Diabetes Briefing
Technical Documentation

August 2012

Contents

How the population prevalence models work 3

Step I: Estimate risk in the population 5

Step II: Modify the initial model and identify the final model 7

Step III: Estimate and forecast the number of people in each risk group in the population 13

Step IV: Estimate and forecast national and sub-national population prevalence 15

References 16

Appendix 1: Coding of the outcomes and explanatory variables 17

Appendix 2: Candidate models 24

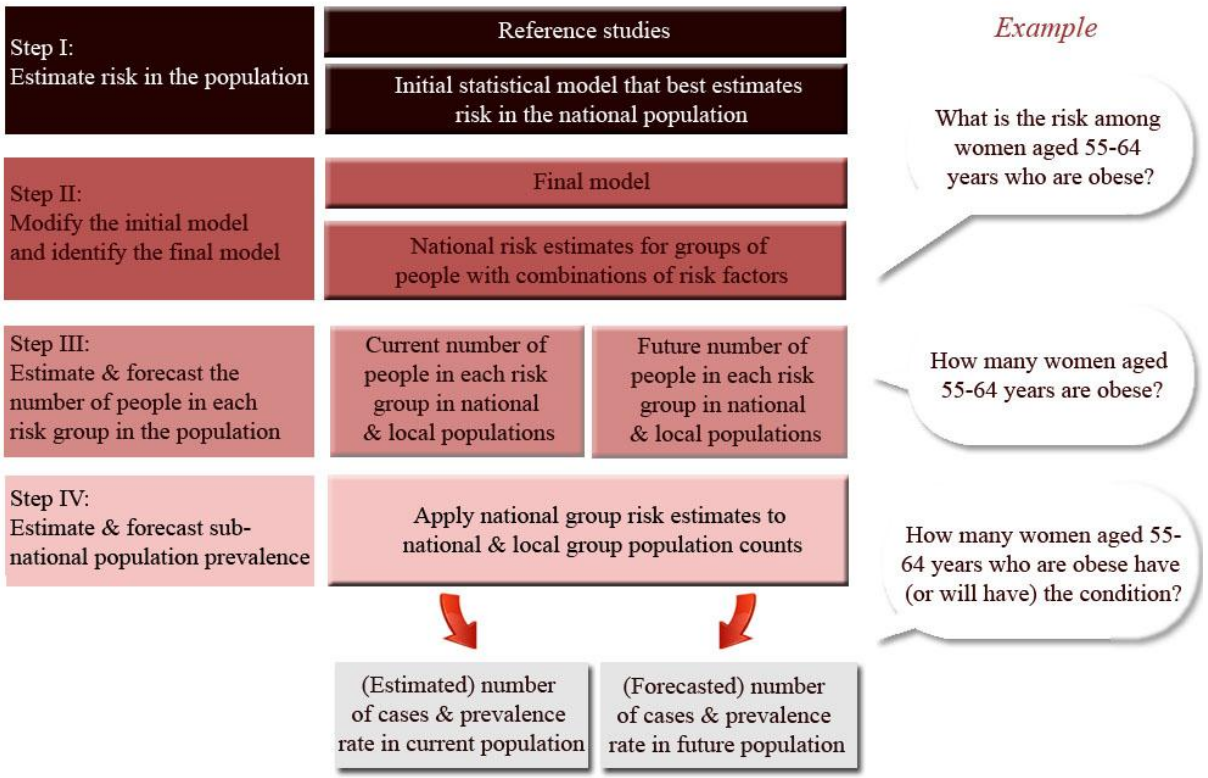
Appendix 3: Definitions of the absolute and comparative criteria 26

Appendix 4: Decision flowcharts for identifying the final models 29

How the population prevalence models work

Estimating and forecasting population prevalence involved four steps that are summarised in this section. The succeeding sections provide more detailed descriptions of each step. The steps were implemented separately in the Republic of Ireland and Northern Ireland using data specific to the particular country. Sub-national areas of the Republic of Ireland were the 32 Local Health Offices (LHO) of the Health Service Executive. Sub-national areas in Northern Ireland were the 26 Local Government Districts (LGDs)

Figure 1: How the population prevalence models work.



Step I:
Estimating risk in the population

A reference study was used to build the best predictive model of risk in the national population. The best predictive model included a number of explanatory variables for the condition. The model:

- Divides the population into risk groups defined by the categories of the explanatory variables
- Provides an estimate of the risk (at national level) of having the condition in each of the risk groups.

Step II:
Modify the initial model and identify the final model

The initial model is the best predictive model of risk based on the reference study. However, it may not be the best model based on other criteria. Specifically, the initial model may be biased due to a small number of observations that are cases or small numbers of observations that comprise the risk groups; it may produce prevalence estimates that are not satisfactorily precise; or it may not be possible to satisfactorily estimate the number of people in the population in all the groups defined by the initial model. In Step II, a 'final model' – a possibly simplified version of the initial model – was identified that:

- Is as close as possible to the initial model
- Provides sub-national estimates of population prevalence that are as unbiased and precise as possible
- Allows the population size of the risk groups in each LHO/LGD to be estimated as satisfactorily as possible

Step III:
Estimate and forecast the number of people in each risk group in the population

Population-based data (for age and sex) and data from the reference studies (for the other explanatory variables in the final model) were used to estimate and forecast the number of people in each risk group in the population, by:

- Disaggregating the reference study's national sample by the risk groups defined by the categories of the explanatory variables in the final model
- Applying the sample's national proportions to LHO/LGD population counts. The proportions were specific to explanatory variables that were included in the final model, and available for populations at sub-national level (ie age and sex).

Step IV:
Estimate and forecast national and sub-national population prevalence

The final model's national group risk estimates (Step II) were multiplied by the corresponding group population count estimates and forecasts (Step III) to estimate and forecast the number of people with the condition.

Step I:
Estimate risk in the population

Reference studies

National health surveys were used as reference studies to identify the best predictive model of risk for diabetes in each country.

In the Republic of Ireland, the Survey of Lifestyle, Attitudes and Nutrition (SLÁN) 2007 was used to identify the models. SLÁN consists of face-to-face interviews with 10,364 adults aged 18 years or more in 10,364 private residences in the Republic of Ireland (individual response rate=62%) and physical measurements of a sub-sample of 1,207 adults aged 45+ years.

The data were weighted to be representative of the age, sex, economic status, education, occupational category, ethnicity, household size, and geographical region distribution of the Republic of Ireland population (Department of Health and Children, 2008).

