**Stroke prevalence model for small populations:**

**Technical Document produced for Public Health England**

**Ben Hollis, Roger Newson, Bowen Su, Ioanna Tzoulaki, Pankaj Sharma, Azeem Majeed & Michael Soljak**

**Department Primary Care & Public Health**

**School of Public Health**



1. Contents

[1 Executive Summary 3](#_Toc456702048)

[2 Background 4](#_Toc456702049)

[2.1 CVD Risk Factors 4](#_Toc456702050)

[2.1.1 Age 5](#_Toc456702051)

[2.1.2 Gender 5](#_Toc456702052)

[2.1.3 Ethnicity 5](#_Toc456702053)

[2.1.4 Hypertension 5](#_Toc456702054)

[2.1.5 Smoking 6](#_Toc456702055)

[2.1.6 Diabetes 6](#_Toc456702056)

[2.1.7 Obesity 6](#_Toc456702057)

[2.1.8 Dyslipidaemia 6](#_Toc456702058)

[2.1.9 Deprivation and socioeconomic status 7](#_Toc456702059)

[2.1.10 Physical Activity 7](#_Toc456702060)

[2.1.11 Inflammatory Markers 7](#_Toc456702061)

[2.1.12 Chronic Kidney Disease 7](#_Toc456702062)

[2.1 Stroke prevalence from the literature 10](#_Toc456702063)

[3 Methods 12](#_Toc456702064)

[3.1 “Data discovery” from UK survey data sources 12](#_Toc456702065)

[3.1.1 Whitehall II 12](#_Toc456702066)

[3.1.2 Airwave Health Monitoring Study 12](#_Toc456702067)

[3.1.3 Health Survey for England 2013 12](#_Toc456702068)

[3.1.4 English Longitudinal Study of Ageing 13](#_Toc456702069)

[3.2 Definition of Outcomes 15](#_Toc456702070)

[3.3 Risk factor variables 16](#_Toc456702071)

[3.3.1 Age 16](#_Toc456702072)

[3.3.2 Sex 16](#_Toc456702073)

[3.3.3 Ethnicity 16](#_Toc456702074)

[3.3.4 Hypertension 16](#_Toc456702075)

[3.3.5 Dyslipidaemia 17](#_Toc456702076)

[3.3.6 Family History (CHD) 17](#_Toc456702077)

[3.3.7 Smoking 18](#_Toc456702078)

[3.3.8 Obesity 18](#_Toc456702079)

[3.3.9 Diabetes 19](#_Toc456702080)

[3.3.10 Chronic Kidney Disease 19](#_Toc456702081)

[3.3.11 “Hypercoagulable state”/on anti-coagulant 19](#_Toc456702082)

[3.3.12 Inflammatory Markers 20](#_Toc456702083)

[3.3.13 Socioeconomic status/deprivation 20](#_Toc456702084)

[3.3.14 Physical Activity 20](#_Toc456702085)

[3.4 Statistical Analyses 20](#_Toc456702086)

[3.4.1 Whitehall II missing data 20](#_Toc456702087)

[3.4.2 Whitehall II descriptive analyses 21](#_Toc456702088)

[3.4.3 Whitehall II regression modelling 21](#_Toc456702089)

[3.4.4 Interactions 21](#_Toc456702090)

[3.4.5 Internal validation 21](#_Toc456702091)

[3.5 Local prevalence estimates 22](#_Toc456702092)

[3.5.1 Method 1: bootstrapping procedure to produce repeated samples 24](#_Toc456702093)

[3.5.2 Method 2: Logistic regression and inverse probability weights 24](#_Toc456702094)

[3.6 Validation of local estimates 26](#_Toc456702095)

[3.6.1 Internal validation 26](#_Toc456702096)

[3.6.2 External validation 26](#_Toc456702097)

[4 Results 28](#_Toc456702098)

[4.1 Stroke prevalence from Whitehall II Data 28](#_Toc456702099)

[4.1.1 Missing data 28](#_Toc456702100)

[4.1.2 Baseline characteristics of Whitehall II Respondents 28](#_Toc456702101)

[4.2 Whitehall II CVD definitions, incidence & prevalence 30](#_Toc456702102)

[4.2.1 Whitehall II prevalence 30](#_Toc456702103)

[4.3 Stroke regression modelling 30](#_Toc456702104)

[4.3.1 Univariate logistic analysis 30](#_Toc456702105)

[4.3.2 Multivariate logistic analysis for stroke 31](#_Toc456702106)

[4.4 Internal validation 34](#_Toc456702107)

[4.4.1 ROC curves 34](#_Toc456702108)

[4.4.2 Probability and sensitivity/specificity analysis 35](#_Toc456702109)

[4.5 Local estimates 35](#_Toc456702110)

