employed or not), relationship status (married, cohabiting or neither), mental health (has condition or not), income, and frequency of GP visits. Hawkins et al did not estimate trends in smoking relapse over calendar time, but when we link the estimates to the HSE data, variation in the above covariates between years of the survey data and between social strata will produce corresponding variation in the probabilities of relapse to smoking.

4.5.3 Quitting smoking

We define quitting smoking by the probability $P_c(\text{quit}|a, y, j)$ that someone who is a current smoker (indicated by subscript $_c$) at age a transitions to being a former smoker at age $a + 1$ years. The schematic in Figure 1 illustrates how we use our various data sources to estimate quit probabilities from the cross-sectional HSE data, with the addition of external sources of data on relapse to smoking and mortality. The mathematical framework for the estimation of quit probabilities comes from a re-arrangement of the formula that we use to track the change in the numbers of current smokers with age, $A_c(a, j)$ (see equation (2) in the STPM methodology report, pp. 10–11). Re-arranging this equation to have only $P_c(\text{quit}|a, j)$ on the left hand side gives

$$P_c(\text{quit}|a, j) = 1 - \frac{A_c(a + 1, j)}{A_c(a, j)P_c(\text{survive}|a, j)} \tag{3}$$

$$+ \frac{A_f(a, j)P_f(\text{survive}|a, j)P_f(\text{relapse}|a, j)}{A_c(a, j)P_c(\text{survive}|a, j)} \tag{4}$$

$$+ \frac{A_n(a, j)P_n(\text{survive}|a, j)P_n(\text{initiate}|a, j)}{A_c(a, j)P_c(\text{survive}|a, j)}. \tag{5}$$

- (3) is the **unadjusted quit probability**. It has the number of current smokers at age $a + 1$ years as its numerator, and the expected number of current smokers from age a years who will survive to age $a + 1$ years as its denominator. The ratio of these terms is the proportion of current smokers who remain current smokers, adjusted for survival. Therefore, one minus this proportion is the proportion of current smokers who do not remain current smokers, i.e. the proportion who quit.

- (4) is the **adjustment for relapse to smoking** by people who were former smokers at age a years. Without this adjustment, the quit probability would be under-estimated because not all of the current smokers at age $a+1$ years were current smokers at age a years. (4) therefore estimates the number of current smokers at age $a+1$ years who were actually former smokers at age a years and subtracts these from the numerator.

- (5) is the **adjustment for smoking initiation** by people who were never-smokers at age a years. As above, without subtracting these individuals from the numerator, the quit probability would be under-estimated.

4.5.4 Inputs into the estimation of the probabilities of quitting smoking

Table 3 gives an overview of our data inputs. Here we explain how those inputs are used in the estimation of the probabilities of quitting smoking.

(i) **Probabilities of smoking initiation**, $P_n(\text{initiate}|a,j)$ (see Section 4.5.1).

(ii) **Probabilities of smoking relapse**, $P_f(\text{relapse}|a,\text{quit-years},j)$ (see Section 4.5.2).

(iii) **Probabilities of survival by smoking status**. Estimates of $P_{\text{never, current or former}}(\text{survive}|a,j)$, the probabilities that individuals aged a years with a certain smoking status will survive to age $a+1$ years, might be obtained in a number of ways (e.g. estimated directly from linked smoking and mortality data). We apply an indirect method of estimation that uses HSE survey data on smoking, the associations between smoking and 52 diseases (Webster et al. 2018), and cause-specific rates of mortality, $m(h,a,j)$. Our method of indirectly estimating the probabilities of survival by smoking status is as follows: We first calculate the relative risks of each disease for each individual, according to their smoking state, $rr_i(h)$. Second, we standardise these individual relative risks so that they sum to one within each age, sex and IMD quintile category, $rr_i^*(h)$ (with categories corresponding to the stratification of mortality rates). Third, we calculate the mean of these standardised relative risks within each category, $\overline{rr^*}(h,a,j)$. Within each category, the mortality rate from each disease to which an individual i is exposed is then