In Northern Ireland, the Health and Social Wellbeing Survey (HSWB) 2005/06 was used to identify the models. The HSWB survey 2005/06 consists of face-to-face interviews with 4,245 adults aged 16 years or more in 2,905 private residences in Northern Ireland (household response rate=66%). The data were weighted to be representative of the age and sex distribution of the Northern Ireland population (Northern Ireland Statistics and Research Agency (NISRA), 2005).

The models in Northern Ireland were adjusted for correlation of responses from people within the same household using Generalized Estimating Equations (GEE) with an exchangeable correlation matrix. This adjustment assumes that responses from people within a household are equally correlated but that there is no correlation between responses from people from separate households.

Outcomes and explanatory variables

An initial set of outcomes and explanatory variables relating to CHD was identified from the reference studies. The outcomes were presence or absence of clinically diagnosed diabetes.[1] The explanatory variables comprised appropriate and available biological, behavioural and social determinants of health. Definitions of the outcomes, explanatory variables and their categories for the diabetes models can be found in Appendix 1.

[1] The diabetes outcomes that are available for adults aged 18+ years in SLÁN 2007 and HSWB 2005/06 are based on self-reported doctor-diagnosed diabetes. By definition, these outcomes exclude undiagnosed diabetes and are clinical diagnosis rates rather than population prevalence rates. Physically measured diabetes (blood HbA1c concentration >=6.5%) was available in SLÁN 2007 for a sub-sample of adults aged 45+ years and this allowed us to report national population prevalence estimates and forecasts for adults aged 45+ years in the Republic of Ireland. There was no physical examination sub-sample available in NIHSWB 2005/06.

Identifying the best predictive model

A forward selection logistic regression procedure was applied to the reference studies to identify the best predictive model of risk for clinically diagnosed diabetes (the 'initial' model) at national level. The forward selection procedure builds a statistical model by identifying explanatory variables that are associated with clinically diagnosed diabetes from the initial set of explanatory variables (Appendix 1). The procedure begins with a null model and selects the explanatory variable with the largest significant association with the outcome. Further explanatory variables are selected in order of the size of their significant association with the outcome (adjusted for the explanatory variables already selected by the procedure). This selection order means that explanatory variables with more explanatory power are selected before explanatory variables with less explanatory power. The procedure stops when no further explanatory variables are significantly associated with the outcome. The forward selection logistic regression procedures were implemented in SAS Version 9.2 with a significance level of 0.05.

Table 1 shows the reference study, the outcome that was modelled and the explanatory variables that were selected for the initial model for diabetes in the Republic of Ireland and Northern Ireland. The initial model:

- Divides the population into risk groups defined by the categories of the explanatory variables
- Provides an estimate of the risk (at national level) of having the condition in each of the risk groups.

Table 1: The reference studies, the outcomes that were modelled and the explanatory variables that were selected for the initial diabetes models in the Republic of Ireland and Northern Ireland

Country	Reference study	Chronic condition	Definition of outcome in the reference study	Explanatory variables selected for the initial model
Republic of Ireland	Survey of Life, Attitudes and Nutrition (SLÁN 2007)	Diabetes	Self-reported, doctor-diagnosed diabetes in the previous 12 months (Yes / No)	Age; Employment; Body Mass Index (BMI); Smoking
Northern Ireland	Health and Social Wellbeing Survey (HSWB) 2005/06	Diabetes	Self-reported, doctor-diagnosed diabetes, ever (Yes / No)	Age; Body Mass Index (BMI); Physical activity

In the Republic of Ireland, SLÁN's physical measurement sub-sample of adults aged 45+ years asked respondents if they were currently taking medication for diabetes. Respondents who did not report a doctor-diagnosis of diabetes but reported that they were currently taking diabetes medication were considered to be 'false negatives'. The percentage of adults aged 45+ years who self-reported a doctor diagnosis of diabetes was adjusted upward to account for false negatives. No adjustment was made for adults aged less than 45 years because no data were available on their use of diabetes medication.

There was no physical examination sub-sample available in HSWB 2005/06 and no adjustment was made for diabetes medication use.

Step II:
Modify the initial model and identify the final model

Initial model and final model

The initial model is the best predictive model of risk based on the reference study. However, it may not be the best model based on other criteria. Specifically, the initial model may be biased due to a small number of observations that are cases or small numbers of observations that comprise the risk groups, it may produce prevalence estimates that are not satisfactorily precise, or it may not be possible to satisfactorily estimate the number of people in the population in all the groups defined by the initial model.

In Step II, a 'final model' – a possibly simplified version of the initial model – was identified that:

- Is as close as possible to the initial model
- Provides sub-national estimates of population prevalence that are as unbiased and precise as possible
- Allows the population size of the risk groups in each LHO/LGD to be estimated as satisfactorily as possible

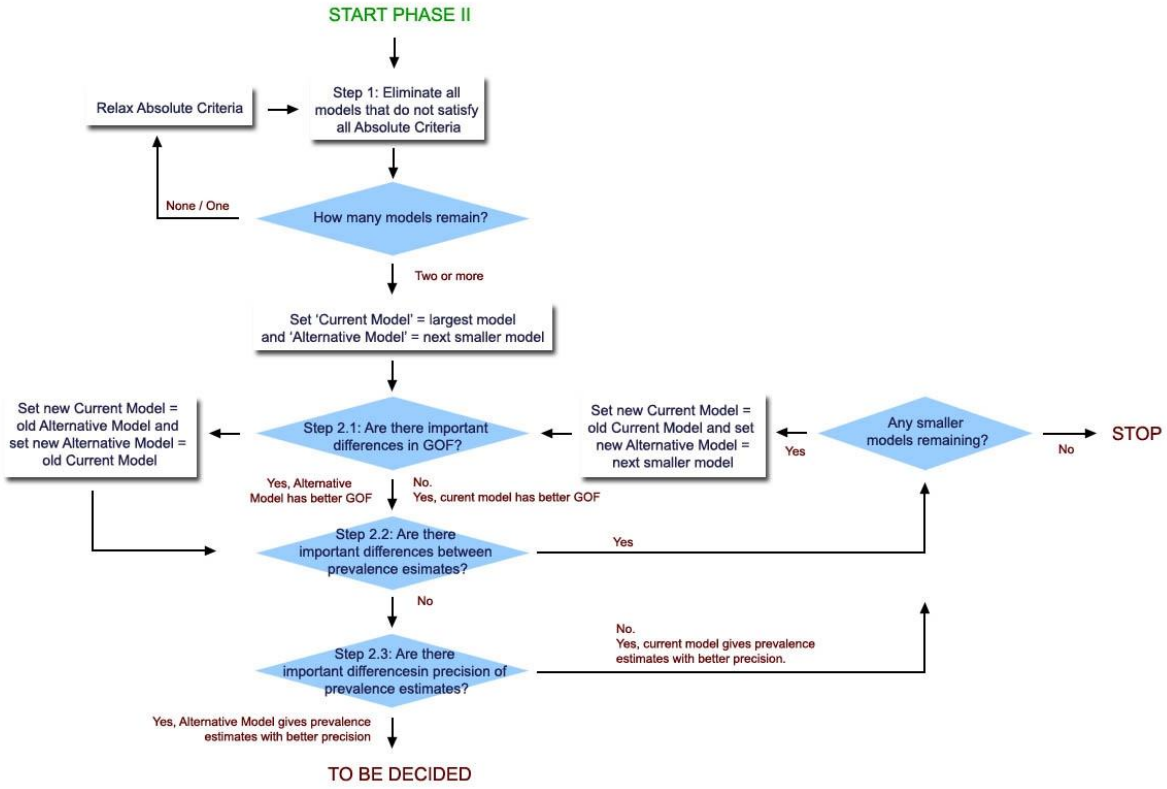
A series of 'candidate' models was generated for evaluation to identify the final model. The series of candidate models began with the initial model and nested models were generated by backward elimination of explanatory variables (ie successively removing the explanatory variable with the least explanatory power) until a null model remained. Appendix 2 shows the series of candidate models for diabetes in the Republic of Ireland and Northern Ireland.

