Risk factor modelling in PHE

DRAFT 0.1

Julian Flowers, Jürgen Schmidt, Brian Ferguson, Annali Belloni, Leoni Belsman, Jake Gommon,

Table of Contents

[Introduction 1](#_Toc492809198)

[Why model? 1](#_Toc492809199)

[A typology of models 3](#_Toc492809200)

[Risk factor estimates and models 3](#_Toc492809201)

[Scenario modeling 8](#_Toc492809202)

[Quality assurance and reproducibility 8](#_Toc492809203)

[Capacity and capability in PHE for NCD and risk factor modelling 9](#_Toc492809204)

[Summary 9](#_Toc492809205)

[Annex 11](#_Toc492809206)

[Some specifics 11](#_Toc492809207)

[Why model - the case for modelling? 12](#_Toc492809208)

[Some modelling principles 13](#_Toc492809209)

[Limited - see predictive modelling current activities in PHE 14](#_Toc492809210)

[Activities in PHE. 14](#_Toc492809211)

# Introduction

This paper update our previous paper (see annex) which described significant non-communicable disease modelling work being undertake in PHE. It is written on behalf of the ‘virtual modelling group’ (see annex 2). The paper focusses on 3 issues:

1. A reiteration of modelling of risk factor modelling (tobacco, obesity and so on)
2. A review of modelling capacity and capability for non-communicable disease in PHE
3. Recommendations about developing modeling in PHE

# Why model?

Models are simplifications of reality which allow us to describe, explain, simulate and predict how systems work. Population and public health systems are inherently complex interplays of individual, environmental and social factors so if we want to try and understand how for example changes in health behaviour may influence changes in health outcomes we inevitably need to construct some kind of model to help us understand what may happen - there is often not a simple linear relationship (for example, changes in smoking behaviour do no translate into concomitant changes in mortality outcomes because there is a more complex shift in the distribution and temporality of risk).

For PHE modelling is increasingly important. PHE is committed to make the case for prevention and support investment by local authorities and others to improve health and tackle health inequalities. To do this we need;

* To understand trends in risk factors and morbidity
* To forecast future levels of risk, disease, burden and outcome
* To use and develop evidence of effective interventions and their uptake to estimate their impact
* To evaluate the effect on cost and return on investment.

To do this we need public health and heath economic modelling and we need quantification.

Modelling can also help improve measurement of key indicators

PHE has prioritised tackling the risk factors with the largest contribution to disease burden or mortality. Some of these are clinical, some behavioural and some environmental.

* Smoking
* Alcohol consumption
* Overweight and obesity
* Diet (especially sugar)
* Underactivity
* Hypertension
* Atrial fibrillation
* Diabetes
* Air pollution
* Early years
* Mental ill health

Often they are related - for example, alcohol consumption increases the prevalence of hypertension, smoking and alcohol (independently and interactively) cause heart disease which underlies atrial fibrillation, overweight and underactivity contribute to diabetes and so on, so models which accounts for these relationships and interactions may be necessary but are inevitably more complex and data hungry.

### A typology of models

For the purposes of this paper we propose the following typology of models. We think this provides a useful framework for assessing our capacity and capability.

1. *Models developed by others* – these tend to be the most complex, and the product of considerable academic input over a period of time. This category includes agent-based simulation models like UKHF MidRif, complex statistical and machine learning models like the Global Burden of Disease, and deterministic spreadsheet models like the FOIH tool. It also includes “data gap filling” models like prevalence estimates which are intended to provide more timely or disaggregated estimates of disease or risk factors which can in turn be fed into other models.
   1. These are the mainstay of work on forecasting future health states and testing potential scenarios and interventions. They are developed by or with others and have been adopted and adapted by PHE.
   2. They often require significant end-user skill to both run and interpret the results, and from a PHE perspective are ideally co-produced with model developers to ensure assumptions are transparent and tested, and models are valid.
   3. Where we are the recipient of the output we need transparency of methods
   4. Increasingly when we commission models – especially prevalence estimates – we need to ensure reproducibility – that is the underlying data, assumptions, code and method.
2. *Statistical models* - this includes regression models, time series models and so on, and is more bread and butter analytical work but needs further development.
3. *Machine learning models* – this is a subset of statistical modelling and is at the core of data science. Machine learning is synonymous with predictive analytics. Work in PHE in this area is in its infancy. (For an example of the application of ML techniques see annex 3 using an ML algorithm called glmnet to predict area suicide rates and identify key population predictors of suicide rate.).
4. *Other mathematical models* - this includes the kinds of modelling done by the EDR team on environmental hazards for example of plumes caused by fires or toxic releases

### Risk factor estimates and models

The starting point for most work attempting to predict future health states is estimates of up-stream risk factors. We need to know 4 things:

1. Current prevalence appropriately stratified by time, place and person
2. Past trends
3. Projections of future trend for the counterfactual (i.e. no deliberative change)
4. Distributions of multiple risk factors, that is smoking *and* alcohol