$$m_i(h) = \frac{m(h,a,j)rr_i^*(h)}{\overline{rr^*}(h,a,j)}. \tag{6}$$

The $m_i(h)$ are converted to individual probabilities of death during a one-year interval, $P_i(\text{death},h)$, assuming that the rate of death increases exponentially over time within the age interval. The individual probabilities of death from all causes, $P_i(\text{death})$, are then calculated by summing the $P_i(\text{death},h)$ across causes. The corresponding individual probabilities of survival, $P_i(\text{survive}) = 1 - P_i(\text{death})$, are averaged for each age and smoking status to give $P_n(\text{survive}|a,j)$, $P_c(\text{survive}|a,j)$, and $P_f(\text{survive}|a,\text{quityears},j)$.

(iv) **Numbers of current, former and never smokers**. In describing the numbers of never, current and former smokers by age and cohort, $A_{\text{never, current or former}}(a,c,j)$, it is important to the accuracy of our calculation to minimise the influence of changes to the number of individuals due to immigration and emigration, and random survey sampling error. In an attempt to minimise these influences, we designed a method that breaks the estimation of the numbers of people in each smoking state into two parts, $A_{\text{never, current or former}}(a) = l(a)\theta_{\text{never, current or former}}(a)$. First, $l(a)$ is the cohort survivorship function (the probability that someone born in a particular year will survive to age a years). We estimate it using historic death rates from the Human Mortality Database (HMD), stratified by sex. We then add socio-economic stratification based on the socio-economic differentials in contemporary death rates. Second, $\theta_{\text{never, current or former}}(a,j)$ is the proportion of individuals aged a years in each smoking state (i.e. the age-specific proportions of never, current and former smokers). There are many possible methods that might be used to smooth the age and year trends in the proportions of current, former and never smokers calculated from the survey data. The current version of our method uses a multinomial linear regression model, stratified by sex and socio-economic conditions, that we initially parameterised as a 3rd order polynomial response surface, before simplifying the parameter structure

and adding an additional 4th order term to improve the description of age-patterns. However, different model fits can be explored.

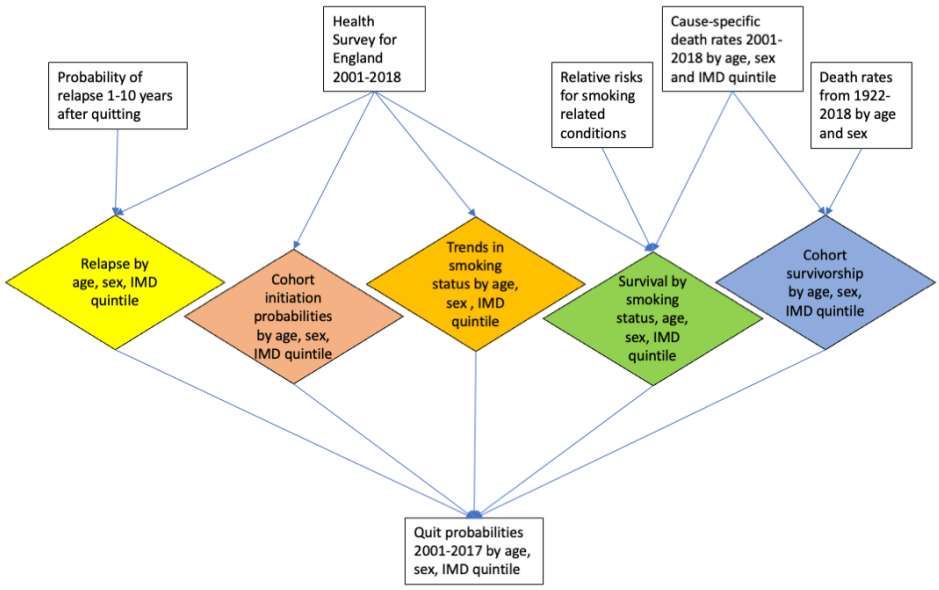

Figure 1: Empirical workflow to estimate quit probabilities. Each diamond is a function in the smktrans R package.