Identifying the final model

Evaluation of the candidate models to identify the final model was a two-part process. In Part 1, candidate models were eliminated based on absolute criteria. In Part 2, the remaining candidate models were compared on comparative criteria and the final model was identified. Appendix 3 defines the metrics and thresholds for the absolute and comparative criteria.

Figure 2 shows the decision flowchart used to evaluate candidate models and identify the final model. Appendix 4 documents the decision flowchart to identify the final model in Northern Ireland and the Republic of Ireland.

Figure 2: Flowchart for evaluating candidate models and identifying the final model.



Part 1:
Eliminate candidate models that do not satisfy all of the absolute criteria

A candidate model was eliminated if it did not satisfy all four of the absolute criterion below.

Criterion A.1: Number of outcomes per explanatory variable in the model (Peduzzi et al, 1996)

Criterion A.2: Percentage of risk groups with a small number of observations (Bishop et al, 1975)

Criterion A.3: Relative standard error of population prevalence estimates (Centers for Disease Control and Prevention, 2010)

Criterion A.4: Utility - Inclusion of modifiable explanatory variables in the model

Appendix 3 defines the metrics and thresholds for the absolute criteria. Table 2 shows the possible results, decisions and rationales when the absolute criteria are applied. Note that

Part 2's assessment of bias and precision required at least two candidate models to go forward from Part 1.

Table 2: Possible results, decisions and rationales for the absolute criteria

Possible results of applying the absolute criteria	Decision for Part 1 based on accumulated results	Rationale
No candidate models or one candidate model remains	Abolish absolute criterion A.4 and check how many candidate models satisfy the other three absolute criteria. If no or one candidate model remains then abolish absolute criterion A.3 and check how many candidate models satisfy the other two absolute criteria.	Applying all four absolute criteria means that fewer candidate models remain than the minimum of two required implementing Part 2. Abolishing absolute criteria A.4 and A.3 in sequence means that at least two candidate models remain to implement Part 2.[2]
At least two candidate models remain	Set Current Model to be the largest remaining candidate model, and set Alternative Model to be the next smaller remaining candidate model. Go onto Part 2.	It is now possible to go on and use the comparative criteria to decide between the remaining candidate models.

Part 2:
Compare remaining candidate models on comparative criteria

The remaining candidate models were compared on comparative criteria to identify the final model. The comparative criteria relate to the models' goodness of fit, the similarity of the sub-national population prevalence estimates they produced, and the precision of the sub-national population prevalence estimates they produced. Appendix 3 defines the metrics and thresholds for the comparative criteria.

The models were compared two at a time and the comparative criteria identified the more appropriate of the two models. At the start of Part 2 the Current Model was the largest remaining candidate model from Part 1 and the Alternative Model was the next smaller remaining candidate model. Part 2 was repeated until all remaining models were considered and the final model was identified. Note that the comparative process favoured the smaller Alternative Model only when it was necessary to meet the criteria.

Compared to the "alternative model", the "current model" model:

- Has the better goodness of fit (in the terms of the LRT statistic used in the forward selection procedure used in Step I)
- Is the less parsimonious but is closer to the best risk prediction model

Part 2.1
Criterion C.1: Are there 'important' differences in the goodness of fit of the Current Model and the Alternative Model?

The larger Current Model should have better goodness fit than the Alternative Model because a larger model always has higher likelihood score that a smaller model nested

[2] Note that this was the situation for the models considered here. It would not necessarily be the situation for other models and other conditions.

within it. Part 2.1 compared the models' goodness of fit to ensure that the Current Model had better goodness of fit as defined in Appendix 3.

Table 3: Possible results, decisions and rationales for the comparative criterion C.1: Are there 'important' differences in the goodness of fit of the Current Model and the Alternative Model?

Possible results of applying the comparative criterion	Decision for Part 2.1 based on accumulated results	Rationale
No important differences	Go to Part 2.2.	No change of the Current Model is necessary.
Yes, Current Model has better goodness of fit	Go to Part 2.2.	Confirms the better goodness of fit suggested by the LRT statistic and no change of the Current Model is necessary.
Yes, Alternative Model has better goodness of fit [3]	Swap Current Model and Alternative Model before continuing: set the new Current Model to be the existing Alternative Model, set new Alternative Model to be the Current Model. Go to Part 2.2.	We now know that the LRT statistic from the forward selection logistic regression procedure hides some important lack of fit of the Current Model, and the Alternative Model now has better goodness of fit and so should be carried forward as the new Current Model.

Part 2.2 Criterion C.2: Are there 'important' differences between the sets of sub-national population prevalence estimates produced by the Current Model and the Alternative Model?

Part 2.2 compared the Current Model's and the Alternative Model's sets of sub-national prevalence estimates. 'Important' differences for Part 2.2 are defined in Appendix 3.

Table 4: Possible results, decisions and rationales for the comparative criterion C.2: Are there 'important' differences between the sets of sub-national population prevalence estimates produced by the Current Model and the Alternative Model?

