Notes on data files for PHIM project

Author : P N Lee
Date : 30th March 2016

Contents

1. Introduction

1.1. Fixed data files

Eight fixed data CSV files may be required for the PHIM Model Program. These eight files have now been prepared as follows:

File description, file name

Population file, POP.csv

Current smoking prevalence file, CSP.csv

Former smoking prevalence file, FSP.csv

Quit time distribution file, QTD.csv

Death file, MORT.csv

Relative risk file, RR.csv

Half-life file, H.csv

Absolute risk in never smoker file, AR.csv

These versions of the files are intended to be adequate for testing purposes, but are subject to limitations as discussed below in sections 2 to 9, which describe the contents of each file, the data sources used, and the extent of checking carried out.

1.2. Other files

Four additional csv files are required when running the PHIM Model Program. Versions of these files have now been prepared as follows, the first two being assumptions files, and the other two controlling the running of the program:

File description, file name

Smoking transition probabilities file for the null scenario, STP_NULL.csv

Smoking transition probabilities file for the MRTP scenario, STP_MRTP.csv

Options control file, OPT.csv

Output choices file, OUTC.csv

Some comments on these files are given in sections 10 to 13.

Note that there is an additional control file used for the stand-alone module CALC_RISK_2.SAS called CR_Control.csv. Information on this module is given in the PHIM Users Guide.

1.3 Abbreviations

Abbreviations used below are as follows:

Country – AT (Austria), CA (Canada), FR (France), DE (Germany), HU (Hungary), IT (Italy), JA (Japan), PL (Poland), SE (Sweden), CH (Switzerland), GB (United Kingdom) and US (United States).

Disease – LC (lung cancer), IHD (ischaemic heart disease), STR (stroke) and COPD (chronic obstructive pulmonary disease).

Data Sources – ISO (International Standards Organization), ISS (International Smoking Statistics), NHIS (National Health Interview Survey), UN (United Nations), WHO (World Health Organization).

2. Population (POP) file

2.1. Datasets

There are two datasets – WHO and UN.

2.2. File layout

The file (lines 2-21589) is sorted by dataset (2), then by country (12), sex (2), year (33: 1980, 1981 … 2012), and age group (14: 10-14, 15-19, … 75-79).

2.3. Sources

All data comes from recent downloads (12-Dec 2014 for WHO and 19-Dec 2014 for UN) from the WHO and UN websites.

2.4. Actual and projected data

No estimations or projections have been made to the data since download. The UN data had projections made before publication and therefore there are no missing data. The WHO data had missing values for the following countries:

Country	Missing years
Canada	2006-2012
Switzerland	2011-2012
France	2011-2012
United Kingdom	2011-2012
Italy	2005, 2012
Japan	2012
United States	2008-2012

2.5. File creation

A SAS file was created to import the data from the WHO and UN websites as they were downloaded. This was a single file from WHO and two files (males and females) from UN. Note that one change was made to the data from UN as it was in Excel format and the cells had to be adjusted to show more decimal places before being imported into SAS. The SAS file combined the data from the two sources, dropped data out of required ranges, sorted what was left and output a CSV file of all the population data.

2.6. Checking

SAS files were created to take in data files and to produce an output CSV file in the required format. This was compared to the CSV file created based on recently downloaded data by using comparison software (UltraEdit). Non-matching results were found to be due to updates to the data on the WHO website. These were for Austria 1982-1999 and 2001-2010 and France for 2008-2010. United Kingdom also had differences for 2000-2010 due to the 2013 download using the sum of England, Scotland, Wales and Northern Ireland and the 2014 download using given United Kingdom values. These differences were checked and it was decided that there were no differences of any consequence.

3. Current smoking prevalence (CSP) file

3.1. Datasets

There is only one dataset – ISS.

3.2. File layout

The file (lines 2-2297) is sorted by country (12), then by sex (2), period (6, 7 or 8; 1980, 1981-1985, 1986-1990, 1991-1995, 1996-2000 and then either 1996-2012, 2001-2012 [CA, GE, GB, US], 2001-2005, 2006-2012 [AT, FR, IT, JA, SE, CH], or 2001-2005, 2006-2010, 2006-2012 [HU, PL]) and age group (14; 10-14 to 75-79).

3.3. Sources

All the data come from the latest update (10-Jul 2014) of Supplement 1 to International Smoking Statistics "Estimation of sex-specific smoking statistics by standardized age groups and time periods" on the P.N. Lee Statistics and Computing Ltd (PNLSC) website: www.pnlee.co.uk.

3.4. Comments on data included

Product smoked All tobacco products

Country Data are available for all of the 12 countries. Estimates for Germany are for West Germany up to 1990 and for Unified Germany afterwards.

Year Data for 1980 entered are those presented for 1976-1980.

For Canada, Germany, Poland, UK and USA, the source data go up to 2001-2005, with estimates for this final period taken to apply to 2001-2012, so that there are 6 relevant periods in the dataset.

For Austria, France, Italy, Japan, Sweden and Switzerland, the source data go up to 2006-2010, with estimates for this final period taken to apply to 2006-2012, so that there are 7 relevant periods.