4.6 Forecasting smoking state transition probabilities

After estimating the smoking state transition probabilities of smoking initiation, relapse and quitting, we forecast the future trends in these transition probabilities separately for each of our 10 sex and IMD quintile subgroups. We could project them up to an extremely far time horizon (e.g. the year 2100) but we are conscious that the further into the future that projections are made, the less useful they are likely to be to current decision-making. Our projections of the probabilities of smoking initiation, relapse and quitting are made using three separate methods, designed to suit the way in which each of the three smoking state transition probabilities was estimated.

(i) Our projection of smoking initiation probabilities is based on extrapolating the linear trends in the proportion of people aged 25-34 years who have ever smoked (i.e. it is a cohort-based projection).

(ii) For smoking relapse probabilities, which rely heavily on the published analysis of Hawkins et al. (2010) and so have a less reliable historic time trend on which to base a projection, we assume that the future values of relapse remain constant at their average values over the last five years of survey data (2014–2018).

(iii) For the probabilities of quitting smoking, we apply a forecasting method based on the Lee-Carter approach (Lee and Carter 1992). Our method first smooths and logit transforms the age-period surface of the adjusted quit probabilities, and then based on a singular value decomposition estimates the overall rate of change over time and the age-emphasis of this change.

As with any forecast, the predicted future trends are sensitive to the time period of past trends that are used to inform the forecast, e.g. the predictions might vary depending on whether the observed trends over the past 5 vs. 10 years are used. The choice of time period might be constrained by the available data, or it might be affected by an understanding of relevant changes to society, e.g. quitting might be informed by trends since 2013, when e-cigarette use began to increase.

5 Discussion

The methodology that we present enables the use of repeat cross-sectional smoking survey data to estimate the smoking state transition probabilities of smoking initiation, quitting and relapse that underlie the population trends in smoking rates, stratified by sex and socio-economic conditions. The method can be applied to cross-sectional smoking survey data from any geographic region provided it has the appropriate smoking and socio-demographic variables. The significance of this methodology is that it facilitates the modelling of population trends in smoking rates as a function of the trends in the underlying smoking behaviours as characterised by initiation, quitting and relapse. As such, it enables modelling of the population-level impacts of policies or interventions that affect one or more of these smoking state transition probabilities.

To develop our methodology we drew heavily on the methods developed in the context of the United States by Mendez et al (1998), Anderson et al. (2012) and Holford et al. (2014), which led the way in showing how to estimate probabilities of smoking initiation and quitting from multiple years of cross-sectional smoking survey data. Holford et al. (2014) used their method to produce annual estimates of the probabilities of smoking initiation and quitting, stratified by age and cohort from age 15 years onwards. We followed Holford et al.'s method to estimate initiation probabilities, who used self-reported data on the ages at which people started smoking, and then corrected their estimates of initiation probabilities for biases by applying an adjustment-factor. We developed extensions to Holford et al.'s method to estimate probabilities of quitting, most notably by explicitly accounting for smoking relapse using published data from Hawkins et al. (2010) to inform the expected annual probabilities of relapse (Holford et al. account for smoking relapse by only recording people as former smokers if they had been quit for two years), and by accounting for differential mortality by smoking status using published data on the risks of smoking for adult diseases and cause-specific mortality rates. These methodological developments improve the accuracy of estimates of the probabilities of quitting smoking from cross-sectional survey data, and allow the use of the estimated probabilities of quitting within a microsimulation model of smoking that accurately projects population trajectories of smoking rates as a function of the long-term trends in the annual probabilities of smoking initiation, quitting and relapse.

Modelling the underlying probabilities of smoking initiation, quitting and relapse is also useful for evaluating the past effects of policies and interventions, and for predicting the potential effects of future policies and interventions. Notably there has been a recent wave of development individual-based microsimulation models of smoking, driven by the need to use modelling to understand the potential impact on smoking and health of the rise of e-cigarettes and related diversifications of the nicotine market (Cobb et al. 2015; Vugrin et al. 2015; Hill and Camacho 2017; Soneji et al. 2018). The challenge is to construct such models in a way that they can provide effective decision-support to policy-makers and effectively inform the societal debate on tobacco control policy. A first step in this is to have reliable estimates of how the trends in smoking rates over time are underpinned by trends in the probabilities of transitioning among smoking states.