Possible results of applying the comparative criterion	Decision for Part 2.2 based on accumulated results	Rationale
Important differences	Retain the Current Model and consider the next smaller candidate model: set the new Current Model to be the existing Current Model, set new Alternative Model to be the next	The two models produce population prevalence estimates that exhibit important differences. In this case we favour the model with better goodness of fit: no change of the

[3] This situation did not arise in any of model comparisons for the conditions considered here.

	smaller candidate model and go back to Part 2.1 with these new settings. If no smaller candidate model exists, then STOP.	Current Model is necessary.
No important differences	Go to Part 2.3.	No change of the Current Model is necessary.

Part 2.3 Criterion C.3: Are there 'important' differences in the precision of the sub-national population prevalence estimates produced by the Current Model and the Alternative Model?

Part 2.3 compared the precision of the Current Model's and the Alternative Model's sets of sub-national prevalence estimates. 'Important' differences for Part 2.3 are defined in Appendix 3.

Table 5: Possible results, decisions and rationales for the comparative criterion C.3: Are there 'important' differences between the sets of sub-national population prevalence estimates produced by the Current Model and the Alternative Model?

Possible results of applying the comparative criterion	Decision for Part 2.3 based on accumulated results	Rationale
No important differences	Retain the Current Model and consider the next smaller candidate model: set the new Current Model to be the existing Current Model, set new Alternative Model to be the next smaller candidate model and go back to Part 2.1 with these new settings. If no smaller candidate model exists, then STOP.	The two models produce similar population prevalence estimates (Part 2.2) with no important differences in their precision (Part 2.3). In this case we favour the model with better goodness of fit: no change of the Current Model is necessary.
Yes, Current Model has better precision	Retain the Current Model and consider the next smaller candidate model: set the new Current Model to be the existing Current Model, set new Alternative Model to be the next smaller candidate model and go back to Part 2.1 with these new settings. If no smaller candidate model exists, then STOP.	The two models produce similar population prevalence estimates (Part 2.2). The Current model has better goodness of fit and provides more precise population prevalence estimates (Part 2.3). In this case, no change of the Current Model is necessary.
Yes, Alternative Model has better	This situation did not arise in any of the model comparisons considered here.	The two models produce similar population prevalence estimates

precision		(Part 2.2). However, the Current Model has better goodness of fit while the Alternative Model provides more precise population prevalence estimates (Part 2.3). We need to decide if we want to change the Current Model

Once we've STOPPED, the final model is the last Current Model. Table 6 shows the outcome that was modelled and the explanatory variables in the initial and final model for diabetes in the Republic of Ireland and Northern Ireland.

Table 6: The reference studies, the outcomes that were modelled and the explanatory variables in the initial and final models in the Republic of Ireland and Northern Ireland

Country	Reference study	Chronic condition	Definition of outcome in the reference study	Explanatory variables selected for the initial model	Explanatory variables included in the final model
Republic of Ireland	Survey of Life, Attitudes and Nutrition (SLÁN 2007)	Diabetes	Self-reported, doctor-diagnosed diabetes in the previous 12 months (Yes / No)	Age; Employment; Body Mass Index (BMI); Smoking	Age; Employment
Northern Ireland	Health and Social Wellbeing Survey (HSWB) 2005/06	Diabetes	Self-reported, doctor-diagnosed diabetes, ever (Yes / No)	Age; Body Mass Index (BMI); Physical activity	Age; Body Mass Index (BMI); Physical activity

The final model:

- Divided the population into risk groups defined by combinations of the categories of the explanatory variables in the final model
- Provided an estimate of the risk (at national level) of having the condition in each of the risk groups

It was then necessary to estimate and forecast the number of people in each of these risk groups (Step III) so that group risk estimates could be multiplied by group population count estimates and forecasts to give the estimated/forecasted number of cases (Step IV).

Step III: Estimate and forecast the number of people in each risk group in the population.

We combined population-based data (for age and sex) and data from the reference studies (for the other explanatory variables) to estimate and forecast the number of people in each risk group in the population.

Population-based data: Republic of Ireland

Population data were provided by the Central Statistics Office (CSO). Population estimates for 2010 and population projections for 2015 and 2020 were based on the usually resident population at Census 2006. CSO (2008) prepared different population projection scenarios based on different assumptions about trends in mortality, fertility, international migration, and internal migration. Four scenarios were prepared at sub-national level:

1. M0F1 Traditional
2. M0F1 Recent
3. M2F1 Traditional
4. M2F1 Recent

where M0: Net international migration=0
M1: Moderately positive but declining net international migration
F1: Fertility rate remains constant at 2006 level (1.9)
Traditional: Internal migration follows the patterns traditionally observed
Recent: Internal migration follows the patterns recently observed

See CSO (2008) for details.

IPH's original population prevalence forecasts (Balanda et al, 2010) were based on the M2F1 Traditional scenario. However, population estimates published since then (CSO, 2011) suggest that net international migration is negative so M0 (international migration=0) may now be the most appropriate international migration assumption available. The CSO did not identify a preferred population projection scenario so we produced population prevalence forecasts based on first three scenarios above: M0F1 Traditional; M0F1 Recent; M2F1 Traditional.

Sub-national population estimates and projections were not available for LHOs but were available for eight Regional Authorities. Age-sex specific changes in population from Census 2006 to 2010 (estimates), 2015 and 2020 (both projections) were calculated for each Regional Authority. These Regional Authority adjustment factors were applied to Census 2006 LHO data. For this we assumed that age-sex specific changes at Regional Authority level apply to each of the LHOs within that Regional Authority.

Population-based data: Northern Ireland

Population data were provided by NISRA. Population estimates for 2010 and population projections for 2015 and 2020 were based on the usually resident population. Population estimates for 2010 were based on Census 2001. Population projections for 2015 and 2020 were based on 2008 population estimates as these were the most up-to-date population projections available for LGDs. Populations projections for LGDs by age and sex are only

produced for a principal projection scenario which incorporates what are considered to be the best assumptions, based on historical trends, about mortality, fertility and migration. The principal scenario assumes declining mortality, small positive net migration, and a fertility rate of 1.95. See NISRA's population projections for details.

Reference study data

Population data were not available for some of the explanatory variables in the final models. In the Republic of Ireland and Northern Ireland, the reference study was used to estimate the distribution of these explanatory variables (ie explanatory variables apart from age and sex). To do this we:

1. Disaggregated the reference study's national sample by all the explanatory variables in the final model.
2. Calculated the reference study's national sample percentages specific to the explanatory variables in the final model for which we had population-based data (ie age and sex).
3. If age and sex were in the final model, we applied the age-sex-specific percentages from the reference study's national sample to the age-sex specific LHO/LGD population counts.
4. If age (but not sex) was in the final model, we applied the age-specific percentages from the reference study's national sample to the age specific LHO/LGD population counts.