For Hungary, the source data go up to 2011-2015, with estimates for this final period taken to apply to 2011-2012, so that there are 8 relevant periods.

Age Prevalence has been entered as zero for all countries for age 10-14, the source data giving results for age groups from 15-19 to 75-79.

At this stage more recent current smoking prevalence estimates are not available on Supplement 1, which is periodically updated.

3.5. Creation of the file and checking

The data file was created from the Excel file on the PNLSC website, replacing the country name with the corresponding ISO country code, removing years and ages outside the required range and reorganising the prevalence data into one column. The creation was carried out initially in Excel, with the result converted into CSV format, and subsequently, for checking purposes, by macros in SAS. It was clear from a comparison of the two versions of the CSV file that they produced identical answers.

4. Former smoking prevalence (FSP) file

4.1. Datasets

There is only one dataset at present – ISS.

4.2. File layout

The file (lines 2-1205) is sorted by country (7), then by sex (2), period (6 or 7; 1980, 1981-1985, 1986-1990, 1951-1995, 1996-2000 and then either 2001-2012 [AT, CA, DE, JA, GP, US] or 2001-2005, 2006-2012 [IT]) and age group (14; 10-14 to 75-79).

4.3. Sources

All the data come from Table 3 of a report (Forey and Lee, 2012), the report describing in detail the sources of these estimates.

4.4. Comments on data included

Country Data are only available for 7 of the 12 countries. Estimates for Germany (West Germany to 2000, Unified afterwards) have been taken as applicable to the other 5 countries (France, Hungary, Poland, Sweden and Switzerland).

Year Data for 1980 entered are those presented for 1976-1980. For Austria, data entered for 2001-2012 are as those presented for 1996-2000. For Canada, Germany, Japan, UK and USA data entered for 2001-2012 are as those presented for 2001-2005. For Italy data entered for 2006-2012 are as those presented for 2006-2010.

Age Prevalence has been entered as zero for all countries for age 10-14, for Japan for age 15-19, and also for Italy for age 15-19 up to 1985. In all cases estimates were not available from the source.

4.5. Checking

The file created was 100% checked by hand by PNL against the source report.

5. Quit time distribution (QTD) file

5.1. Dataset

There is only one dataset – NHIS2006

5.2. File layout

The file (lines 2-27) contains data only for the US (National Health Interview Survey – www.cdc.gov/nchs/hnis.htm) only and for 2006, and is taken as applicable to all other countries and for the whole of the time period 1980-2012. It is sorted by sex (2) then age group (13; 10-14 to 75-79). Data are given for age 10-14 as, even though quitting is not allowed for this age group, this file has to match the age formats used in the other files.

5.3. File creation

Firstly, data on population estimates by sex and age group were taken from a PNLSC file (N:\RLMETA\Dynamic simulation\Dynamic simulation_v2.xlsx) derived from the 2006 National Health Interview Survey. Then, based on midpoint estimates for years quit and age, cells were corrected to zero if the estimated age of quit was less than 18. For each age group, the distribution (as percentages) was then estimated from the populations to one decimal place. Where the totals did not add to 100%, the largest percentages were corrected up or down by 0.1 as appropriate.

5.4. Checking

The file created was 100% checked by hand by PNL against the derived distributions.

6. Death (MORT) file

6.1. Datasets

There is only one dataset – "Dec-14" (from WHO).

6.2. File layout

The file (lines 2-52361) is sorted by country (12), then by sex (2), year (33: 1981, 1982, … 2012), disease (5; all causes, COPD, IHD. Lung cancer, stroke), and age group (14; 10-14, 15-19, … 75-79).

6.3. Sources

The file derives from a combination of WHO data up to 2010 previously downloaded on 13 September 2013, for which some missing values were estimated (see section 6.4), and WHO data up to 2012 which were downloaded on 14 December 2014.

Note that data for age groups 10-29 were added on 23rd February 2016 to enable the PHIM system to calculate estimates of absolute risk in never smokers, which are used in the Calc_Risk_2 SAS module associated with the main PHIM system.

6.4. Data gaps

In the previously downloaded Sep-13 file some gaps in the data were estimated using linear interpolation and extrapolation, as described in the table below.

Country	Missing mortality data	Method used to estimate
Canada	2010	Linear extrapolation from 2005-2009
France	2010	Linear extrapolation from 2005-2009
Italy	2004, 2005	Linear interpolation using data for 1999-2003 and 2006-2010
Poland	1997, 1998	Linear interpolation using data for 1992-1996 and 1999-2003

6.5. Missing data

There were no 2012 data for Canada, Switzerland, France, Italy, Japan, and United States, and no 2011 data for Switzerland and United States. No attempt has been made to estimate these missing values.

6.6. File creation

Two SAS files were created to import the two versions of the WHO data and output the data in CSV files in the required format. A third SAS file combined these two CSV files by keeping the new data where it was available and adding in any extrapolations or previously available data from the old data file. This combined data was then exported into another CSV file.