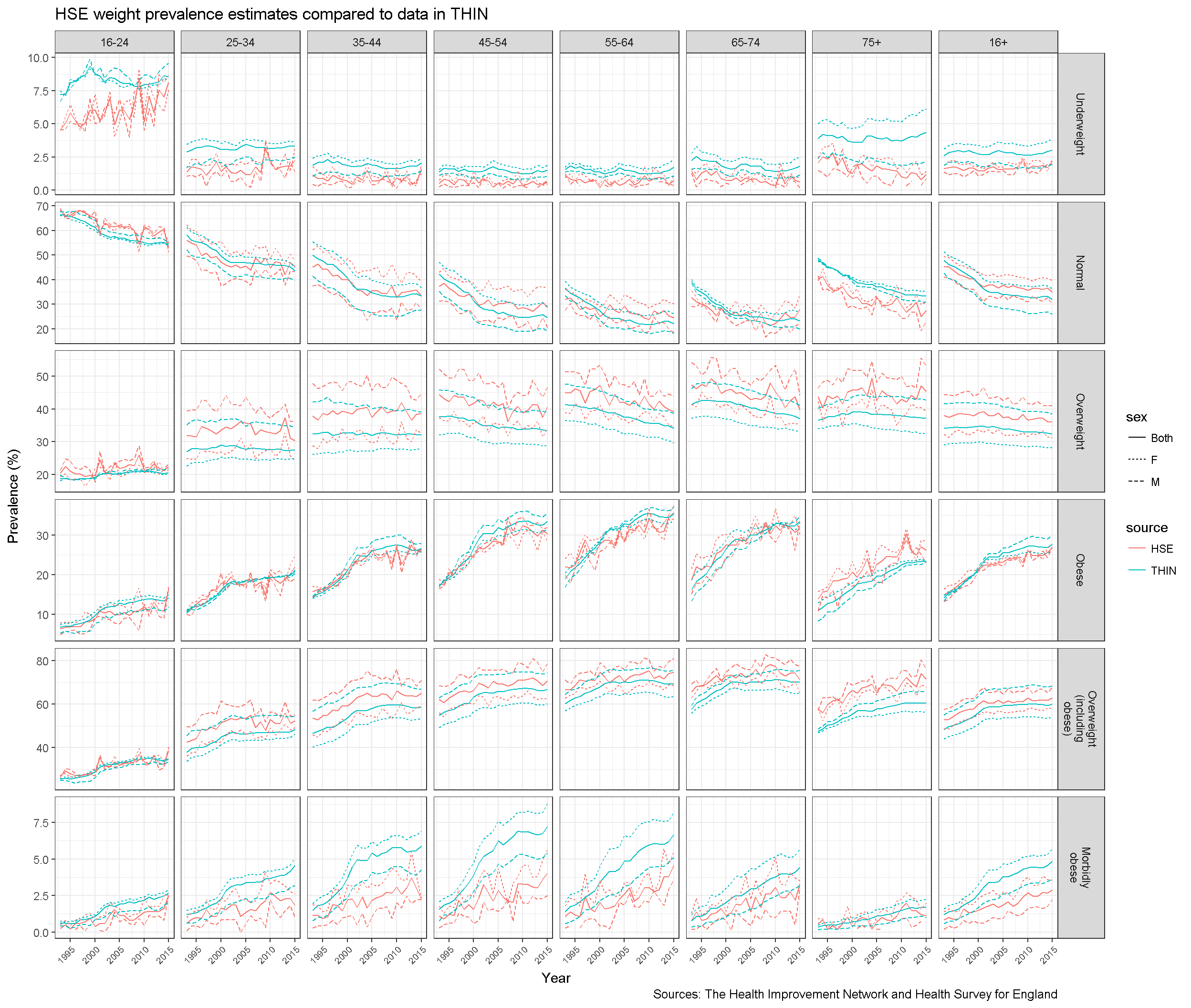
#### Current prevalence and past trends

Our current understanding of a range of contemporary risk factor prevalence and incidence, either directly measured or estimated through modelling, comes from a range of datasets and sources:

* Surveys
  + Health survey for England – Hypertension, disease prevalence models, smoking, obesity, diabetes, CKD
  + Annual population survey - smoking
  + Sport England Active People’s survey – diet, overweight, physical activity rates
  + What about Youth (WAY)
* Clinical datasets (direct and modelled estimates)
  + GP data – obesity, smoking, diabetes, pre-diabetes,
  + Audit data – AF, CKD
* Surveillance systems
  + NCMP
* Extrapolations from population based research
* GBD publishes a set of modelled and extrapolated risk factor prevalence estimates

For some sources we have historical trend data but this is not universal, and there have been a number of changes in national surveys which have created trend discontinuities in some estimates. For some major national surveys PHE is now the main commissioner.

PHE also holds one of the large GP research datasets (THIN) which has allowed us to use a ‘big data’ approach to prevalence estimation (at national level). For example, the THIN dataset contains some 36 million BMI measurements and gives estimates of obesity and overweight which compare favourably with estimates derived from the Health Survey for England but are more contemporary and have greater precision (see fig).



One challenge we have is that there is no single place where we store all the estimates from different sources, and we are aware that different modelling tools may use different inputs. This creates potential inconsistency and duplication.

As part of efforts to overcome this we have:

* Consolidated all the published prevalence estimates currently available in Fingertips into a single prevalence profile (<https://fingertips.phe.org.uk/profile/prevalence)>
* Posted all the 2013 GBD estimates as a large table in the ‘data lake’.