[4.5.1 Internal validation 35](#_Toc456702111)

[4.5.2 External validation of practice estimates against QOF prevalence 35](#_Toc456702112)

[5 References 38](#_Toc456702113)

[6 Appendix: additional information 44](#_Toc456702114)

Stroke prevalence model Technical Document

# Executive Summary

# Background

The Department of Primary Care & Public Health (PCPH) in the School of Public Health (SPH) at Imperial College London (ICL) has tendered successfully to Public Health England (PHE) to develop small population prevalence models for several chronic diseases. PHE has requested a single cardiovascular disease (CVD) prevalence model. CVD is a term used to define disorders which affect the heart and blood vessels. These may be broadly subcategorised into two classes: 1) atherosclerotic CVDs including coronary heart disease (CHD), cerebrovascular disease more commonly called stroke, and transient ischaemic attacks (TIAs), and peripheral arterial disease (PAD); and 2) CVDs of other aetiology including congenital and rheumatic heart disease and cardiac arrhythmias. We considered that it would be feasible, and more useful at the local level, to develop three separate prevalence models for CHD, stroke and TIA, and PAD, nevertheless using the same data source and exactly the same methods. There is a large but not complete overlap in risk factors for the three diseases, but the prevalence of each is distinct. The prevalence of each disease in each local population can be summed to provide an overall prevalence of CVD. (For the present models, we focus on the atherosclerotic CVDs (CHD, CD/stroke, and PAD) as these represent the vast majority of the CVD disease burden in the UK.) This Technical Document considers all CVD risk factors, but then describes the derivation of the model for CHD.

Collectively, CVD is the leading cause of mortality worldwide, accounting for 17.5 million deaths in 2012.[[1](#_ENREF_1)] In the UK, CVD remains among the leading causes of death and presents a tremendous financial burden, with the National Health Service (NHS) having spent £6.8 billion on CVD treatment in 2012/13.[[2](#_ENREF_2)] The prominent risk factors for CVD are well established. While many of these cannot be altered (e.g. age, gender, family history) there are numerous examples of modifiable risk factors including smoking, alcohol consumption and obesity. These represent targets for primary prevention strategies which can significantly reduce the risk of an individual developing CVD.

## CVD Risk Factors

A non-systematic literature search was conducted for known CVD risk factors. These are shown in the following table, with associated references (Table 1):

Table 1: CVD risk factor list

|  |  |
| --- | --- |
| Risk factor | References |
| Age | [[3](#_ENREF_3)] [[4-8](#_ENREF_4)] |
| Gender | [[9-20](#_ENREF_9)] [[21](#_ENREF_21)] |
| Ethnicity | [[22](#_ENREF_22) [23](#_ENREF_23)] [[24](#_ENREF_24) [25](#_ENREF_25)] [[26-28](#_ENREF_26)] |
| Smoking |  |
| Diabetes |  |
| Hypertension |  |
| Dyslipidaemia |  |
| Obesity |  |
| Physical activity |  |
| Family history |  |
| Socioeconomic status |  |
| Previous stroke/TIA |  |
| Atrial fibrillation |  |
| Inflammatory markers |  |
| Hyperviscosity/Hypercoagulable state |  |
| Hyperchromocysteinaemia |  |
| Chronic renal insufficiency |  |

### Age

Age is the most significant non-modifiable risk factor for developing cardiovascular disease. Data from the WHO and UN suggest that the risk of mortality from ischaemic heart disease triples with each decade of life (2.3-2.7-fold increase per decade for men, 2.9-3.7 for women).[[3](#_ENREF_3)] Similar gradients are seen in carotid disease and peripheral arterial disease,[[4-8](#_ENREF_4)] with risk of developing these conditions increasing exponentially with age. The mechanisms for this are debatable, but may include diminishing structural and mechanical integrity of the heart and blood vessels as well as the increased risk of exposure to other common risk factors.

### Gender

Men are at higher risk of developing CVD than women, although the risk for women increases substantially after menopause.[[9-20](#_ENREF_9)] Over time and at different ages, independent of diagnostic and treatment practices, women have a similar or slightly higher prevalence of angina than men across countries with widely differing myocardial infarction mortality rates.[[19](#_ENREF_19)] There are also gender-specific differences in the effects of other risk factors. Guideline-recommended treatments for angina are underused in women and older patients.[[21](#_ENREF_21)] These suboptimal practice patterns, which are worst in older women, are of particular concern. The prevalence of minor ECG changes is slightly higher among men (10.4%v 9.5% in women). The occurrence of ischaemia-like findings on the ECG was comparable between men and women (9.0% v 9.8%).[[20](#_ENREF_20)]