6.7. Checking

The two import SAS file outputs were compared to each other using comparison software (UltraEdit). This found that 2014 data had values for 2010-2012 which had been estimated in the 2013 version. The 2014 data had missing values which had also been estimated for the 2013 version (Italy 2004-2005, Poland 1997-1998, Germany 1980-1990).

The combined output was compared to the output from the 2014 data and the only differences were where gaps had been in the original download.

7. Relative risk (RR) file

7.1. Datasets

There is only one – PNLEST.

7.2. File layout

The file (lines 2-3273) is sorted by disease (4), then by country (12), sex (2), year (3; 1980-1989, 1990-1999, 2000-2012), and age group (14; 10-14 to 75-79). Exceptionally for age 10-29 RRs are only given for age 1980-2012 combined.

7.3. Sources

The sources used for the random-effects RR estimates for current smoking are:

Lung cancer – Lee et al., 2012a; Table 8 (All lung cancer)

Ischaemic heart disease (IHD) and stroke – Lee, 2010; Table 2

Chronic obstructive pulmonary disease (COPD) – Forey et al., 2011; Table 7 (COPD)

All these analyses give variation in RR for the factors of interest (country, sex, year and age group) individually, but not in combination. RRs have therefore been entered according to the factor most affecting the RR.

For IHD and stroke this was clearly age group, so the estimates cited for the age groups <54, 55-64, 65-74, 75+ have been used.

For lung cancer and COPD the most dominant factor was location, so the RRs entered are country specific, with the estimate taken for the relevant location. For example, estimates for North America have been applied to Canada and USA.

Note that the RRs for age 10-29 were set to be the same as those for age 30-34. It is convenient to have data for these ages as it facilitates the calculation of an individual's excess risk for older ages. However, the actual values entered for age 10-29 will have no effect on the calculated excess risks for ages 30+.

7.4. Checking

The file created was 100% checked by hand by PNL against the data in the source publications.

8. Half-life (H) file

8.1. Datasets

There is only one – PNLEST.

8.2. File layout

The file (lines 2-113) is sorted by disease (4), then by sex (2) and age group (14; 10-14 to 75-79).

8.3. Sources:

Lung cancer – Fry et al., 2013; Table 6

Ischaemic heart disease (IHD) – Lee et al., 2012b; Table 5

Stroke – Lee et al., 2014a; Table 5

Chronic obstructive pulmonary disease (COPD) – Lee et al., 2014b; Table 4

All these analyses give variation to H for the factors of interest (sex and age group) individually, but not in combination.

For stroke and COPD there was no evidence of variation in H by age group or sex so the overall estimate has been used for all sex/age combinations.

For IHD there was clear evidence of variation by age group, but not by sex, so the estimates cited for the age groups <50, 50-59, 60-69, 70+ have been used.

For lung cancer there was clear evidence of variation by both age group and sex, but estimates were not available jointly by age group and sex, and some estimates were for sexes combined. The variation by sex, which is not that great, has been ignored and the estimates cited for the age groups <50, 50-59, 60-69, 70+ have been used.

Note that H values for age 10-29 were set to be the same as those for age 30-34.

Note also that though estimates of H were derived from studies of quitting, the values of H are also used to determine changes in relative risk following reduction in exposure, initiation or re-initiation.

8.4. Checking

The file created was 100% checked by hand by PNL against the data in the source publications.

9. Absolute risk in never smokers (AR) file

9.1. Datasets

There is only one dataset – P4_Appendix 1_Basic analyses

9.2. File layout

The file (lines 2-2761) is sorted by sex (2), disease (4), year (30; 1980, 1981, … 2009) and age group (14; 10-14 to 75-79). Exceptionally, for years 1985-1989 data are only from ages 15-19, for years 1990-1994 from ages 20-24, for years 1995-1999 from ages 25-29, for years 2000-2004 from ages 30-34, and for years 2005-2009 from ages 35-39. The file currently only includes data for US.

9.3. Sources

Absolute risks for an individual can be estimated by multiplying never smoker risks for a given age, sex, year and disease by the corresponding estimated RR for the individual. Absolute risks for never smokers are not readily available in the literature, so were estimated by PHIM based on the Null Scenario and a run of 100,000 males and 100,000 females over the period 1980-2009, with the fixed age of starting to smoke taken to be 16 years. For each combination of age, sex, disease and year (in the given range), data are available on RM (the estimated mean RR) and on A (the absolute overall US mortality rate, which is calculated from the fixed files POP.CSV and MORT.CSV referred to in sections 2 and 6 respectively). The absolute mortality rate for never smokers is then estimated as A/RM, with the saved results for males and for females combined into the single file AR.csv.

9.4. Comment

At present the data are not used as input to PHIM, but to an additional SAS module, CALC_RISK_2, which allows estimation (and plotting) of RRs and absolute risk for individuals with specified histories of tobacco use. However the file is available for possible future use in PHIM.