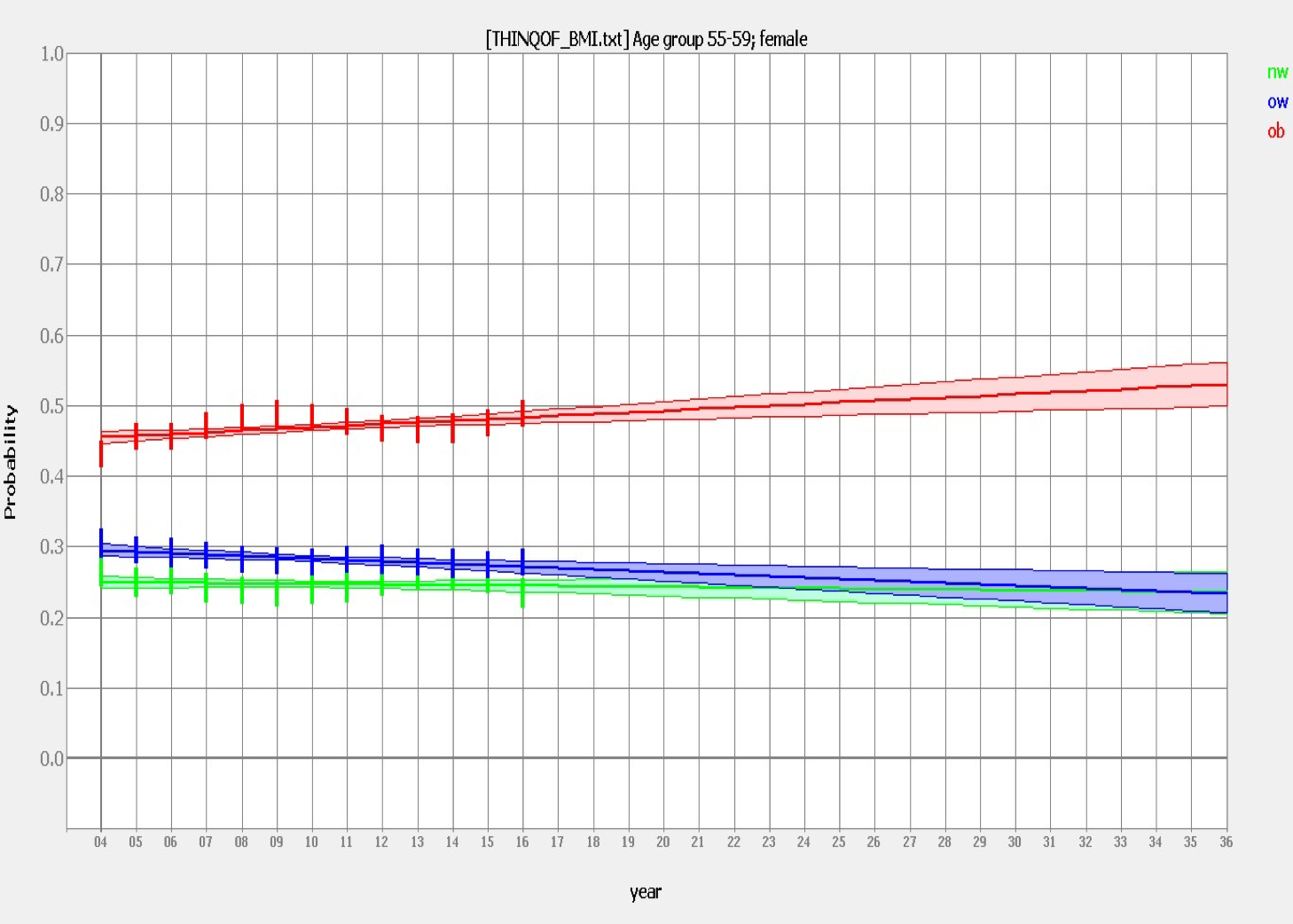
But there is more work to do.

PHE is also commissioning a suite of small area prevalence risk factor estimates (ward and LSOA) from Southampton University.

#### Projections

Our current approaches to projecting risk factor estimates are based on modelling. The first step of the UKHF tool projects individual risk factors up to 30 years ahead using a nonn-linera regression model and is a relatively easy task.

The figures below show obesity projections based no model fitting to 12 years of THIN from 2004 (when QOF began) to 2016 data for men and women aged 55-59. If trends continue as they are by 2035 60% of men and 55% of women in this age group will be obese.