### Ethnicity

Migrants of South Asian descent worldwide have elevated risks of morbid and mortal events because of ischaemic heart disease.[[22](#_ENREF_22) [23](#_ENREF_23)] In the UK, mortality from IHD in both South Asian men and women is 1.5 times that of the general population, and South Asians have not benefited to the same extent from the general decline in deaths caused by IHD over the last few decades.[[24](#_ENREF_24) [25](#_ENREF_25)] Declines in stroke incidence have been observed the UK in men, women, White groups, and those aged >45 years, but not in Black groups. The reduction in prevalence of before-stroke risk factors was mostly seen in White patients aged >55 years, whereas an increase in diabetes mellitus was observed in younger Black patients.[[26](#_ENREF_26) [27](#_ENREF_27)] Incidence rates of first ever stroke adjusted for age and sex are twice as high in Black people compared with White people. This excess incidence cannot be accounted for by differences in social class in ages 35-64. Black people tend to have their first stroke at a younger age than White people. The excess incidence is found in all pathological types of stroke but is greatest for primary intracerebral haemorrhage.[[28](#_ENREF_28)]

### Hypertension

Hypertension ranks among the most significant risk factors for cardiovascular disease. Willey et al. estimated hypertension to account for almost one quarter of the population risk of CVD (population attributable risk (PAR) = 24.3%, 95% CI 13.2-35.4%).[[29](#_ENREF_29)] Severe hypertension (systolic blood pressure, SBP≥160mm Hg) has been associated with a two-fold increase in risk of CHD.[[30](#_ENREF_30)] The risk among pre-hypertensive individuals (BP 120-139/80-89 mm Hg) is debatable, with different studies drawing different conclusions. However, Huang et al. recently performed a meta-analysis including over 500,000 individuals from 17 studies and demonstrated a significant association of pre-hypertension with risk of CHD (RR 1.43, 95% CI 1.26-1.63).[[25](#_ENREF_25)] Hypertension has a stronger association with stroke, with PAR estimates as high as 36%.[[5](#_ENREF_5)] Sun et al. showed an increased risk of stroke across several different categories of hypertension, with combined systolic and diastolic hypertension presenting the greatest risk (OR 2.13, 95% CI 1.78-2.75). Similarly for PAD, Fowkes et al. calculated an odds ratio of 1.47 (95% CI 1.37-1.57) among hypertensive patients (BP ≥ 140/90 mm Hg) compared to normotensive subjects.[[7](#_ENREF_7)]

### Smoking

Smoking has long been established as a risk factor for CHD, and represents perhaps the most easily modifiable risk factor for CVD. In ranking the major risk factors for coronary heart disease, Schnohr et al. determined that smoking was the most important risk factor for women (OR 1.74, 95% CI 1.43-2.11) and the fifth most important for men (OR 1.62, 95% CI 1.25-2.09) when comparing non-smokers to individuals smoking >15g of tobacco per day.[[30](#_ENREF_30)] Similarly, Tolstrup et al. showed a significant correlation between the amount of tobacco smoked and the odds of developing coronary heart disease in a pooled analysis, and that the risk was higher for women than in men.[[31](#_ENREF_31)] Furthermore, there is ample evidence suggesting that cessation of smoking greatly reduces the risk of developing CHD.[[32](#_ENREF_32)] Similar associations are seen in stroke, with Peters et al. reporting an 83% increased risk of stroke among women who currently smoke compared to those who have never smoked, and 67% increase among men.[[13](#_ENREF_13)] There is an approximate two-fold increase in the risk of PAD for both current smokers (OR 2.09, 95% CI 1.91-2.29) and former smokers (OR 1.87, 95% CI 1.64-2.18) compared to non-smokers.[[7](#_ENREF_7)]

### Diabetes

Diabetes is associated with a 2 to 4-fold increase in the risk of developing cardiovascular disease.[[7](#_ENREF_7) [8](#_ENREF_8) [11](#_ENREF_11) [29](#_ENREF_29) [33-35](#_ENREF_33)] While this association may be somewhat confounded by the fact that diabetes is linked to several other CVD risk factors (e.g. obesity, hypertension, dyslipidaemia), there is evidence to suggest that diabetes is an independent risk factor. There is a graded positive association between HbA1c levels and risk of CHD, which holds true after adjustment for blood pressure, BMI and blood lipids, suggesting that poor blood glucose control increases risk of CHD.[[36](#_ENREF_36)] (Zhao et al. 2014). Yusuf et al. estimated the odds of an individual with diabetes having a heart attack are 2.37 times greater (99% CI 2.07-2.71) than a non-diabetic when adjusting for other factors.[[34](#_ENREF_34)]