This method assumes that:

- Each LHO/LGD has the same national age-specific and age-sex-specific distribution of explanatory variables. This was necessary because of limited availability of data on explanatory variables at sub-national level – sample sizes were not large enough to provide robust sub-national disaggregation of the reference study's sample by all the explanatory variables in the final model (see 1 above).
- The age-specific and age-sex-specific distribution of explanatory variables will not change in future years (ie the age-sex-specific prevalence of explanatory variables will remain constant at current levels).

Step IV:
Estimate and forecast national and sub-national population prevalence

Population prevalence estimates and forecasts

The final model's national group risk estimates (Step II) were multiplied by the corresponding group population count estimates and forecasts (Step III) to estimate and forecast the number of people with the condition.

Confidence intervals

The statistical models of risk are based on reference studies that use samples from the population rather than the whole population. Therefore the population prevalence estimates and forecasts provide an imprecise estimate of the true population value. To quantify the imprecision of the estimates and forecasts we calculated 95% confidence intervals.

Because the estimates and forecasts use the same reference studies to develop the risk model and to estimate the number of people in the corresponding risk groups, assumptions had to be made to allow the standard errors of the population prevalence estimates and forecasts to be calculated. In particular, we assumed that:

- The group risk estimates and the group population count estimates and forecasts in each LGD/LHO are statistically independent
- The number of people in each risk group in each LHO/LGD population is known without error.

References

Balanda, K. P., Barron, S., Fahy, L., McLaughlin, A. (2010). *Making Chronic Condition Count: Hypertension, Coronary Heart Disease, Stroke, Diabetes. A systematic approach to estimating and forecasting population prevalence on the island of Ireland.* Dublin: Institute of Public Health in Ireland, 2010.
http://www.publichealth.ie/sites/default/files/documents/files/Making%20Chronic%20Conditions.pdf

Barron, S., Balanda, K. P. (2010). *Making Chronic Condition Count: Chronic Airflow Obstruction. A systematic approach to estimating and forecasting population prevalence on the island of Ireland.* Dublin: Institute of Public Health in Ireland, 2010.
http://www.thehealthwell.info/sites/all/libraries/tinymce/files/5_MCCC_Chronic_Airflow_Obstruction_Dec_2010.pdf

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Centers for Disease Control and Prevention (2010). *Reliability of estimates*. Atlanta: Centers for Disease Control and Prevention.
http://www.cdc.gov/nchs/ahcd/ahcd_estimation_reliability.htm

Central Statistics Office (2008). *Population and Labour Force Projections 2011-2041*. Dublin: Central Statistics Office.
http://www.cso.ie/en/releasesandpublications/population/populationandlabourforceprojections2011-2041/

Central Statistics Office (2011). *Population and Migration Estimates April 2011*. Dublin: Central Statistics Office.
http://www.cso.ie/releasespublications/documents/population/2011/popmig_2011.pdf

Department of Health and Children (2008). *Survey of Lifestyle, Attitudes and Nutrition 2007.* Dublin: Department of Health and Children.
http://www.slan06.ie/index.htm

Haase, T. and Pratschke, J. (2008). *New Measures of Deprivation for the Republic of Ireland*. Dublin: Pobal.
http://www.pobal.ie/WhatWeDo/Deprivation/Pages/InformationforBeneficiaries.aspx

Northern Ireland Statistics & Research Agency (2005). *Northern Ireland Multiple Deprivation Measure 2005*.
http://www.nisra.gov.uk/deprivation/nimdm_2005.htm

Northern Ireland Statistics and Research Agency Central Survey Unit (2005). *Northern Ireland Health and Social Wellbeing Survey 2005/06.* Belfast: Northern Ireland Statistics and Research Agency.
http://www.csu.nisra.gov.uk/survey.asp46.htm

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R. (1996). *A simulation study of the number of events per variable in logistic regression analysis.* Journal of Clinical Epidemiology, Vol 49, No 12, pp1373-9.
http://www.ncbi.nlm.nih.gov/pubmed/8970487

Appendix 1:
Coding of the outcomes and explanatory variables

TABLE A1: DIABETES **CODING OF THE DIABETES MODELS' OUTCOMES AND EXPLANATORY VARIABLES INCLUDED IN THE VARIABLE SELECTION PROCEDURE**				
Outcomes and explanatory variables	**Recoding used in models for estimating risk and estimating the number of people in each risk group in the population**		**Original coding in reference study**	
	Republic of Ireland	**Northern Ireland**	**Republic of Ireland**	**Northern Ireland**
Outcomes				
Diabetes	Self-reported doctor-diagnosed diabetes in the previous 12 months (Yes / No)	Self-reported doctor-diagnosed diabetes, ever (Yes / No) (Excludes women who were pregnant at the time of diagnosis)	Self-reported doctor-diagnosed diabetes in the previous 12 months (Yes / No)	Self-reported doctor-diagnosed diabetes, ever (Yes / No) (Excludes women who were pregnant at the time of diagnosis)
Explanatory variables				
Sex	Male Female	Male Female	Male Female	Male Female
Age	18-34 years 35-44 years 45-54 years 55-64 years 65-74 years 75+ years	18-34 years 35-44 years 45-54 years 55-64 years 65-74 years 75+ years	Single year 18+	Single year 16+
Ethnicity	White	White	Irish	White

TABLE A1: DIABETES **CODING OF THE DIABETES MODELS' OUTCOMES AND EXPLANATORY VARIABLES INCLUDED IN THE VARIABLE SELECTION PROCEDURE**				
Outcomes and explanatory variables	**Recoding used in models for estimating risk and estimating the number of people in each risk group in the population**		**Original coding in reference study**	
	Republic of Ireland	**Northern Ireland**	**Republic of Ireland**	**Northern Ireland**
	Non-white	Non-white	Irish Traveller Any other white background African Any other black background Chinese Any other Asian background Other	Chinese Irish traveller Indian Black - Caribbean Mixed ethnic Other
Body Mass Index (BMI)	Underweight / Normal <25 Over weight 25-29.99 Obese >=30	Underweight/Normal <25 Over weight 25-29.99 Obese >=30	Physically measured and self-reported BMI value	Physically measured BMI value
Physical activity	Low Moderate High	Sedentary Intermediate Above recommended levels	Low Moderate High	Sedentary Intermediate Above recommended levels
Cigarette smoking	Former smoker Never smoked Current smoker	Former smoker Never smoked Current smoker	Former smoker Never smoked Current smoker	Former smoker Never smoked Current smoker