#### 

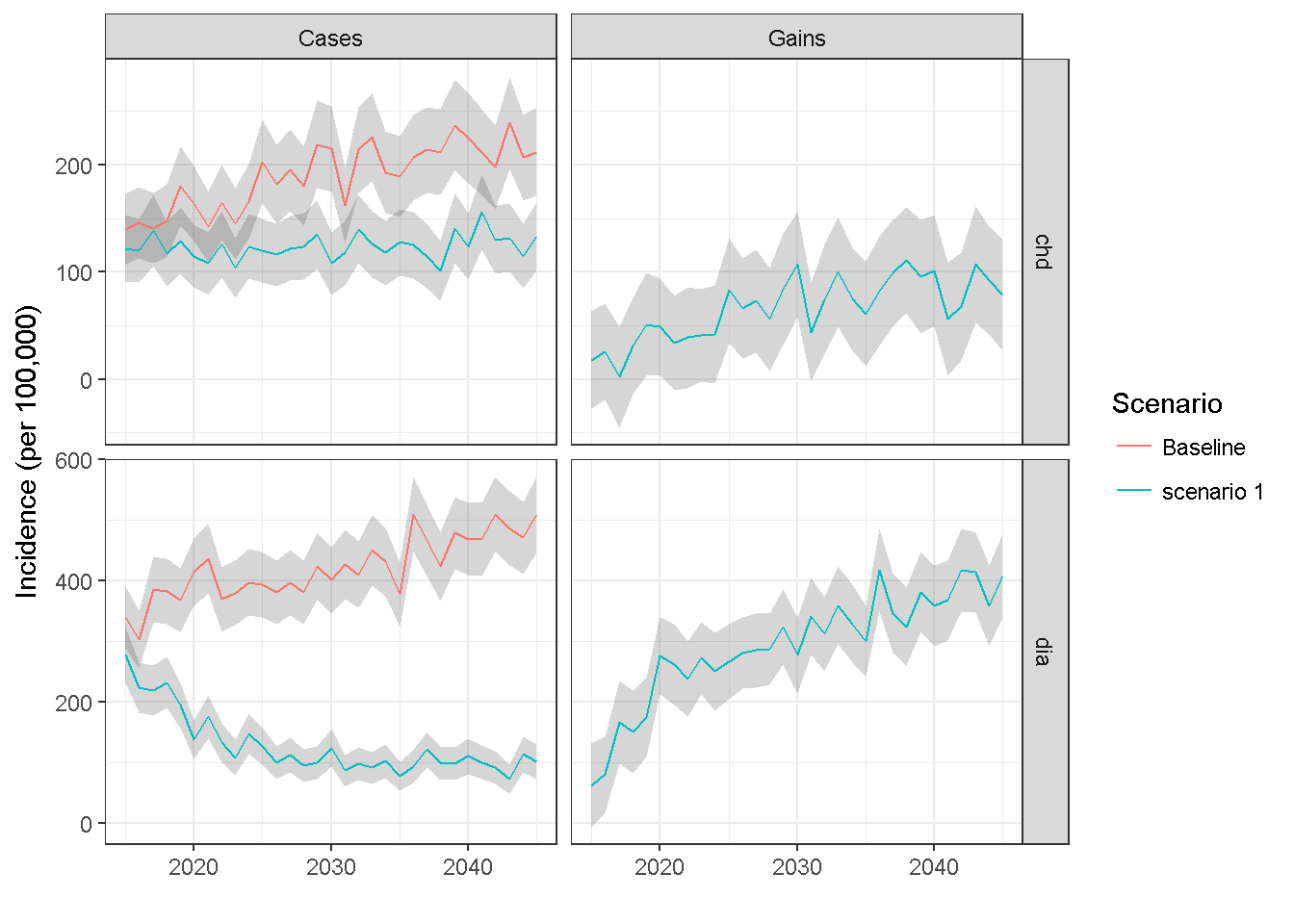
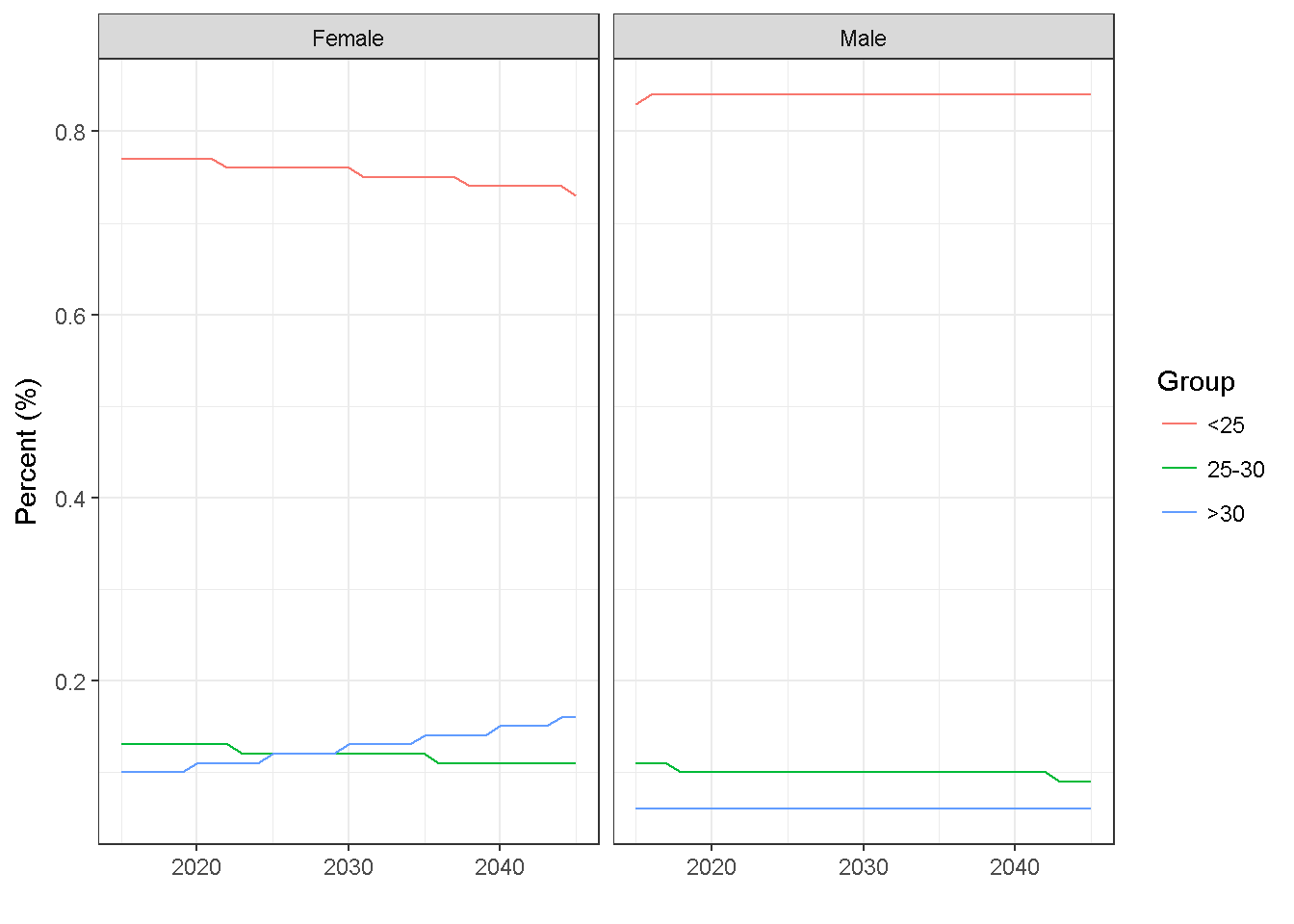
## 

#### FOIH model projections

To be completed

## Scenario modeling

The UKHF model and FOIH are the two tools available in house for PHE to develop scenarios and examine future health states and costs. Both are not without their issues.



As an example, the figure shows an illustrative small scale simulation (50,000 runs) of the MidRif model, comparing the possible impact of continued trend in obesity levels (left figure) on disease outcomes against a steady state counterfactual (right figure).

Given sufficient improvements in the model user interface and analytical capacity it would be possible to develop a range of scenarios which PHE can run itself, bearing in mind that the model is currently designed to run single rather than multiple risk factors.

[Something about FOIH DN]

## Quality assurance and reproducibility

If PHE is to become a major creator and commissioner of models it needs to comply with expected standards and quality assurance. Advice to government following the Macpherson review is encapsulate in guidance on Quality analysis - the so called [Aqua Book](https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government) which sets out assurance processes and reporting.

There is a developing range of guidance on good practice in modelling including:

* TRIPOD guidelines [ref]
* REPORT [ref]
* GATHER [ref]
* Brighton declaration [ref]

Finally, there is growing scientific consensus that analysis should be **reproducible** (to add refs and examples). This means the data, method, and code for the model should be published so that others can achieve the same answers. This requires full transparency and the ability to share data.