9.5. Checking

Initially, results in an AR.CSV file produced with the fixed age of starting to smoke taken as 14 years were checked back against the original WHO data. For eight selected sex, year and disease specific sets of age-specific estimates (4 for each sex, 2 for each year) PNL extracted corresponding data on deaths and populations from the fixed data files, and then estimated overall population mortality rates. Dividing these by the estimated rates for never smokers then gave age-specific estimates of X, the average RR for the population. From base case scenario output (for a sex and disease) of the total number of deaths, N, and the total number attributable to smoking, NA, PNL then calculated an alternative all-age estimate of X, by Y = N/(N-NA). For each block, the age-specific estimates of X were compared with the all age estimate of Y, to ensure that the values of X were plausible. Thus, for the first block, male COPD, where Y was estimated as 2.37, the estimates of X rose from values just over 1.00 at younger ages (consistent with the short period of smoking) to a maximum of 2.78 at age 45-49 and then fell (consistent with less current and more former smokers) to 2.29 at age 75-79. Since the great majority of male COPD deaths were at age 70+, there seemed no reason to believe there might be errors in the estimates of X. For the other seven blocks the estimated values of X again seemed consistent with the value of Y. Note that as the RR calculation based on individual smoking histories had already been checked in the validation of PHIM, it was not considered necessary to check the whole of the smoking histories and RR calculations for the 100,000 individuals in the run generating the absolute risks in never smokers, and PNL considered that the estimates of absolute risk in never smokers were valid. Fuller details are given in the file CHECK_AR.CSV and the associated note "Checking of absolute rate for never smokers.docx."

Subsequently, when the AR.CSV file was re-created assuming the fixed age of starting to smoke was 16 rather than 14 years, PNL extended the file CHECK_AR.CSV to include the new estimates of absolute risks in never smokers, and recalculated X. The estimated absolute risks in never smokers were very similar in the two runs, as were the estimates of X. Further detailed checking was considered unnecessary.

10. STP_NULL file

10.1. Datasets

There is only one dataset– GUESSNULL1

10.2. File layout

The file (lines 2-29) is sorted by period of follow-up (months) (2; 1-3 and 4+ months) and then by age group (14; 10-14, 15-19 …… 75-79).

10.3. Sources

In January 2015 estimates were derived of the following monthly smoking transition probabilities (STPs).

Initiation rate: the probability, P_{NC}, that someone who has never smoked (N) becomes a cigarette smoker (C) .

Quitting rate: the probability, P_{CF}, that a cigarette smoker becomes a former smoker (F).

Re-initiation rate: the probability, P_{FC}, that a former smoker becomes a cigarette smoker.

Note that these rates only refer to the smoking of conventional cigarettes and that it is assumed that only one transition is possible in a month. Thus the STP from never smoker to former smoker (P_{NF}) is assumed to be zero.

These STPs are educated guesses reflecting age-specific patterns of cigarette initiation, quitting and restarting which seemed plausible and which produced not unreasonable estimates of prevalences of current and former smoking at different ages. These were used in the initial testing of the program.

Subsequently, attention was drawn to a recent paper (Weinberger et al., 2014) which described results of a study in which interviews were carried out 3 years apart in a representative sample of the US, and gave a table from which rates of initiation, quitting and restarting could be calculated. Converting these rates (which were not age or sex specific) to monthly rates and comparing them with those that we had used

suggested that though the rates of initiation and quitting that we had used seemed reasonable, the rates of re-initiation that we had used were clearly too low. Accordingly we multiplied our re-initiation rates by a factor of 2.4 to try to bring them into line. These revised rates will be used in the simulation and sensitivity analyses, and produced not unreasonable estimates of prevalences of current and former smoking at different ages as will be demonstrated in the report on these analyses.

10.4. The rates

The revised values are given, expressed as monthly rates per million to avoid decimals, in the table below. The actual data in the file are given as the actual rates. Note that they do not vary by period of follow-up or sex. The initiation rate, 2000 per million per month at age 10-14, rises to a maximum of 3500 at age 15-19 and then falls steadily, being assumed to be zero at ages 35+. The quitting rate is low initially, starting at 500 per million per month at age 10-14, and then rising, being assumed a constant 2000 between ages 20-24 and 50-54, before rising again to 4000 at age 75-79. Re-initiation rates are assumed to be 48% of quitting rates.

Period of follow-up	Age	Initiation P_{NC}	Quitting P_{CF}	Re-initiation P_{FC}
1-3	10-14	2000	500	240
	15-19	3500	1500	720
	20-24	2000	2000	960
	25-29	1000	2000	960
	30-34	500	2000	960
	35-39	0	2000	960
	40-44	0	2000	960
	45-49	0	2000	960
	50-54	0	2000	960
	55-59	0	2500	1200
	60-64	0	2500	1200
	65-69	0	3000	1440
	70-74	0	3500	1680
	75-79	0	4000	1920
4+	10-14	2000	500	240
	15-19	3500	1500	720
	20-24	2000	2000	960
	25-29	1000	2000	960
	30-34	500	2000	960
	35-39	0	2000	960
	40-44	0	2000	960
	45-49	0	2000	960
	50-54	0	2000	960
	55-59	0	2500	1200
	60-64	0	2500	1200
	65-69	0	3000	1440
	70-74	0	3500	1680
	75-79	0	4000	1920

11. STP-MRTP file

11.1. Datasets

The dataset used for the base case scenario will be GUESSMRTP9

11.2. File layout

GUESSMRTP9 (lines 2-29) is sorted by period range (2; 1-24 and 25+ months) and then by age group (14; 10-14, 15-19 ….. 75-79).