TABLE A1: DIABETES **CODING OF THE DIABETES MODELS' OUTCOMES AND EXPLANATORY VARIABLES INCLUDED IN THE VARIABLE SELECTION PROCEDURE**				
Outcomes and explanatory variables	**Recoding used in models for estimating risk and estimating the number of people in each risk group in the population**		**Original coding in reference study**	
	Republic of Ireland	**Northern Ireland**	**Republic of Ireland**	**Northern Ireland**
Alcohol consumption	Never / Monthly or less / 2-4 times a month 2-3 times a week >=4 times a week	Never / Not at all in last 12 months / 1 or 2 times a year / Once every couple of months / 1 or 2 times a month 1 or 2 times a week / 3 or 4 times a week 5 or 6 times a week / Almost every day	Never Monthly or less 2-4 times a month 2-3 times a week 4 or more times a week	Never Not at all in last 12 months 1 or 2 times a year Once every couple of months 1 or 2 times a month 1 or 2 times a week 3 or 4 times a week 5 or 6 times a week Almost every day
Fruit and vegetable consumption	< 5 a day >= 5 a day	< 5 a day >= 5 a day	Number of portions derived from Food Frequency Questionnaire	Number of portions
Highest level of education	Primary level Secondary level Third level	Primary (No Qualifications/Other Qualifications) Secondary (GCSE D-G, GCSE, A-C, A GCE) Third level (Higher education/Degree)	Some primary Primary Intermediate/Junior/Group Leaving Certificate Diploma/Certificate Primary degree	No qualifications Other qualifications GCSE D-G, GCSE, A-C, A GCE Higher education

TABLE A1: DIABETES **CODING OF THE DIABETES MODELS' OUTCOMES AND EXPLANATORY VARIABLES INCLUDED IN THE VARIABLE SELECTION PROCEDURE**				
Outcomes and explanatory variables	**Recoding used in models for estimating risk and estimating the number of people in each risk group in the population**		**Original coding in reference study**	
	Republic of Ireland	**Northern Ireland**	**Republic of Ireland**	**Northern Ireland**
		Unknown (People aged 70 years or more)	Postgraduate/Higher degree	Degree
Employment status	Employed (Employee; Self-employed outside farming; Farmer) Unemployed (Unemployed, actively looking for a job) Economically inactive (Student; State training scheme; Long-term sickness or disability; Home duties/looking after home or family; Retired; Other)	Employed (Worked last week / Away from work last week) Unemployed (Waiting / Looking / Not looking) Economically inactive	Employee Self-employed outside farming Farmer Student (full time) State training scheme Unemployed, actively looking for job Long-term sickness or disability Home duties/ looking after home or family Retired Other (please specify)	Worked last week Away from work last week Waiting to take up job Looking for work Not looking sick Economically inactive
Social class	SC 1-2 (Professional and managerial) SC 3-4 (Non-manual and skilled manual) SC 5-6 (Semi-skilled and unskilled)	Professional / Managerial Skilled non-manual and skilled manual Partly skilled / Unskilled Unclassified	SC 1-2 (Professional / Managerial) SC 3-4 (Skilled non-manual and Skilled manual) SC 5-6 (Semi-skilled and unskilled)	Professional / Managerial Skilled non manual Skilled manual Semi-skilled Unskilled

TABLE A1: DIABETES **CODING OF THE DIABETES MODELS' OUTCOMES AND EXPLANATORY VARIABLES INCLUDED IN THE VARIABLE SELECTION PROCEDURE**				
Outcomes and explanatory variables	**Recoding used in models for estimating risk and estimating the number of people in each risk group in the population**		**Original coding in reference study**	
	Republic of Ireland	**Northern Ireland**	**Republic of Ireland**	**Northern Ireland**
	Unclassified		Unclassified	Economically inactive
Area deprivation	1 to 5 (Least deprived to Most deprived)	1 to 5 (Least deprived to Most deprived)	Observations assigned to one of 32 LHOs	Observations assigned to one of 890 SOAs

Notes

Republic of Ireland's SLÁN 2007 Body Mass Index

Physically measured BMI was available for 2,170 respondents and self-reported BMI was available for all respondents. We used physically measured BMI where available and adjusted self-reported BMI for the other respondents. Self-reported BMI was adjusted by age- sex-specific factors that were calculated by comparing measured BMI with self-reported BMI for the 2,170 respondents who had both.

Northern Ireland's HSWB 2005/06 Highest level of education

Highest level of education of people aged 70 years or more was coded as 'unknown' because this question was not asked of people aged 70 years or more.

Assigning observations to area-based deprivation categories

In the Republic of Ireland, deprivation scores for Electoral Divisions (EDs) were taken from New Measures of Deprivation for the Republic of Ireland (Haase and Pratschke, 2008). Five deprivation categories were created by ordering the deprivation scores for all EDs and identifying cut-off scores that created five categories with approximately equal numbers of EDs. SLÁN 2007 observations were assigned to one of 32 LHOs as ED-level data were not available. An LHO's deprivation score was calculated as the population weighted average of the deprivation scores of the EDs within that LHO. The LHO deprivation score was then assigned to an ED based deprivation category. Therefore, there was not an equal number of LHOs within each category.

In Northern Ireland, deprivation scores for Super Output Areas (SOAs) were taken from the Northern Ireland Multiple Deprivation Measure 2005 (NISRA, 2005). Five deprivation categories were created by ordering the deprivation scores for all SOAs and identifying cut-off scores that created five categories with approximately equal numbers of SOAs**.** HSWB 2005/06 observations were assigned to an SOA and to an SOA based deprivation category.