## Capacity and capability in PHE for NCD and risk factor modelling

A stock take of modelling activity across PHE in 2014 revealed consolidated capacity in communicable disease and environmental modelling but very little in non-communicable and risk factor modelling.

Our assessment is that this has not changed although there has been no systematic attempt to identify people within (particularly) the Knowledge and Intelligence service who have skills and interest in this area.

The Pubic Health Data Science teams dedicates about 0.5 wte data science capacity to modelling work of this nature – largely working to improve and be trained on the UKHF models so we can use them in house, and developing data to feed to the Global Burden of Disease programme. There is some exploratory work looking at the application of predictive analytics and machine learning and to developing data for reuse.

## Summary

PHE has a vital role to play in public health modelling to help central and local government invest appropriately in prevention and population health improvement.

PHE lacks a coherent modelling strategy for what is rapidly becoming core business. PHE does have access to a range of models of variable utility and quality, and although it has some capacity and capability to run and interpret models developed by others it has limited in-house development capacity. As a result it has relied on re-using models produced by others, or commissioning work from academics or private providers. There is inevitable duplication and inconsistency, for example different models coming to widely different conclusions, models using different input data and so on.

PHE needs to take advantage of the guidance on Quality Analysis provided by the [Aqua Book](https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government) to assure the quality of models PHE uses for its business requirements.

A strategic approach to modelling would read across to wider strategies on, for example:

* Data - quality assured and open data, building a more comprehensive and systematic population health information system to ensure we have access to all the data we need to deliver our remit
* Data science - developing the platforms, tools, capacity and skills to make the best use of data we collect or reuse
* Workforce development giving opportunities to analysts to develop modeling skills, run models
* Building partnerships with academic modelling units

## Annex

#### Types of model

There are many types of quantitative modelling which are relevant to the work of Public Health England and public health practice including:

* *System dynamic modelling* is a methodology and mathematical modeling technique to frame, understand, and discuss complex issues and problems. Originally developed in the 1950s to help corporate managers improve their understanding of industrial processes, SD is currently being used throughout the public and private sector for policy analysis and design
* *Predictive modelling*. Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. It is synonymous with (supervised) machine learning
* *Microsimulation* (from microanalytic simulation) is a category of computerized analytical tools that perform highly detailed analysis of activities such as highway traffic flowing through an intersection, financial transactions, or pathogens spreading disease through a population. Microsimulation is often used to evaluate the effects of proposed interventions before they are implemented in the real world. For example, a traffic microsimulation model could be used to evaluate the effectiveness of lengthening a turn lane at an intersection, and thus help decide whether it is worth spending money on actually lengthening the lane.
* An *agent-based model (ABM)* is one of a class of computational models for simulating the actions and interactions of autonomous agents (both individual or collective entities such as organizations or groups) with a view to assessing their effects on the system as a whole. It combines elements of game theory, complex systems, emergence, computational sociology, multi-agent systems, and evolutionary programming. Monte Carlo methods are used to introduce randomness.
* *Forecasting* is the process of making predictions of the future based on past and present data and most commonly by analysis of trends.
* *Econometric models* are statistical models used in econometrics. An econometric model specifies the statistical relationship that is believed to hold between the various economic quantities pertaining to a particular economic phenomenon under study. An econometric model can be derived from a deterministic economic model by allowing for uncertainty, or from an economic model which itself is stochastic.

#### Why model - the case for modelling?

PHE's remit on behalf of government is to improve and protect health and reduce health inequality. It acts through a range of levers and tools including influencing policy and practice, prioritisation, resource allocation, service design, intervention, behaviour change and so on. PHE needs to be able to systematically monitor current health, identify priorities for action, project and predict future health states, test scenarios and hypotheses - all of these may need modelling.

In public health practice and policy making we use or may need to models and modelling techniques for a range of purposes [@Webber14]:

* **Filling data gaps** - for example obtaining estimates of disease frequency or burden where direct measurement maybe difficult or too costly. For example, to monitor adult obesity as part of the Public Health Outcomes Framework (PHOF) we currently rely on an (expensive) survey. This does not provide data which is sufficiently granular or timely for our requirements. To fill the gap work on small area estimation of prevalence has been commissioned. There has been a limited amount of work on using existing sources and novel datasets to develop more timely estimates. We have developed national obesity estimates from a large sample GP dataset and cross checked these with the Health Survey for England.
* **Projecting future health states** - trying to understand current and potential future trends in risk factors, health determinants and disease outcomes. Some work has been done using the projection tool in the UKHF model to predict future rates of obesity by age. These have then been used to try and simulate future health states related to obesity such as diabetes, heart disease and overall life expectancy.
* **Testing interventions or policy options** - it can be very difficult to design controlled trials or 'gold standard' methodological design for evaluating interventions, so models can help in devising and testing scenarios, or altering levels of intervention, assessing impact and so on. Previous work was done using the UKHF model (see below) to test scenarios of obesity, salt consumption and smoking. The Future of Ill Health Model also provides tools to help model scenarios.
* **Understanding systems** - models (both qualitative and quantitative) can help us understand how systems *work* - the dynamics and interactions of complex systems - often present in public health problems and solutions and this may help in understanding why interventions may or may not work and devise new innovations
* **Testing and challenging assumptions** - all models are based on assumptions and sometimes these are not explicit and therefore not open to challenge. For example,
* **Understanding costs** of future health states and savings from intervention
* **Driving improvements in measuring health states and outcomes**. For example, if we build forecasts of future prevalence of adult obesity, we need a means of monitoring adult obesity rates which we can compare against our estimates, both to evaluate our model performance, and evaluate our efforts to tackle the problem.

#### Some modelling principles

There are some important principles:

1. Models are only as good as the data feeding them
2. Models are only as good as the underlying assumptions (implicit or explicit) that are built in to the model

# Briefing note: Modelling in Knowledge and Intelligence

This document provides a brief summary of public health modelling activity in PHE, but focuses mainly on modelling non communicable disease outcomes.