11.3. Sources

There are 15 STPs in the MRTP Scenario, based on transitions between five states – never smoked (N), current smoker of conventional cigarettes only (C), current smoker of MRTP only (M), current dual user (D), and former smoker (F). Three STPs (P_{NC}, P_{NM}, P_{ND}) relate to initiation, three (P_{CF}, P_{MF}, P_{DF}) to quitting, three (P_{FC}, P_{FD}, P_{FM}) to re-initiation, and six (P_{CM}, P_{CD}, P_{MC}, P_{MD}, P_{DC}, P_{DM}) to switching within the current smoking group.

Initial estimates of the STPs were educated guesses, though they were revised for various reasons. First, initial estimates of re-initiation rates were found to be low, as judged by results from a recent paper (Weinberger et al., 2014), so were revised upwards. Second, output generated using earlier STP estimates suggested that they overestimated the likely uptake of PMI's MRTP IQOS. GUESSMRTP9 was developed so that, 10 years after MRTP introduction, current smokers would be distributed approximately 84% CC, 10% MRTP and 6% dual use, considered to be a plausible level of uptake. Third, the rates were designed to reflect the fact that younger people were considered less likely than older people to initiate with or switch to IQOS, because of cost consideration.

To ensure comparability with the rates on the STP_NULL file various constraints were applied:

1. Initiation rates. The sum of the three STPs for initiation should be equal to that for P_{NC} in the NULL Scenario. As for the NULL Scenario, the initiation rates are age-dependent, rising to a maximum for age 15-19 years and then falling to age 30-34 years, subsequently being zero.

2. Re-initiation rates. The sum of the three STPs for re-initiation should be equal to that for P_{FC} in the NULL Scenario. The three STPs for re-initiation (P_{FC}, P_{FM}, P_{FD}) are in the ratio 60:20:20 for 1-24 months after MRTP introduction, and 40:40:20 for 25+ months after. As for the NULL Scenario, re-initiation rates rise with age.
3. Quitting. The three STPs for quitting (P_{CF}, P_{MF}, P_{DF}) should be equal to that for P_{CF} in the NULL Scenario, so do not vary by time of follow-up. As for the NULL Scenario, quitting rates rise with age.
4. Switching. The six STPs which relate to switching between current smoking status do not vary by period of follow-up, and only two of them (P_{CM} and P_{CD}) vary by age, rising from age 10-14 to age 25-29 and then being constant.

11.4. The rates

The rates shown in the table below are expressed as monthly rates per million and are independent of sex.

Period of follow-up	Age	Initiation			Quitting			Re-initiation		
		P_{NC}	P_{NM}	P_{ND}	P_{CF}	P_{MF}	P_{DF}	P_{FC}	P_{FM}	P_{FD}
1-24	10-14	1840	80	80	500	500	500	144	48	48
	15-19	2940	280	280	1500	1500	1500	432	144	144
	20-24	1520	240	240	2000	2000	2000	576	192	192
	25-29	680	160	160	2000	2000	2000	576	192	192
	30-34	300	100	100	2000	2000	2000	576	192	192
	35-39	0	0	0	2000	2000	2000	576	192	192
	40-44	0	0	0	2000	2000	2000	576	192	192
	45-49	0	0	0	2000	2000	2000	576	192	192
	50-54	0	0	0	2000	2000	2000	576	192	192
	55-59	0	0	0	2500	2500	2500	720	240	240
	60-64	0	0	0	2500	2500	2500	720	240	240
	65-69	0	0	0	3000	3000	3000	864	288	288
	70-74	0	0	0	3500	3500	3500	1008	336	336
	75-79	0	0	0	4000	4000	4000	1152	384	384
25+	10-14	1680	160	160	500	500	500	96	96	48
	15-19	2380	560	560	1500	1500	1500	288	288	144
	20-24	1040	480	480	2000	2000	2000	384	384	192
	25-29	360	320	320	2000	2000	2000	384	384	192
	30-34	100	200	200	2000	2000	2000	384	384	192
	35-39	0	0	0	2000	2000	2000	384	384	192
	40-44	0	0	0	2000	2000	2000	384	384	192
	45-49	0	0	0	2000	2000	2000	384	384	192
	50-54	0	0	0	2000	2000	2000	384	384	192
	55-59	0	0	0	2500	2500	2500	480	480	240
	60-64	0	0	0	2500	2500	2500	480	480	240
	65-69	0	0	0	3000	3000	3000	576	576	288
	70-74	0	0	0	3500	3500	3500	672	672	336
	75-79	0	0	0	4000	4000	4000	768	768	384