Appendix 2:

Candidate models

APPENDIX 2: CANDIDATE MODELS FOR DIABETES			
Country	**Condition**	**Candidate models and final model**	**Explanatory variables**
Republic of Ireland	**Diabetes**	Diabetes_1	Age
		Diabetes_2 (Final model)	Age; Employment
		Diabetes_3	Age; Employment; BMI
		Diabetes_4 (Initial model)	Age; Employment; BMI; Smoking
Northern Ireland	**Diabetes**	Diabetes_1	Age
		Diabetes_2	Age; BMI
		Diabetes_3 (Initial and final model)	Age; BMI; Physical activity

Appendix 3:

Definitions of the absolute and comparative criteria

APPENDIX 3: DEFINITIONS OF THE MODELS' ABSOLUTE AND COMPARATIVE CRITERIA			
Criterion	**Metric**	**Cut-off**	**Result**
Absolute criteria			
A.1 Number of outcomes per explanatory variable in the model (Peduzzi et al, 1996)	Number of outcomes in the sample ---------- Number of explanatory variables in the model	>= 10	Don't eliminate the model
		< 10	Eliminate the model
A.2 Percentage of risk groups with a small number of observations (Bishop et al, 1975)	100 * Number of risk groups with < 5 observations ---------- Number of risk groups	<= 5%	Don't eliminate the model
		> 5%	Eliminate the model
A.3 Relative Standard Error (RSE) of population prevalence estimates (Centers for Disease Control and Prevention, 2010)	100 * Max RSE (over all LHOs/LGDs)	<= 30%	Don't eliminate the model
		> 30%	Eliminate the model
A.4 Utility: Inclusion of modifiable explanatory variables in the model	Number of explanatory variables other than age or sex in the model	>= 1	Don't eliminate the model
		0	Eliminate the model
Comparative criteria			
C.1 Goodness of fit criterion: Are there 'important' differences between the goodness of fit of the current and the Alternative Model?	Indications of goodness of fit for a model: a) Likelihood Ratio Test significant at 5% level b) Area under Receiver Operating Characteristic curve (c index) is significantly larger (at 5% level) l c) A more acceptable residual plots (based on visual assessments) To each model: • Assign 1 point if favoured by a) • Assign 1 point if favoured by b) • Assign ½ point if favoured by c) Assign no points if an indication is inconclusive. Sum scores for each model. Then the decision in next column is based on size of total score.	Models have equal cores	No Important differences

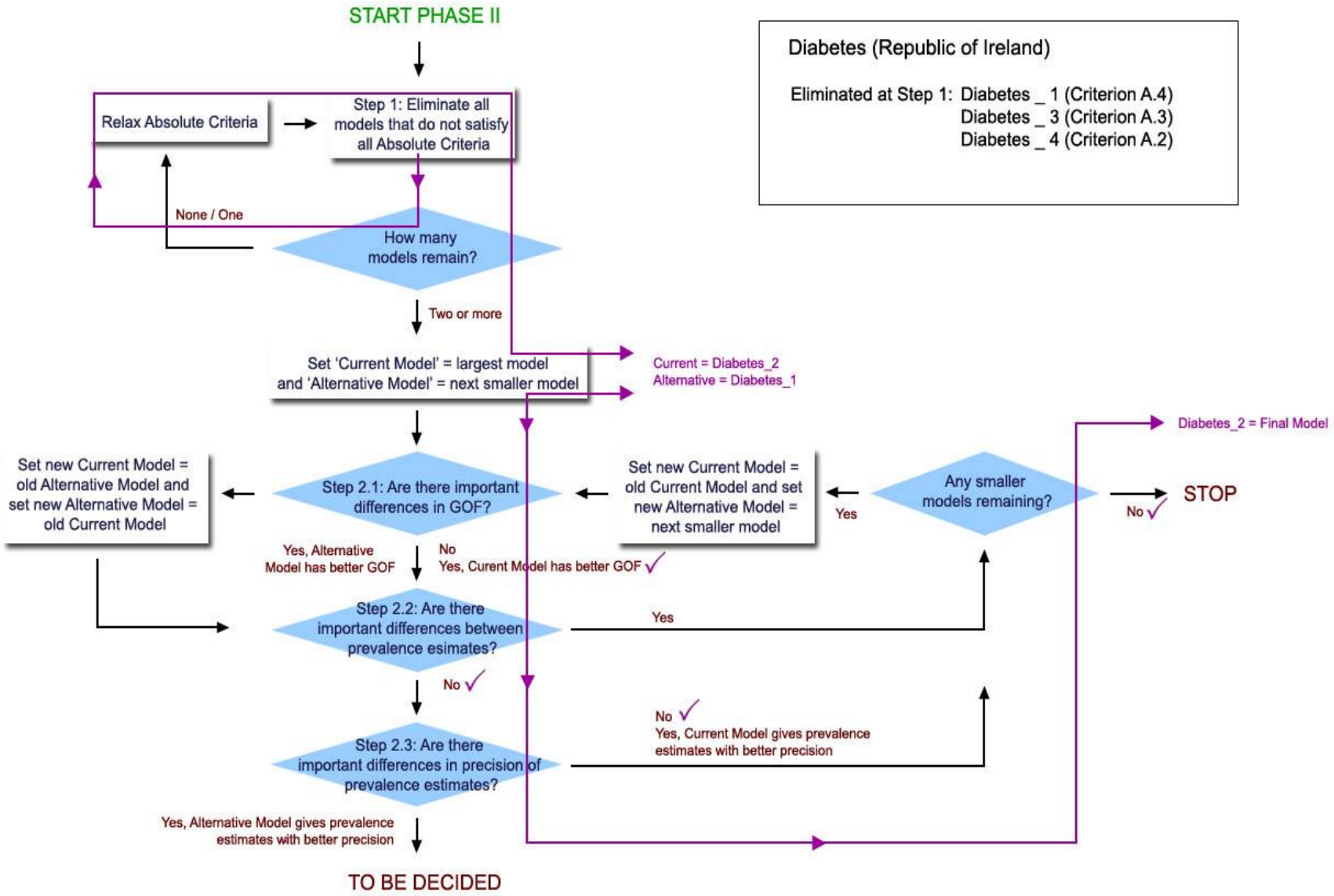
APPENDIX 3: DEFINITIONS OF THE MODELS' ABSOLUTE AND COMPARATIVE CRITERIA			
Criterion	**Metric**	**Cut-off**	**Result**
		Models have unequal scores	Important differences: model with the highest score has better goodness of fit
C.2 Similarity of prevalence estimates: Are there 'important' differences between the sets of sub-national population prevalence estimates produced by the Current Model and the Alternative Model?	Set of values (over all LHOs / LGDs) of: 100*Estimated number of cases (Alternative Model) ---------- Estimated number of cases (Current Model)	All the values are >= 99% and a <= 101%	No important differences[4]
		One or more of the values are < 99% or at least one value > 101%	Important differences
C.3: Precision of sub-national population prevalence estimates: Are there 'important' differences between the precision of the sub-national population prevalence estimates based on the Current Model and the Alternative Model	Set of values (over all LHOs/LGDs) of : RSE (Alternative Model) ---------- RSE (Current Model) Relative standard error = Square root (variance of estimated number of cases) / Estimated number of cases	More than half of the areas have values >120%	Important differences: Current Model has better precision
		More than half of areas have values < 80%	Important differences: Alternative model has better precision
		All other circumstances	No important differences[5]