By modelling we mean a simplification of reality, often represented quantitatively, which helps make predictions, understand interactions or relationships, under a set of assumptions. Applying this approach to public health problems allows us to project future health states and test interventions or policies.(Webber et al., 2014)

NCD modelling provides methods for estimating the extent to which changes in one or more risk factors (e.g. smoking and diet) affects disease and health. It has two broad uses, health impact modelling (understanding the effect of prevention, screening or treatment interventions on health) and forecasting (estimating disease trends based on demographic change or predicted changes in risk factors). It considers the costs to society or health systems that arise from disease, and the potential savings that may accrue from interventions.

## Background

Modelling is undertaken at various levels of the organisation, such as for communicable diseases within Health Protection, or Health Checks within Health and Wellbeing, or indeed economic and population health models within the CKO Directorate. The reasons for modelling and the resources available (in terms of staff capacity and IT support) vary considerably. This note will address the current state of affairs within Knowledge and Intelligence.

A previous review (September 2014) identified a number of disparate modelling activities and outputs across (mainly) the CKO directorate: for example various prevalence models, a predictive model of future child health and wellbeing outcomes, synthetic drinking estimates and an obesity modelling tool. As a consequence it was agreed to bring together those modelling activities more coherently.

Although there are significant modelling activities across the whole of PHE, this note focuses on the main non- communicable disease modelling activities undertaken in the CKO Directorate. There are four key areas of work:

1. Future of Ill-Health model (FOIH)
2. Global Burden of Disease (GBD)
3. UK Health Forum Multiple Interacting Diseases and Risk Factors MIDRIF model
4. Disease prevalence models

## Modelling roles

### Role of the Health Economics Team

Working with partner organisations and stakeholders, PHE’s health economics team commissions through the Health Economics Commissioning Framework a series of projects that involve modelling aimed largely at supporting local commissioners to make the case for prevention and early intervention. An exception to this is the work that PHE’s previous Director of Strategy wished to see commissioned: namely to develop a national level model aimed at scenario modelling of health and the future costs of health and social care across the system. This was commissioned to address the ‘Office for Budgetary Responsibility (OBR) for Health’ role that PHE could potentially play. This modelling work has become known as the ‘Future of Ill-Health (FOIH) model.

The Health Economics team also quality assures and provides input on projects that are commissioned by other parts of PHE (e.g. Oral Health ROI tool) and the wider health system (for example, tools produced by NICE).

### Role of Public Health Data Science

Following Securing Our Future, which developed the Public Health Data Science (PHDS) team, some responsibility for modelling of non-communicable diseases passed to the function with the aim of “providing PHE with governance, capacity and purpose around modelling of non-communicable diseases”.

## Key NCD models

A brief synopsis of these key models is presented here with more detail in the annex.

### FOIH- Future levels of ill-health and impact on health services, expenditure and outcomes

PHE commissioned McKinsey to produce an interactive, user-friendly model to give a view, at a macro-level, on future levels of ill-health, and the impact of changes in risk factors (namely alcohol, obesity and tobacco) on disease prevalence, future health and social care spend and outcome measures. The project steering group included representatives from PHE (Health Economics, PHDS, Strategy and Well-being teams), Department of Health, NHS England, NHS Improvement and NICE – reflecting the main intended audience of national policy and decision makers across the health and social care system.

The model can produce the following outputs:

Epidemiology outputs

* Risk factor prevalence in years 2020, 2025 and 2035
* Disease prevalence (number of people) for 11 key diseases plus comorbidities in years 2020, 2025 and 2035

Health Economics Outputs

* Total (and by diseases) NHS and Social care spend in years 2020, 2025 and 2035
* DALYs for 11 key diseases plus comorbidities in years 2020, 2025 and 2035
* EQ5D score (per 100k) in years 2020,2025 and 2035

*What kind of questions can it answer?*

* What is the long term change in cost of ill health in the population?
* If we tackled a risk factor (obesity, smoking, alcohol) what would be the impact in terms of cost and quality?

### UKHF

PHE have acquired the licence from UK Health Forum to utilise a computer model developed by UKHF in-house. It simulates the lives and related medical costs of virtual people in the presence of “Multiple Interacting Diseases and Risk Factors” (MIDRIF model).

It can provide forecasts of the likely, dynamical health outcomes, costs, changes in QALYs, DALYs and life expectancies resulting from targeted health care and upstream interventions. The original project steering group comprised K&I (KIT East of England, KIT London, Health Economics), Strategy, Business Development, Finance and Commercial Directorate Office. The intended internal audience is policy makers/decision makers, and Health Economics, Health and Wellbeing, Strategy, Health Protection, PHE Centres. The external audience is Local Authority Public Health teams, local authorities, LGA, NHS England, OGD, in general health care providers and commissioners.

The intention is to ensure that the MidRif model

* is used by number of trained PHE super-users,
* is further developed to include multiple risk factors and multiple conditions,
* finds a place in policy and strategy development.

PHE collaborates with UKHF and others on a consensus on Modelling of Non-Communicable Diseases.

The model can produce the following outputs:

Epidemiology outputs

BMI Distributions: probability of rf-group by age and sex

Cumulative Incidence Cases by disease by year

Cumulative Incidence Gain by disease by year

Incidence Gain by disease by year

Incidence by disease by year

Prevalence by disease by year

Prevalence Gain by disease by year

Health Economics Outputs

Direct Costs by disease by year by scenario

Indirect Costs by disease by year by scenario

ICERs by disease by year (recap of all the variables used for the calculation)

Intervention Costs by disease by year

QALY by sex

*What kind of questions can it answer?*

* *What is the future population risk profile, given an assumed trend in one or more risk factors?*
* *What are future population health outcomes, based on above projections of risk factors?*
* *What Health and Wellbeing, and what economic savings, can be achieved given certain public health interventions?*

### Global Burden of Disease

GBD is a comparative assessment methodology, supporting the systematic effort to quantify the comparative magnitude of health loss due to diseases, injuries and risk factors by age, sex and geographies for specific points in time. Originally conceived within WHO at a much smaller scale (first report in 1993), the project was taken up by the University of Washington with the creation of the Institute of Health Metrics and Evaluation (IHME). 2010 saw the first GBD output. This method presents estimates of all-cause mortality, deaths by cause, Years of Life Lost (YLL), Years Lived with Disability (YLD), and Disability-Adjusted Life Years (DALY) by country, age, and sex. It consists of data from 188 countries from 1990-2013 and uses the currently highest life expectancy (86 years for Japanese women) as reference of for its YLL calculations. It is a global scientific collaboration involving some 1,000 researchers, 100 countries. The current revision of data covers 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, and 315 diseases and injuries and healthy life expectancy (HALE).