Period of follow-up	Age	Switching within current smoking groups					
		P_{CM}	P_{CD}	P_{MC}	P_{MD}	P_{DC}	P_{DM}
1-24	10-14	150	30	400	400	400	400
	15-19	200	40	400	400	400	400
	20-24	300	60	400	400	400	400
	25-29	450	90	400	400	400	400
	30-34	450	90	400	400	400	400
	35-39	450	90	400	400	400	400
	40-44	450	90	400	400	400	400
	45-49	450	90	400	400	400	400
	50-54	450	90	400	400	400	400
	55-59	450	90	400	400	400	400
	60-64	450	90	400	400	400	400
	65-69	450	90	400	400	400	400
	70-74	450	90	400	400	400	400
	75-79	450	90	400	400	400	400
25+	10-14	150	30	400	400	400	400
	15-19	200	40	400	400	400	400
	20-24	300	60	400	400	400	400
	25-29	450	90	400	400	400	400
	30-34	450	90	400	400	400	400
	35-39	450	90	400	400	400	400
	40-44	450	90	400	400	400	400
	45-49	450	90	400	400	400	400
	50-54	450	90	400	400	400	400
	55-59	450	90	400	400	400	400
	60-64	450	90	400	400	400	400
	65-69	450	90	400	400	400	400
	70-74	450	90	400	400	400	400
	75-79	450	90	400	400	400	400

12. Options Control File for Main PHIM

The Options Control file describes the analyses to be performed by the PHIM system, and will vary depending on the analysis. It is a comma separated file with an initial heading record followed by multiple records each defining one of the options for a run. Each record defines the options for a run or sequence of runs. For some options (indicated by * below), multiple entries in quotes and separated by a comma imply that separate runs are to be conducted based on each value in the list. Where there is more than one set of multiple entries, separate runs will be carried out for each combination of the values. For other parameters, alternatives can be specified in different records. The structure of a record is summarized briefly below, and described in more detail in section 10 of the PHIM Users Guide.

The legitimate entries for each record are given below.

VARIABLE PARAMETERS

1.* Country – e.g. "US". Based on the ISO short code for countries.

2.* Sex – "M" (Male), "F" (Female)

3. Year of start of process - An integer in the range 1980-2012.

4. Number of months of follow-up - An integer from 1 to the maximum allowable months of follow-up, noting that follow-up cannot continue past December 31st 2012.

5. Follow-up interval length (in months) - 1, 3, 6 or 12.

6. Lower age for risk estimation (in years) - 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70 or 75.

7. Upper age for risk estimation (in years) - 34, 39, 44, 49, 54, 59, 64, 69, 74 or 79.

The upper age must be greater than the lower age.

8.* The effective dose for MRTP only (F) - Decimal value in the range 0 to 1.

9. The effective dose for dual use (G) - Decimal value in the range 0 to 2, or the text "FROMF" which indicates the value (1+F)/2.

10.* STP factor 1 - Decimal value $\geq$1.0. STP Factor 1 is used to multiply the smoking transition probability P_{MC} by for someone who has previously been a current conventional cigarette smoker. (See also items 39 and 40 for the associated STP factor 4).

11.* STP factor 2 - Decimal value ≥1.0. STP Factor 2 is used to multiply the STPs P_{CF}, P_{MF} and P_{DF} by for someone who has previously quit smoking. This applies to the NULL and MRTP Scenarios.

12.* STP factor 3 - Decimal value ≥ 1.0. STP Factor 3 is used to multiply the STPs P_{FC}, P_{FM} and P_{FD} by for a short-term quitter of smoking. This applies to the NULL and MRTP Scenarios.

13. Definition of short-term - Integer between 0 and 10. This defines the upper limit of years of quitting smoking for which an individual is regarded as a short-term quitter and hence have STP factor 3 applied.

14. Number in population to be simulated - Integer between 10,000 and 100,000.

15. Number of simulations - Integer between 1 and 100.

16. The random number seed for the first simulation - Integer from 0 to 2^{31}-1.

For items 17 to 21 the texts that can be entered are currently only those shown, as described in sections 2 to 6 of this note.

17. Source for the population file (POP) - "UN" or "WHO".

18. Source for the current smoking prevalence file (CSP) – "ISS".

19. Source for the former smoking prevalence file (FSP) – "ISS".

20. Source for the quit time distribution file (QTD) – "NHIS2006".

21. Source for the death file (MORT) – "Dec-14"

For items 22 and 23, the text that can be entered corresponds to the data described in sections 7 and 8 of this note. Alternative files have been set up with different texts for use in the sensitivity analyses.

22. Source for the relative risk file (RR) – "PNLEST"

23. Source for the half-life file (H) – "PNLEST"

STP ASSUMPTIONS

For items 24 and 25, the text that can be entered corresponds to the data described in sections 9 and 10 of this note. Alternative files have been set up with different texts for use in the sensitivity analyses.