The main limitations of our methodology stem from the fact that it attempts to infer smoking transition probabilities from cross-sectional data, rather than using data sources from which transition probabilities might be estimated directly. Due to the annual nature of the cross-sectional data that we used, we were constrained to using a one year time step and so our results are not detailed enough to allow us to model the micro-dynamics of individuals initiating, quitting and relapsing several times within the year. Our methodology also does not allow us to estimate how smoking behaviour might be affected by the dynamics of characteristics such as income, family relationships or mental and physical health (this more detailed modelling will likely have to be informed by specialist datasets and analytical methods such as agent based modelling). The adjustments for biases that we apply as part of our method are also unlikely to be perfect

and that means that our estimates remain vulnerable to several sources of error (such as errors in survey respondent recall that we were not able to correct for).

References

Anderson, Christy M, David M Burns, Kevin W Dodd, and Eric J Feuer. 2012. "Chapter 2: Birth-Cohort-Specific Estimates of Smoking Behaviors for the Us Population." *Risk Analysis: An International Journal* 32: S14–S24.

Barbieri, Magali, John R Wilmoth, Vladimir M Shkolnikov, Dana Glei, Domantas Jasilionis, Dmitri Jdanov, Carl Boe, Timothy Riffe, Pavel Grigoriev, and Celeste Winant. 2015. "Data Resource Profile: The Human Mortality Database (Hmd)." *International Journal of Epidemiology* 44 (5): 1549–56.

Berg, Marrit L, Kei Long Cheung, Mickaël Hiligsmann, Silvia Evers, Reina JA de Kinderen, Puttarin Kulchaitanaroaj, and Subhash Pokhrel. 2017. "Model-Based Economic Evaluations in Smoking Cessation and Their Transferability to New Contexts: A Systematic Review." *Addiction* 112 (6): 946–67.

Boshuizen, Hendriek C, Stefan K Lhachimi, Pieter HM van Baal, Rudolf T Hoogenveen, Henriette A Smit, Johan P Mackenbach, and Wilma J Nusselder. 2012. "The Dynamo-Hia Model: An Efficient Implementation of a Risk Factor/Chronic Disease Markov Model for Use in Health Impact Assessment (Hia)." *Demography* 49 (4): 1259–83.

Buuren, S van, and Karin Groothuis-Oudshoorn. 2010. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 1–68.

Capannesi, Maurizio, Hendriek C Boshuizen, Marc C Willemsen, and Hans C Van Houwelingen. 2009. "How to Obtain Long Term Projections for Smoking Behaviour: A Case Study in the Dutch Population." *Computational and Mathematical Methods in Medicine* 10 (2): 155–64.

Christopoulou, Rebekka, Jeffrey Han, Ahmed Jaber, and Dean R Lillard. 2011. "Dying for a Smoke: How Much Does Differential Mortality of Smokers Affect Estimated Life-Course Smoking Prevalence?" *Preventive Medicine* 52 (1): 66–70.

Cobb, Caroline O, Andrea C Villanti, Amanda L Graham, Jennifer L Pearson, Allison M Glasser, Jessica M Rath, Cassandra A Stanton, David T Levy, David B Abrams, and Raymond Niaura. 2015. "Markov Modeling to Estimate the Population Impact of Emerging Tobacco Products: A Proof-of-Concept Study." *Tobacco Regulatory Science* 1 (2): 129–41.

Deen, Frederieke S van der, Takayoshi Ikeda, Linda Cobiac, Nick Wilson, and Tony Blakely. 2014. "Projecting Future Smoking Prevalence to 2025 and Beyond in New Zealand Using Smoking Prevalence Data from the 2013 Census." *NZ Med J* 127 (1406): 71–79.

Doll, R., and A. B. Hill. 1954. "THE Mortality of Doctors in Relation to Their Smoking Habits - a Preliminary Report." Journal Article. *Bmj-British Medical Journal* 1 (4877): 1451–5. https://doi.org/10.1136/bmj.1.4877.1451.

———. 1964. "MORTALITY in Relation to Smoking - 10 Years Observations of British Doctors." Journal Article. *British Medical Journal* 1 (539): 1399 –&. https://doi.org/10.1136/bmj.1.5395.1399.