The risk for stroke is similarly elevated, as studies show an approximately two-fold increase in the risk of stroke in diabetics compared to non-diabetics. Data from the Clinical Practice Research Datalink (CPRD) in the UK and estimated an odds ratio 2.19 (95% CI 2.09-2.32).[[37](#_ENREF_37)] Peters et al. determined a similar risk in a recent meta-analysis, while noting an increased risk in women compared to men (OR 2.28, 95% CI 1.93-2.69 vs. 1.83, 95% CI 1.60-2.08).[[12](#_ENREF_12)] Diabetes is also a key risk factor for peripheral arterial disease. Estimates range from 10-40% for the prevalence of PAD among individuals with diabetes, and PAD is the leading cause of amputation in diabetics.[[38](#_ENREF_38)] Fowkes et al. showed a 68% increased risk of PAD among diabetic subjects in their meta-analysis (OR 1.68, 95% CI 1.53-1.84).[[7](#_ENREF_7)]

### Obesity

We are now beginning to understand the underlying mechanisms as well as the ways in which smoking and dyslipidaemia increase, and physical activity attenuates, the adverse effects of obesity on cardiovascular health.Obesity is associated with other risk factors – diabetes, dyslipidaemia etc- which may mediate its effect on CVD. [[39](#_ENREF_39)] Adipose tissue releases a large number of bioactive mediators that influence not only body weight homeostasis but also insulin resistance - the core feature of type 2 diabetes - as well as alterations in lipids, blood pressure, coagulation, fibrinolysis and inflammation, leading to endothelial dysfunction and atherosclerosis.[[40](#_ENREF_40)] Body mass index (BMI) shows a positive association with cerebrovascular risk which is non-significant after adjustment for physical inactivity, smoking, hypertension, and diabetes (odds ratio 1.18; 95% CI, 0.77 to 1.79).[[41](#_ENREF_41)] Markers of abdominal adiposity are strongly associated with the risk of stroke/TIA. For the waist-to-hip ratio, adjusted odds ratios for every successive tertile were greater than that of the previous one.

### Dyslipidaemia

High total cholesterol, low high density lipoproteins (HDL), and high low density lipoproteins (LDL) are all well-established CVD risk factors and are included in both the Framingham and QRISK2 CVD risk scores.[[42-44](#_ENREF_42)]

### Deprivation and socioeconomic status

In the UK material deprivation is a well-established CVD risk factor and is included in the QRISK2 CVD risk score.[[45](#_ENREF_45) [46](#_ENREF_46)] Favourable population-wide trends in smoking, blood pressure and cholesterol are consistent with falling CHD death rates. However, adverse trends in obesity and diabetes in deprived populations are likely to counteract some of these gains. Furthermore, little progress over the last 15 years has been made towards reducing risk factor inequalities.[[47](#_ENREF_47)] Approximately half the recent CHD mortality fall in England is attributable to improved treatment uptake. This benefit occurred evenly across all social groups. However, the opposing trends in major risk factors means that their net contribution amounted to just over a third of the CHD deaths averted; and these also varied substantially by socioeconomic group.[[48](#_ENREF_48)]

### Physical Activity

In a recent meta-analysis, compared with individuals reporting no leisure time physical activity, there was a 20% lower overall mortality risk among those performing more than the recommended minimum, a 31% lower risk at 1 to 2 times the recommended minimum, and a 37% lower risk at 2 to 3 times the minimum. A similar dose-response relationship was observed for mortality due to CVD.[[49](#_ENREF_49)] Sedentary time is independent of physical activity associated with an increased risk of diabetes, CVD and CVD mortality.[[50](#_ENREF_50)] Compared with inactive individuals, those who exercise for an average of 15 min a day have a 14% reduced risk of all-cause mortality, and a 3 year longer life expectancy.[[51](#_ENREF_51)].

### Inflammatory Markers

Although there is a growing evidence base showing the importance of other inflammatory marker data, [[52](#_ENREF_52)] they are either not available in the national data sources available for use, or (more commonly) data are not available for small local populations, so we have not considered them further here.

### Chronic Kidney Disease

Chronic kidney disease (CKD) is associated with a number of other CVD risk factors, notably hypertension and diabetes, which contribute to the increased risk of CVD observed in CKD patients. However CKD has been shown to be an independent risk factor for CVD after adjustment for these related factors.[[53](#_ENREF_53)] There is a relative risk of 1.4 (95% CI 1.3-1.5) for CHD development in individuals with CKD.[[54](#_ENREF_54)] The data on stroke is somewhat less clear, with different studies arriving at dissimilar conclusions on the independence of CKD as a risk factor.[[55](#_ENREF_55)] However in a large meta-analysis, Lee et al. showed a statistically significant 43% increase in risk of stroke among patients with a glomerular filtration rate (eGFR) < 60 mL/min/1.73m2 (RR 1.43, 95% CI 1.31-1.57), which held even after adjustment for traditional risk factors.[[56](#_ENREF_56)] CKD is an established risk factor for PAD, with a two-fold increase in odds of developing PAD among patients with an eGFR < 60 mL/min/1.73m2 (OR 2.0, 95% CI 1.4-2.7).[[6](#_ENREF_6)]

Table 2 summarises all the CHD risk factors we reviewed with their recent pooled, matched or adjusted odds ratios.