In 2015 the collaboration between PHE and IHME allowed an analysis of morbidity, mortality and risk factors specifically for England, covering twenty years and distinguishing by geography and levels of poverty. The study allows examining main risk factors, broader co-variates, outcomes, and their relationship across 45 subnational areas (9 regions by 5 deprivation levels). Data allow the description and initial analysis of variation of population outcomes in England, and their comparison to European and non-European countries. The 2016 edition (covering data up to 2015) will allow a breakdown of analysis at local authority level.

PHE are leading the way (internationally) in terms of developing methods for the analysis and inclusion of primary care data for Global Burden of Disease. All England GBD data produced so far are updated and available to IHME, integrated into the PHDS data lake, accessible to PHE users, and have started being used.

*What kind of questions can it answer?*

* *What are the main causes of health loss in a country or subnational area?*
* *Which causes are getting worse or are improving?*
* *How do causes compare between different areas and countries?*
* *Where is the greatest potential to reduce burden?*
* *What is the effect of deprivation and other risk factors on disease patterns?*

### Prevalence models

The frequency of disease in the UK population are really only enumerated with any completeness and accuracy for cancer – for the majority of other diseases (or risks), it needs to be estimated.

PHE have commissioned updates to a range of pre-PHE non-communicable disease models through an open tender process which was won by Imperial College (who developed the original estimates). The overall aim is to refresh or develop a set of web-based small population chronic disease and risk factor (hypertension) prevalence models in conjunction with NHSE.

These models are a series of prevalence estimates proportions stratified by age, sex and location. PHE also produces other disease prevalence models in house (e.g. diabetes) and by other routes (e.g. dementia).

The main objectives are:

1. To methodologically update and refresh risk factor data in Quality & Outcomes Framework (QOF)- compatible chronic disease prevalence models where they currently exist

2. To develop new QOF-compatible chronic disease prevalence models, for both registered and resident populations

3. To validate the models internally, and where possible externally.

For each model the expected outputs are GP practice, LA and CCG level estimates for prevalence value, numerator, denominator, and confidence intervals.

The non-communicable diseases modelled are:

* Hypertension (diagnosed/undiagnosed)
* CVD, estimated by combining
  + CHD
  + Stroke
  + Peripheral arterial disease separately
* Asthma
* COPD
* Depression
* Heart failure

The Public Health Data Science team will store the data in its Data Lake and store the relevant code (supplied in Stata and in R) in a GitHub repository in line with the concept of reproducible research <https://www.ctspedia.org/do/view/CTSpedia/ReproducibleResearchStandards>.

It is intended to publish these data via PHE tools and NHSE websites.

*What kind of questions can these models answer?*

* *What is the expected prevalence of a particular non-communicable disease at national, regional and local level?*
* *How does this prevalence compare to disease registers held at GP practice level?*
* *Which non-communicable diseases differ significantly in prevalence between regions and locations?*

## Fitting it all together

The suite of models outlined have had separate evolutions but we are increasingly bringing them together in a number of ways:

1. Estimates of disease and risk factor prevalence and frequency are essential inputs to FOIH, GBD and Midrif and we are trying to make the inputs coherent and consistent, and consolidate the necessary data.
2. Increasing our understanding of the ‘fit’ between the different modelling strands – for example, the GBD helps us understand the relative contribution of disease and risk factors to the burden of ill-health and may help us focus our preventive efforts or prioritisation which may in turn be supported by the FOIH or Midrif models for testing interventions or making projections
3. Sharing and testing assumptions
4. Cross-validating and refining models.
5. the FOIH and MIDRIF models are different in that the former was developed over a short period of time, has greater ease of use and can be used for rapid evaluations, whereas the latter is the product of years of development, requires considerable operator skill but is more powerful and flexible. Both however complement each other well.

## Quality assurance

If we are to use these models appropriately, they need to be quality assured in line with the Aqua Book and Macpherson review( HM Treasury, 2015). There are also emerging guidelines in the scientific literature on standards or conduct and reporting for NCD modelling work which PHE is involved in developing.(Collins, Reitsma, Altman, & Moons, 2015; Webber et al., 2014).

Public Health Data Science have produced a draft set of QA guidelines for NCD modelling which are being revised by key informants.

PHE has established a modelling strategy group to provide some focus and governance around modelling work.

## References

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. European Urology. http://doi.org/10.1016/j.eururo.2014.11.025

Treasury, H. (2015). The Aqua Book: London: HM Treasury.