Etter, Jean-François, and John A Stapleton. 2006. "Nicotine Replacement Therapy for Long-Term Smoking Cessation: A Meta-Analysis." *Tobacco Control* 15 (4): 280–85.

Feirman, Shari P, Elisabeth Donaldson, Allison M Glasser, Jennifer L Pearson, Ray Niaura, Shyanika W Rose, David B Abrams, and Andrea C Villanti. 2015. "Mathematical Modeling in Tobacco Control Research: Initial Results from a Systematic Review." *Nicotine & Tobacco Research* 18 (3): 229–42.

Feirman, Shari P, Allison M Glasser, Shyanika Rose, Ray Niaura, David B Abrams, Lyubov Teplitskaya, and Andrea C Villanti. 2017. "Computational Models Used to Assess Us Tobacco Control Policies." *Nicotine & Tobacco Research* 19 (11): 1257–67.

Gartner, Coral E, Jan J Barendregt, and Wayne D Hall. 2009. "Predicting the Future Prevalence of Cigarette Smoking in Australia: How Low Can We Go and by When?" *Tobacco Control* 18 (3): 183–89.

Gillespie, Duncan, Laura Webster, and Alan Brennan. 2020. *smktrans: Smoking Initiation, Quit and Relapse Probabilities from Cross-Sectional Survey Data (R Package Version 1.1.0)*. School of Health & Related Research, University of Sheffield. https://stapm.gitlab.io/r-packages/smktrans/.

Gillespie, Duncan, Laura Webster, Damon Morris, Colin Angus, and Alan Brennan. 2020. *hseclean: Health Survey Data Wrangling (R Package Version 1.4.0)*. School of Health & Related Research, University of Sheffield. https://stapm.gitlab.io/r-packages/hseclean/.

Hawkins, James, William Hollingworth, and Rona Campbell. 2010. "Long-Term Smoking Relapse: A Study Using the British Household Panel Survey." *Nicotine & Tobacco Research* 12 (12): 1228–35.

Hill, Andrew, and Oscar M Camacho. 2017. "A System Dynamics Modelling Approach to Assess the Impact of Launching a New Nicotine Product on Population Health Outcomes." *Regulatory Toxicology and Pharmacology* 86: 265–78.

Hiscock, R., L. Bauld, A. Amos, and S. Platt. 2012. "Smoking and Socioeconomic Status in England: The Rise of the Never Smoker and the Disadvantaged Smoker." Journal Article. *Journal of Public Health* 34 (3): 390–96. https://doi.org/10.1093/pubmed/fds012.

Holford, Theodore R, David T Levy, Lisa A McKay, Lauren Clarke, Ben Racine, Rafael Meza, Stephanie Land, Jihyoun Jeon, and Eric J Feuer. 2014. "Patterns of Birth Cohort–Specific Smoking Histories, 1965–2009." *American Journal of Preventive Medicine* 46 (2): e31–e37.

Hoogenveen, Rudolf T, Pieter HM van Baal, Hendriek C Boshuizen, and Talitha L Feenstra. 2008. "Dynamic Effects of Smoking Cessation on Disease Incidence, Mortality and Quality of Life: The Role of Time Since Cessation." *Cost Effectiveness and Resource Allocation* 6 (1): 1.

Hughes, John R, Erica N Peters, and Shelly Naud. 2008. "Relapse to Smoking After 1 Year of Abstinence: A Meta-Analysis." *Addictive Behaviors* 33 (12): 1516–20.

Hunt, Daniel, André Knuchel-Takano, Abbygail Jaccard, Arti Bhimjiyani, Lise Retat, Chit Selvarajah, Katrina Brown, Laura L Webber, and Martin Brown. 2018. "Modelling the Implications of Reducing Smoking Prevalence: The Public Health and Economic Benefits of Achieving a 'Tobacco-Free'UK." *Tobacco Control* 27 (2): 129–35.

Kenkel, Donald S, Dean R Lillard, and Alan D Mathios. 2004. "Accounting for Misclassification Error in Retrospective Smoking Data." *Health Economics* 13 (10): 1031–44.