Table 2: CVD risk factors with their pooled, matched or adjusted odds ratios

| Risk factor | Type of Odds Ratio & references | Odds Ratio | 95% CI | Effect on Outcome |
| --- | --- | --- | --- | --- |
| Hypertension | HR (adjusted)[[30](#_ENREF_30)] |  |  |  |
| SBP<120 mm Hg |  | 1.00 |  | Reference |
| SBP 120-139 mm Hg |  | 1.32 | [0.96-1.82] | NS |
| SBP 140-159 mm Hg |  | 1.63 | [1.18-2.24] | Risk Factor |
| SBP≥160 mm Hg or BP medication |  | 2.07 | [1.48-2.88] | Risk Factor |
|  |  |  |  |  |
| Smoking | HR (adjusted)[[30](#_ENREF_30)] |  |  |  |
| Never smoker  |  | 1.00  |  | Reference |
| Former smoker |  | 1.52 | [1.18-1.97] | Risk Factor |
| 1-4g tobacco/day |  | 1.63 | [0.91-2.94] | NS |
| 5-14g tobacco/day |  | 1.46 | [1.09-1.95] | Risk Factor |
| ≥15g tobacco/day |  | 1.62 | [1.25-2.09] | Risk Factor |
|  |  |  |  |  |
| Diabetes | For risk of MI – adjusted for all other risk factors[[34](#_ENREF_34)] |  |  |  |
| No |  | 1.00 |  | Reference |
| Yes |  | 2.37 | [2.07-2.71] (99% CI) | Risk Factor |
| Total cholesterol | HR (adjusted)[[30](#_ENREF_30)] |  |  |  |
| 1st quartile |  | 1.00 |  | Reference |
| 2nd quartile |  | 1.09 | [0.88-1.35] | NS |
| 3rd quartile |  | 1.18 | [0.95-1.46] | NS |
| 4th quartile |  | 1.77 | [1.42-2.21] | Risk Factor |
| HDL cholesterol | HR (adjusted)[[30](#_ENREF_30)] |  |  |  |
| ≥1.5 mmol/L |  | 1.00 |  | Reference |
| 1.0-1.4 mmol/L |  | 1.39 | [1.18-1.64] | Risk Factor |
| <1.0 mmol/L |  | 1.65 | [1.30-2.09] | Risk Factor |
| Family history | HR (adjusted)[[30](#_ENREF_30)] |  |  |  |
| No  |  | 1.00 |  | Reference |
| Yes |  | 1.49 | [1.04-2.14] | Risk Factor |
| Physical activity | HR (adjusted)[[30](#_ENREF_30)] |  |  |  |
| High activity in leisure time |  | 1.00 |  | Reference |
| Moderate activity in leisure time |  | 1.00 | [0.85-1.18] | NS |
| Low activity in leisure time |  | 1.30 | [1.03-1.65] | Risk Factor |
| Obesity | RR (adjusted) |  |  |  |
| BMI 18.5-22.9 (Men) |  | 1.00 |  | Reference |
| BMI 23.0-24.9 (Men) |  | 1.22  | [1.04-1.43] | Risk Factor |
| BMI 25.0-26.9 (Men) |  | 1.53 | [1.31-1.78] | Risk Factor |
| BMI 27.0-29.9 (Men) |  | 1.71  | [1.44-2.02] | Risk Factor |
| BMI 30+ (Men) |  | 1.81 | [1.48-2.22] | Risk Factor |
| BMI 18.5-22.9 (Women) |  | 1.00 |  | Reference |
| BMI 23.0-24.9 (Women) |  | 1.10 | [0.93-1.30] | NS |
| BMI 25.0-26.9 (Women) |  | 1.34 | [1.11-1.61] | Risk Factor |
| BMI 27.0-29.9 (Women) |  | 1.53 | [1.27-1.84] | Risk Factor |
| BMI 30+ (Women) |  | 2.16 | [1.81-2.58] | Risk Factor |
| Inflammatory Markers |  |  |  |  |
| IL-6 | HR per 1-SD increase (adjusted)[[57](#_ENREF_57)] | 1.46 | [1.30-1.64] | Risk Factor |
| CRP | RR per 3-fold increase (adjusted) (Emerging risk factors collaboration – 2010)[[52](#_ENREF_52)] | 1.64 | [1.54-1.75] | Risk Factor |
| Chronic Kidney Disease | Relative rate (adjusted)[[54](#_ENREF_54)] | 1.4 | [1.3-1.5] | Risk Factor |

Peters et al have conducted several systematic reviews, one for each major stroke risk factor.[[12](#_ENREF_12) [13](#_ENREF_13)] Table 3 provides a summary of stroke risk factors specifically with their pooled, matched or adjusted odds ratios from these and other reviews.