Webber, L., Mytton, O. T., Briggs, A. D. M., Woodcock, J., Scarborough, P., McPherson, K., & Capewell, S. (2014). The Brighton declaration: the value of non-communicable disease modelling in population health sciences. European Journal of Epidemiology, 29(12), 867–870. http://doi.org/10.1007/s10654-014-9978-0

1. **Appendix: what models what?**

This is a brief synopsis of the main features of the models listed above.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelling tools |  | **FOIH** | **Prevalence** | **UKHF** | **GBD** |
| Model type |  | Excel spreadsheet | Logistic regression based | Microsimulations | Various |
| Multiple morbidity |  | x | - | planned | - |
| Interacting risk factors |  | x ('overlap') | - | planned | x ('overlap') |
| Data sources |  | Multiple | HSE; ONS; QOF; CPRD | Multiple | Cancer registry; HSE; ONS mortality; HES; QOF |
| PHE input |  | Data; evidence; staff time | monitoring | Data; evidence; staff time | data; staff time |
| Outputs |  | Count, Rate/ratio, DALY, EQ5D score, £ | Count; Rate / ratio | Rate/ ratio; Cost; LE | Count, rate/ratio; HLE; DALY |
| Granularity |  | national or local (user data input-for baseline scenarios only) | GP practice, LA and CCG level | national or local (user data input) | currently 45 sub-areas; in future: local authority |
| Uses |  | Policy; ROI | JSNA and similar | Policy, ROI | Policy, JSNA and similar |
| **Socioeconomic adjustment** | Deprivation | - | - | - | x |
|  | Ethnicity | - | - | - | - |
|  | Other | - | location | - | location |
| **Wider determinants** | Education | - | - | - | - |
|  | Housing | - | - | - | - |
|  | Environment | - | - | - | (x) |
|  | Employment | - | - | - | (x) |
|  | Workplace | - | - | - | - |
| **Risks** | Diet | - | - | - | x |
|  | Salt | - | - | x | x |
|  | BP | - | - | - | x |
|  | BMI | x | - | x | x |
|  | Glucose | - | - | - | x |
|  | Phys act | - | - | x | x |
|  | Multi | x | - | x | overlap |
|  | Alcohol | x | - | x | x |
|  | Tobacco | X (smokers and ex-smokers combined) | - | x | x |
|  | Air pollution | - | - | - | x |
|  | gfr | - | - | x | x |
| **Disease** | Cancer | x (lung, breast) | - | x (lung) | x |
|  | CVD | x (CHD) | x | x | x |
|  | Stroke | x | x | x | x |
|  | Hypertension | x | x | x | x |
|  | MSK | - | - | - | x |
|  | Diabetes | x | x | x | x |
|  | Respiratory | x (COPD) | x | x | x |
|  | Comm. disease | - | - | - | x |
|  | GI/liver | x | - | - | x |
|  | Dental | - | - | - | x |
|  | Infection | - | - | - | x |
|  | Injury | - | - | - | x |
|  | Kidney | - | - | x | x |
|  | Mental health | x (depression and other) | x (depression) | - | x |
|  | Dementia | x | x | - | x |
|  | Comorbidities | x |  |  |  |

All enquiries about FOIH model to Panos Zerdevas, Panos [Zerdevas@phe.gov.uk](mailto:Zerdevas@phe.gov.uk)  
All enquiries about UKHF, Global Burden of Disease, and Prevalence modelling to Jürgen Schmidt, [Jurgen.Schmidt@phe.gov.uk](mailto:Jurgen.Schmidt@phe.gov.uk)

1. Producing quantitative outputs is important - numerical scale matters (how big is the problem? What practical scope is there for change?)
2. Models should be as simple as they need to be but no simpler - if models are too complex they may be unstable or give unpredictable results or be too difficult to explain - if they are too simple they may miss important interactions which affect the results or the model may lack credibility
3. Models are representations not actualities - the results will be inherently uncertain even if presented as a point estimate - as far as possible results should be given with uncertainty estimates or sensitivity analysis
4. The best models should be a tradeoff between
   * parsimony (so we can understand the model),
   * accuracy (so we can trust the results),
   * interpretability (so we can understand the results)
   * and be implementable (so we can act on the results)
5. Models should behave - that is if we change the data the results should change in a plausible way. For example reducing smoking rates will create more ex-smokers who carry residual risk of a range of disease so we would expect prevalence’s of these disease attributable to smoking to continue to rise for a period until, if we found that in 20 year’s time the model predicted increases in the population admission rate, this would suggest some issue with the model. Similarly if we found that the outcome was very sensitive to one or more inputs, this may suggest a problem with the way the model captures the relationship between the input variables and the output.

\*\* Examples

* Agent based (< 80 PubMed articles pa with recent peak)
  + Examples are mainly focused on the spread of communicable disease including flu and HIV, but there are a small number of examples of its application to physical activity, especially promoting walking, tackling obesity, smoking. Agent-based modelling approaches use microsimulation.
* Microsimulation
* Predictive modelling

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model. Type | | Uses | | Examples | | | In.house.Capacity.and.Capability |
| Agent-based | | Complex systems and dynamic modelling | | Interventions to promote physical activity or tackle obesity, Complex community interventions, Built environment impacts on health | | | No - no-in house capacity for application in non-communicable disease |
| Predictive modelling and machine learning | | Prevalence estimates, trends, predictions, explanatory analysis | | Estimating and predicting disease and risk factor prevalence, identifying important variables to explain variation in activity, Estimating impacts of changes in risk factors | | | Some - need data science platform and wider use of R/ Python |
| Forecasting | | Projections, time series, future states | | Future health states, Activity trends | |

## Limited - see predictive modelling current activities in PHE

There are a number of modelling teams or functions in PHE:

1. Health Protection Economic Modelling Team (check) led by Peter White and Andre Charlett. They create several models related to pandemic flu, vaccine control
2. The EDR team of mathematical modellers who model environmental and other hazards likely to affect crowds...
3. CRCE?
4. Economics team
5. Data science team in K&I
6. Other

## Activities in PHE.

Much of the modelling work done in PHE involves using or contributing to models created by others - there is very little in-house model development for non-communicable disease or public health risks and determinants.

|  |  |
| --- | --- |
| Modelling.activity | PHE.input |
| Global burden of disease | Data provision to IHME, commissioning, interpetation, networks, explanation, publication |
| MidRif (UKHF) | Data input, software development (by UKHF) to make model user friendly for PHE to run, limited projection and simulation |
| Economic models | Range of externally developed models for ROI and economic evaluation |
| Predictive models and forecasts | Limited activity - early experimentation with machine learning, Commissioned prevalence estimates from Imperial and Southampton |