Lee, Ronald D, and Lawrence R Carter. 1992. "Modeling and Forecasting Us Mortality." *Journal of the American Statistical Association* 87 (419): 659–71.

Levy, David T, Rafael Meza, Yian Zhang, and Theodore R Holford. 2016. "Gauging the Effect of Us Tobacco Control Policies from 1965 Through 2014 Using Simsmoke." *American Journal of Preventive Medicine* 50 (4): 535–42.

Mendez, David, Kenneth E Warner, and Paul N Courant. 1998. "Has Smoking Cessation Ceased? Expected Trends in the Prevalence of Smoking in the United States." *American Journal of Epidemiology* 148 (3): 249–58.

Mindell, Jennifer, Jane P Biddulph, Vasant Hirani, Emanuel Stamatakis, Rachel Craig, Susan Nunn, and Nicola Shelton. 2012. "Cohort Profile: The Health Survey for England." *International Journal of Epidemiology* 41 (6): 1585–93.

NHS Digital. 2019. "Statistics on Smoking: England."

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Royal College of Physicians. 1962. "Smoking and Health. A Report on Smoking in Relation to Lung Cancer and Other Diseases." London.

Secretary of State for Health. 1998. "Smoking kills: A white paper on tobacco." Generic. The Stationery Office London.

Singh, Ankur, Nick Wilson, and Tony Blakely. 2020. "Simulating Future Public Health Benefits of Tobacco Control Interventions: A Systematic Review of Models." *Tobacco Control*. https://doi.org/http://dx.doi.org/10.1136/tobaccocontrol-2019-055425.

Soneji, Samir S, Hai-Yen Sung, Brian A Primack, John P Pierce, and James D Sargent. 2018. "Quantifying Population-Level Health Benefits and Harms of E-Cigarette Use in the United States." *PLoS One* 13 (3).

Song, Fujian, Tim Elwell-Sutton, and Felix Naughton. 2020. "Impact of the Nhs Stop Smoking Services on Smoking Prevalence in England: A Simulation Modelling Evaluation." *Tobacco Control* 29 (2): 200–206.

Stapleton, John A, Gay Sutherland, and Michael AH Russell. 1998. "How Much Does Relapse After One Year Erode Effectiveness of Smoking Cessation Treatments? Long Term Follow up of Randomised Trial of Nicotine Nasal Spray." *Bmj* 316 (7134): 830–31.

Tam, Jamie, David T Levy, Jihyoun Jeon, John Clarke, Scott Gilkeson, Tim Hall, Eric J Feuer, Theodore R Holford, and Rafael Meza. 2018. "Projecting the Effects of Tobacco Control Policies in the Usa Through Microsimulation: A Study Protocol." *BMJ Open* 8 (3): e019169.

Tengs, Tammy O, Nathaniel D Osgood, and Ting H Lin. 2001. "Public Health Impact of Changes in Smoking Behavior: Results from the Tobacco Policy Model." *Medical Care*, 1131–41.

Van de Kassteele, Jan, RT Hoogenveen, PM Engelfriet, PHM Van Baal, and HC Boshuizen. 2012. "Estimating Net Transition Probabilities from Cross-Sectional Data with Application to Risk Factors in Chronic Disease Modeling." *Statistics in Medicine* 31 (6): 533–43.

Verbeek, Marno, and Francis Vella. 2005. "Estimating Dynamic Models from Repeated Cross-Sections." *Journal of Econometrics* 127 (1): 83–102.

Vugrin, Eric D, Brian L Rostron, Stephen J Verzi, Nancy S Brodsky, Theresa J Brown, Conrad J Choiniere, Blair N Coleman, Antonio Paredes, and Benjamin J Apelberg. 2015. "Modeling the Potential Effects of New Tobacco Products and Policies: A Dynamic Population Model for Multiple Product Use and Harm." *PloS One* 10 (3).

Webster, Laura, Colin Angus, Alan Brennan, and Duncan Gillespie. 2018. "Smoking and the Risks of Adult Diseases." The University of Sheffield. https://doi.org/10.15131/shef.data.7411451.v1.