Table 3: stroke risk factors with their pooled, matched or adjusted odds ratios

| Risk Factor | Type of Odds Ratio & references | Odds Ratio | 95% CI | Effect on Outcome |
| --- | --- | --- | --- | --- |
| Sex |  |  |  |  |
| Women  | HR (adjusted)[[13](#_ENREF_13)] | 1.00 |  | Reference |
| Ethnicity |  |  |  |  |
| Men (black) |  | 1.39 | [1.12-1.74] | Risk Factor |
| Men (white) |  | 1.50 | [1.25-1.78] | Risk Factor |
| Hypertension | RR (adjusted) |  |  |  |
| Normotensive |  | 1.00 |  | Reference |
| Isolated systolic hypertension |  | 1.42 | [1.16-1.75] | Risk Factor |
| Isolated diastolic hypertension |  | 1.58 | [1.13-2.20] | Risk Factor |
| Systolic and diastolic hypertension |  | 2.13 | [1.78-2.75] | Risk Factor |
| Managed hypertension |  | 1.19 | [0.48-2.94] | NS |
| Smoking |  |  |  |  |
| Not current | RR (Pooled)[[13](#_ENREF_13) [37](#_ENREF_37)] | 1.00 |  | Reference |
| Current (Women) |  | 1.83 | [1.58-2.12] | Risk Factor |
| Current (Men) |  | 1.67 | [1.49-1.88] | Risk Factor |
| Never |  | 1.00 |  | Reference |
| Former (Women) |  | 1.17 | [1.12-1.22] | Risk Factor |
| Former (Men) |  | 1.08 | [1.03-1.13] | Risk Factor |
| Diabetes |  |  |  |  |
| No | HR (adjusted) )[[37](#_ENREF_37)] | 1.00 |  | Reference |
| Yes |  | 2.19 | [2.09-2.32] | Risk Factor |
| No (Women) | RR (Pooled) [[13](#_ENREF_13)] | 1.00 |  | Reference |
| Yes (Women) |  | 2.28 | [1.93-2.69] | Risk Factor |
| No (Men) |  | 1.00 |  | Reference |
| Yes (Men) |  | 1.83 | [1.60-2.08] | Risk Factor |
| Dyslipidaemia |  |  |  |  |
| Men ≤ 45 years | OR (adjusted)[[58](#_ENREF_58)] | 1.65 | [0.84-3.22] | NS |
| Men 45-65 years |  | 1.19 | [0.95-1.50] | NS |
| Men ≥66 years |  | 1.63 | [1.27-2.08] | Risk Factor |
| Women ≤ 45 years |  | 1.23 | [0.28-5.52] | NS |
| Women 45-65 years |  | 0.85 | [0.61-1.19] | NS |
| Women ≥66 years |  | 1.18 | [0.95-1.46] | NS |
| Obesity |  |  |  |  |
| BMI 30.0–34.9 | AOR, [[41](#_ENREF_41)] | 2.77 (1.52 to 4.24) |  |  |
| Atrial Fibrillation |  |  |  |  |
| Age 50-59 | OR (adjusted) – Risk for Ischemic stroke[[59](#_ENREF_59) [60](#_ENREF_60)] | 2.7 | [1.5-4.7] | Risk Factor |
| Age 60-69 |  | 1.9 | [1.4-2.4] | Risk Factor |
| Age70-79 |  | 1.8 | [1.6-2.1] | Risk Factor |
| Age 80-89 |  | 2.1 | [1.9-2.3] | Risk Factor |

## Stroke prevalence from the literature

Worldwide, a systematic analysis for the Global Burden of Disease (GBD) Study 2013 showed CVD made up 21,177 (14 947 to 28 436) years of life disabled in 2013.[[1](#_ENREF_1)] In the UK, GBD showed that the main drivers for improvement in life expectancy in nearly all countries and in the English regions have been declines in CVD and, to a lesser extent, cancer mortality.[[61](#_ENREF_61)] Throughout the UK, prevalence of stroke in men was almost twice as high than that for women in 2013, although this data was obtained from GP diagnoses in the Clinical Practice Research Datalink (CPRD), which underestimates actual prevalence (Table 4).[[2](#_ENREF_2)] Applying country-specific and age-specific population estimates, obtained from national statistics agencies, to prevalence data from CPRD suggests that >915 000 people in the UK have suffered an MI and >1.3 million are living with angina.[[2](#_ENREF_2)] Consequently, if we combine estimates for MI and angina, we find that almost 2.3 million people in the UK are living with some form of CHD.[[19](#_ENREF_19)]