24. STP file for the Null Scenario (STP_NULL) – "GUESSNULL1"

25. STP file for the MRTP Scenario (STP_MRTP) – "GUESSMRTP9"

OUTPUT REQUIRED

26. Output choice source – "BASIC" or other text code. See section 12.
27. Output 1 - "Y", "N" or "-". Indicates whether the first output for development testing purposes from the P-Component is required. "-" is acceptable when only the E-component is being run.
28. Output 2 - "Y", "N" or "-". Indicates whether the second output for development testing purposes from the P-Component is required. "-" is acceptable when only the E-component is being run.
29. Output 3 - "Y", "N" or "-". Indicates whether the main output from the P-Component is required. "-" is acceptable when only the E-component is being run.
30. Output 4 - "Y", "N" or "-". Indicates whether the first output for development testing purposes from the E-Component is required. "-" is acceptable when only the P-component is being run.
31. Output 5 - "Y", "N" or "-". Indicates whether the second output for development testing purposes from the E-Component is required. "-" is acceptable when only the P-component is being run.
32. Output 6 - "Y", "N" or "-". Indicates whether the third output for development testing purposes from the E-Component is required. "-" is acceptable when only the P-component is being run.
33. Output 7 - "Y", "N" or "-". Indicates whether the main output from the E-Component is required. "-" is acceptable when only the P-component is being run.
34. Output file name – This is a text holding the name of the file to which the output is to be sent. No extension should be added to this file name, the output will be stored in RTF format. The text must be a legitimate name for a Windows file.

If the text is the same for successive records (or is left blank), the output will continue to be sent to that file.

RESULTS FILES

35. Results for P-Component – "NA". This field is not used.
36. Results for E-Component – "NA". This field is not used.

COMPONENTS

37. Components required – "P", "E" or "PE" depending on whether one or both of the two components are to be run.

38. P-Component results file for use in E-Component

    Where P or PE is selected as the components, this must be "-".

    Where E is selected, the directory name of the relevant results file should be entered.

EXTRA STP FACTOR 4

39. STP factor 4 - Decimal value $\geq$ 0.0. The STPs P_{MC} and P_{MD} STPs are multiplied by STP Factor 4 for a long-term user of MRTP. When this factor applies STP factor 1 will not also be applied. Note that this factor will be ignored if the parameter is left blank.

40. Definition of short-term user of MRTP - Integer between 0 and 10. This defines the upper limit of years of using MRTP for which an individual is regarded as a short-term user. Using MRTP for longer than this will results in STP factor 4 being applied. Note that STP factor 4 will be ignored if this parameter is left blank.

Where E only is the component selected, a number of the above entries (8-16, 18-20, 24, 25, 27-29, 35, 39, 40) become irrelevant. The user may enter these as a dash, though other legitimate entries may also be included.

Example: (Note that the rows and columns have been transposed in order to make it easier to read)

Country	US
Sex	M
Year at start	1982
Number of months of follow-up	96
Follow-up interval length (months)	12
Lower age for risk estimation	30
Upper age for risk estimation	59
Effective dose for MRTP only	0.2
Effective dose for dual use	FROMF
STP factor 1	1.8
STP factor 2	2.5
STP factor 3	3.2
Definition of short-term	2
Number in population to be simulated	10000

Number of simulations	1
Random number seed for the first simulation	15975263
Source for population file	UN
Source for current smoking prevalence file	ISS
Source for former smoking prevalence file	ISS
Source for quit time distribution file	NHIS2006
Source for death file	Dec-14
Source for relative risk file	PNLEST
Source for half-life file	PNLEST
Assumption set for STP file for Null Scenario	GUESSNULL1
Assumption set for STP file for MRTP Scenario	GUESSMRTP1
Output choice source	BASIC
Development test output P1	Y
Development test output P2	Y
Main output PNULL and PMRTP	Y
Development test output E1	Y
Development test output E2	Y
Development test output E3	Y
Main output E	Y
Output file name	Test_CF1
Results for P-Component	SASdataP_1
Results for E-Component	SASdataE_1
Components required	PE
P-Component results file for use in E-Component	-
STP factor 4	0
Definition of short-term use of MRTP	2

13. Output Choices Files

The file OUTC_SSA listed below consists of the basic set of choices for each type of output that have been used in the Simulation and Sensitivity Analyses. It is a CSV file which has an initial heading record, and consists of multiple records each giving the values of the output source, output type, user choice and value selected. The first set of records has source BASIC and give a set of default choices for each of the output types. Subsequent records with a given, different, source name can define alternatives for some or all of the output choices. Listed below are the default choices for source basic, followed by an explanation of the alternative choices for each output type.