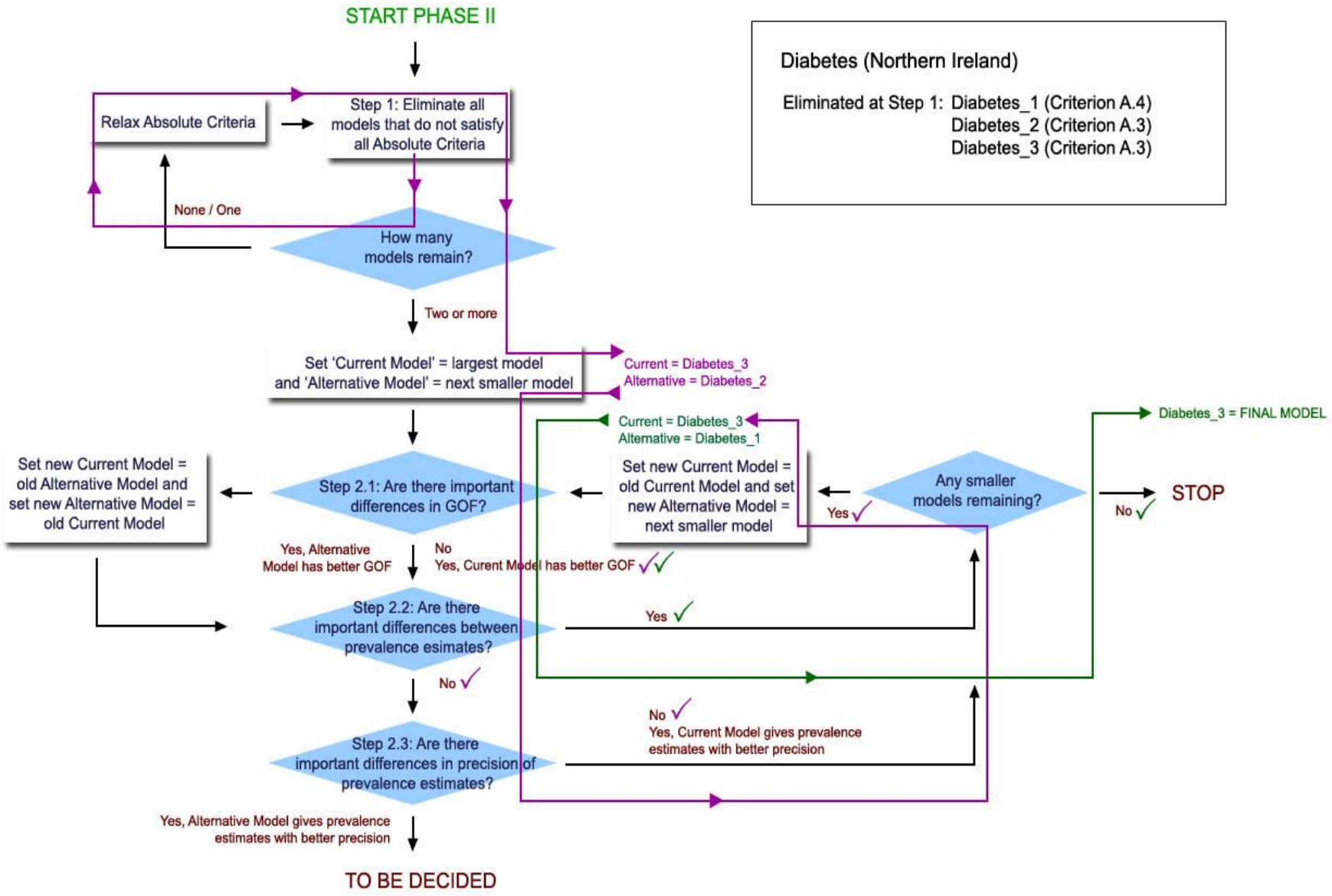
[4] So there is no important differences if no estimate from the Alternative Model differs from the corresponding estimate from the Current Model by more than 1%

[5] If more than half of all areas have an RSE (Alternative Model) and an RSE (Current Model) that differ by at most 20%, there are 'no important differences'. The converse is not necessarily true as long as neither of the critical ranges < 80%, > 120% dominates

Appendix 4:

Decision flowcharts for identifying the final models


START PHASE II
Diabetes (Republic of Ireland)
Eliminated at Step 1: Diabetes _ 1 (Criterion A.4)
Diabetes _ 3 (Criterion A.3)
Diabetes _ 4 (Criterion A.2)
Relax Absolute Criteria
Step 1: Eliminate all models that do not satisfy all Absolute Criteria
None / One
How many models remain?
Two or more
Set 'Current Model' = largest model and 'Alternative Model' = next smaller model
Current = Diabetes_2
Alternative = Diabetes_1
Diabetes_2 = Final Model
Set new Current Model = old Alternative Model and set new Alternative Model = old Current Model
Step 2.1: Are there important differences in GOF?
Set new Current Model = old Current Model and set new Alternative Model = next smaller model
Yes
Any smaller models remaining?
No
STOP
Yes, Alternative Model has better GOF
No
Yes, Curent Model has better GOF
Step 2.2: Are there important differences between prevalence esimates?
Yes
No
Step 2.3: Are there important differences in precision of prevalence estimates?
No
Yes, Current Model gives prevalence estimates with better precision
Yes, Alternative Model gives prevalence estimates with better precision
TO BE DECIDED


START PHASE II
Diabetes (Northern Ireland)
Eliminated at Step 1: Diabetes_1 (Criterion A.4)
Diabetes_2 (Criterion A.3)
Diabetes_3 (Criterion A.3)
Relax Absolute Criteria
Step 1: Eliminate all models that do not satisfy all Absolute Criteria
None / One
How many models remain?
Two or more
Set 'Current Model' = largest model and 'Alternative Model' = next smaller model
Current = Diabetes_3
Alternative = Diabetes_2
Current = Diabetes_3
Alternative = Diabetes_1
Diabetes_3 = FINAL MODEL
Set new Current Model = old Alternative Model and set new Alternative Model = old Current Model
Step 2.1: Are there important differences in GOF?
Set new Current Model = old Current Model and set new Alternative Model = next smaller model
Any smaller models remaining?
Yes
No
STOP
Yes, Alternative Model has better GOF
No
Yes, Curent Model has better GOF
Step 2.2: Are there important differences between prevalence esimates?
Yes
No
No
Yes, Current Model gives prevalence estimates with better precision
Step 2.3: Are there important differences in precision of prevalence estimates?
Yes, Alternative Model gives prevalence estimates with better precision
TO BE DECIDED