Table 4: prevalence of cardiovascular conditions, UK 2013 (from CPRD)

|  | MI | Angina | Heart failure | Atrial fibrillation | Stroke |
| --- | --- | --- | --- | --- | --- |
| *Men* |
| 0–44 | 0.06 | 0.05 | 0.05 | 0.09 | 0.11 |
| 45–54 | 1.14 | 0.92 | 0.33 | 0.76 | 0.89 |
| 55–64 | 3.55 | 3.60 | 1.12 | 2.28 | 2.69 |
| 65–74 | 7.05 | 8.83 | 2.92 | 6.20 | 6.40 |
| 75+ | 12.08 | 16.96 | 7.84 | 15.38 | 14.89 |
| All ages | 2.46 | 3.05 | 1.22 | 2.47 | 2.53 |
| *Women* |
| 0–44 | 0.02 | 0.03 | 0.04 | 0.03 | 0.11 |
| 45–54 | 0.29 | 0.50 | 0.15 | 0.26 | 0.79 |
| 55–64 | 0.89 | 1.74 | 0.45 | 0.91 | 1.96 |
| 65–74 | 2.06 | 4.66 | 1.32 | 3.28 | 4.39 |
| 75+ | 5.50 | 11.15 | 5.89 | 11.71 | 12.43 |
| All ages | 0.87 | 1.79 | 0.76 | 1.56 | 1.99 |
| *Number of cases in sample* |
| Men | 47 449 | 57 927 | 22 954 | 46 597 | 47 888 |
| Women | 19 747 | 41 840 | 18 201 | 36 967 | 46 54 |

Prevalence estimates of stroke worldwide are shown in Table 5.

Table 5: prevalence estimates of stroke worldwide

| Publication | Country | Type of study | Total\* | Male\* | Female\* | Population age (years) |
| --- | --- | --- | --- | --- | --- | --- |
| Bensenor et al. 2015 | Brazil | Cross-sectional | 1.5\*\* | 1.6\*\* | 1.4\*\* | ≥18 |
| MMWR 2012 | USA | Cross-sectional | 2.6\*\* | 2.7\*\* | 2.6\*\* | ≥18 |
| Chang et al. 2015 | Sri Lanka | Cross-sectional | 10\* | 15\* | 7\* | ≥18 |
| Onwuchekwa et al. 2014 | Nigeria | Cross-sectional | 8.5\* | 12.9\* | 5.1\* | ≥15 |
| D’Allessandro et al. 2009 | Italy | Cross-sectional | 1.47\*\* | 1.33\* | 1.59\*\* | ≥15 |
| BHF 2009 | UK | Cross-sectional | 1.7\*\* | NR | NR | All |

\* Prevalence (cases/103 inhabitants)

\*\* Crude rates

Prevalence of stroke by age, sex and UK country is shown in Table 6 from the British Heart Foundation (2014).[[2](#_ENREF_2)]

Table 6: Prevalence of stroke by age and sex from British Heart Foundation (2014)

|  |  |  |
| --- | --- | --- |
|  | Female age groups (yr) | Male age groups (yr) |
|  | **≤44** | **45-54** | **55-64** | **65-74** | **≥75** | **≤44** | **45-54** | **55-64** | **65-74** | **≥75** |
| England | 0.11 | 0.75 | 1.80 | 4.16 | 12.17 | 0.10 | 0.85 | 2.60 | 6.08 | 14.55 |
| Scotland | 0.11 | 0.96 | 2.57 | 5.33 | 14.05 | 0.13 | 1.14 | 3.26 | 7.51 | 16.92 |
| Wales | 0.11 | 0.86 | 2.18 | 4.77 | 12.27 | 0.12 | 0.90 | 2.70 | 7.15 | 15.21 |
| Northern Ireland | 0.13 | 0.92 | 2.43 | 5.26 | 13.74 | 0.10 | 0.88 | 2.82 | 7.39 | 15.42 |
| United Kingdom | 0.11 | 0.79 | 1.96 | 4.39 | 12.43 | 0.11 | 0.89 | 2.69 | 6.40 | 14.89 |

# Methods

## “Data discovery” from UK survey data sources

We considered four potential data sources from different health surveys conducted in the UK. Using these we conducted a data discovery procedure, comparing each source for their coverage of risk factor and outcome variables for CVD. The results are summarised in Table 5.