Output source,Output type,User choice,Value
BASIC,DEVELOPMENT TEST OUTPUT P1,N INDIVIDUALS,20
BASIC,DEVELOPMENT TEST OUTPUT P2,N INDIVIDUALS,20
BASIC,MAIN OUTPUT PNULL,AGE GROUPS,ALSO BY 5-YEAR GROUPS
BASIC,MAIN OUTPUT PNULL,INTERVALS,5-YEARLY
BASIC,MAIN OUTPUT PNULL,PLOT,YES
BASIC,MAIN OUTPUT PNULL,SIMULATIONS,ALL COMBINED
BASIC,MAIN OUTPUT PNULL,STANDARD ERRORS,YES
BASIC,MAIN OUTPUT PMRTP,AGE GROUPS,ALSO BY 5-YEAR GROUPS
BASIC,MAIN OUTPUT PMRTP,INTERVALS,5-YEARLY
BASIC,MAIN OUTPUT PMRTP,PLOT,YES
BASIC,MAIN OUTPUT PMRTP,SIMULATIONS,ALL COMBINED
BASIC,MAIN OUTPUT PMRTP,STANDARD ERRORS,YES
BASIC,DEVELOPMENT TEST OUTPUT E1,N INDIVIDUALS,20
BASIC,DEVELOPMENT TEST OUTPUT E1,PLOT,YES
BASIC,DEVELOPMENT TEST OUTPUT E2,SIMULATIONS,ALL COMBINED
BASIC,DEVELOPMENT TEST OUTPUT E3,SIMULATIONS,ALL COMBINED
BASIC,MAIN OUTPUT E,YEARS,EACH YEAR
BASIC,MAIN OUTPUT E,AGES,COMBINED ONLY
BASIC,MAIN OUTPUT E,DEATHS OR RATES,BOTH
BASIC,MAIN OUTPUT E,DISEASES,ALSO FOR FOUR DISEASES

BASIC,MAIN OUTPUT E,ADJUSTMENTS FOR POPULATION SIZE,UNADJUSTED ONLY
BASIC,MAIN OUTPUT E,SIMULATIONS,ALL COMBINED
BASIC,MAIN OUTPUT E,CUMULATIVE,FOR TIME ONLY
BASIC,MAIN OUTPUT E,SMOKING GROUP,NO
BASIC,MAIN OUTPUT E,STANDARD ERRORS,YES

The possible alternative entries are given in the following:

Where the basic entry is YES, NO is the only valid alternative, and vice versa.

Where the entry is for N INDIVIDUALS, a positive integer must be entered that is no greater than the number of simulations.

Where the entry is for AGE GROUPS, alternatives are AGES COMBINED, ALSO BY 5-YEAR AGE GROUPS.

Where the entry is for EACH INTERVAL, alternatives are EACH INTERVAL, 5-YEARLY, FINAL YEAR ONLY

Where the entry is for SIMULATIONS, alternatives are EACH, ALL COMBINED, BOTH

In the E-Component output other alternatives are:

YEARS: EACH YEAR, 5-YEARLY, FINAL YEAR ONLY

AGES: COMBINED ONLY, COMBINED AND PREMATURE, COMBINED PREMATURE AND BY AGE

DEATHS OR RATES: DEATHS ONLY, RATES ONLY, BOTH

DISEASES: ALL CAUSE ONLY, ALSO FOR FOUR DISEASE

ADJUSTMENTS FOR POPULATION SIZE: UNADJUSTED ONLY, ADJUSTED ONLY, BOTH

CUMULATIVE: FOR TIME ONLY, CUMULATIVE, BOTH

14. References

Forey, B.A., Lee, P.N., 2012. A comparison of smoking prevalence and quitting between countries which use either Virginia or blended tobacco cigarettes. P N Lee Statistics and Computing Ltd, Sutton, Surrey. Available: www.pnlee.co.uk/Reports.htm [Download FOREY2012].

Forey, B.A., Thornton, A.J., Lee, P.N., 2011. Systematic review with meta-analysis of the epidemiological evidence relating smoking to COPD, chronic bronchitis and emphysema. BMC Pulm. Med. 11, 36.

Fry, J.S., Lee, P.N., Forey, B.A., Coombs, K.J., 2013. How rapidly does the excess risk of lung cancer decline following quitting smoking? A quantitative review using the negative exponential model. Regul. Toxicol. Pharmacol. 67, 13-26. DOI:10.1016/j.yrtph.2013.06.001.

Lee, P.N., 2010. Dynamic simulation of potential reduction of disease risk associated with a next generation product (NGP). Providing required data for several potential countries in order to perform NGP dynamic simulation. Milestone 4. Provide CVD relative risk data by country/region for current cigarette smokers by age and sex and for quitters by sex and time of quitting. pp. 28.

Lee, P.N., Forey, B.A., Coombs, K.J., 2012a. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. BMC Cancer. 12, 385. DOI:10.1186/1471-2407-12-385.

Lee, P.N., Fry, J.S., Forey, B.A., 2014a. Estimating the decline in excess risk of chronic obstructive pulmonary disease following quitting smoking - a systematic review based on the negative exponential model. Regul. Toxicol. Pharmacol. 68, 2, 231-239. DOI:10.1016/j.yrtph.2013.12.006.

Lee, P.N., Fry, J.S., Hamling, J.S., 2012b. Using the negative exponential distribution to quantitatively review the evidence on how rapidly the excess risk of ischaemic heart disease declines following quitting smoking. Regul. Toxicol. Pharmacol. 64, 51-67.

Lee, P.N., Fry, J.S., Thornton, A., 2014b. Estimating the decline in excess risk of cerebrovascular disease following quitting smoking - A systematic review based on the negative exponential model. Regul. Toxicol. Pharmacol. 68, 1, 85-95.

Weinberger, A.H., Pilver, C.E., Mazure, C.M., McKee, S.A., 2014. Stability of smoking status in the US population: a longitudinal investigation. Addiction. 109, 9, 1541-1553. DOI:10.1111/add.12647.