### Whitehall II

The Whitehall II study is a prospective cohort study of individuals employed by the British Civil Service at the time of recruitment, between 1985 and 1988. This study follows from the first Whitehall study initiated 1967, with the broad aim of determining disparities in health equality as they relate to socioeconomic factors. The first phase recruited 10,308 individuals (3414 women and 6985 men) between the ages of 35 and 55 years. The cohort is invited to clinical research screenings at 5-year intervals with postal questionnaires sent to participant between clinical screenings. Telephone questionnaires are administered to those unwilling or unable to attend the clinic or complete the full postal questionnaire, and as of phase 7 home nurse visits were offered to those unable or unwilling to attend the clinic. To date there are nine waves of data collection available.

Questionnaires cover a broad range of health, socioeconomic, and demographic information, including self-reported disease diagnosis, medication, as well as food frequency questions. Clinical measures include a wide range of biomarkers from blood samples, anthropometric measurements, and blood pressure. Participants are also administered an electrocardiogram. Clinical records for cardiovascular events self-reported on the questionnaire or determined by ECG are verified through hospital and primary care records. Mortality is followed through the NHS Central Registry, providing date and cause of death for Whitehall study participants.

### Airwave Health Monitoring Study

The Airwave Health Monitoring Study is a prospective cohort study commissioned in 2003 for the original purpose of assessing the long-term health effects of the Terrestrial Trunked Radio (TETRA) communication system among British police officers. It has subsequently been expanded to look at the general health of the police force in Britain. Participation is open to all 54 police forces in Great Britain, and all police staff are eligible to enrol. Since the initial pilot studies began in 2004, 28 forces have since been enrolled. The organisers aim to recruit a total of 60,000 participants in total. As of December 2012, there were 42,897 individuals in the database, of which 42,112 have been linked to National Health Service (NHS) records.

Data collection for the Airwave study is divided into two phases. The first phase consists of a questionnaire to be completed upon enrolment in the study. Questions cover demographic, health and lifestyle information. The second phase involves a second, more in depth questionnaire as well as a health screening. Nurses performed clinical examination of participants, collecting measurements of blood pressure, arterial stiffness, weight, height, waist and hip circumference and body composition as well as administering and electrocardiogram (ECG). Systolic and diastolic blood pressure (SBP and DBP) were also measured. Blood samples are collected, providing biomarkers including lipid profiles, inflammatory markers, and measures of glucose tolerance.

### Health Survey for England 2013

Health Survey for England (HSfE) is an annual cross-sectional survey carried out in England since 1991, designed to assess changes in health outcomes and lifestyle among the British population. The survey design incorporates both children and adults, representative of the general population at a national and regional level. Households and individuals are randomly selected. The 2013 survey included 8795 adults (aged 16 years and over) and 2185 children (age 0-15 years), of which 6183 adults and 1455 children also received a nurse visit for clinical measurements. These included anthropometric and blood pressure measurements as well as blood samples for various biomarkers. The survey questionnaire relies largely on self-reported diagnosis of medical conditions as well as various risk factors.

### English Longitudinal Study of Ageing

The English Longitudinal Study of Ageing (ELSA) is a prospective cohort study designed to assess the health and well-being of the English population over the age of 50. Study participants for the initial wave were drawn from HSfE who were over the age of 50 at the time of recruitment, and included 12,099 individuals aged 50 to 100 years. The study population was refreshed during the third wave of data collection to replenish the sample in the younger age range. Participants complete a face-to-face interview and self-completion questionnaire every two years, and are assessed by a nurse every four years. To date, seven waves of data are available. Questionnaires cover a broad range of characteristics related to physical and mental health, economics and general well-being. The nurse assessment includes clinical measurements of anthropometry, blood pressure, lung function, as well as taking a blood sample which provides data on various biomarkers of disease.

Table 7 shows coverage of risk factor variables from the potential data sources we examined.

Table 7: coverage of risk factor variables from potential data sources

| Risk Factor | Whitehall II | Airwave | HSfE 2013 | ELSA |
| --- | --- | --- | --- | --- |
| Age | Yes – May need to be derived | Yes | Yes – age and age band | Yes |
| Gender | Yes | Yes | Yes | Yes |
| Ethnicity | Yes- Only white/non-white | Text variable from questionnaire | Yes – 18 categories | Yes – Only white/non-white |
| Smoking | Yes/No; cigarettes/day; current/former | Yes/No; cigarettes/day; current/former | Yes/No; cigarettes/day; current/former; cotinine as biomarker | Yes/No; cigarettes/day; current/former |
| Diabetes | Self-reported diagnosis; HbA1c, HOMA-IR, FPG clinical measurements; family history; medication | Self-reported diagnosis – type 1 vs. type 2; HbA1c, FPG and C-peptide clinical measurements | Self-reported diagnosis – type 1 and type 2; HbA1c clinical measure; medications | Self-reported diagnosis; HbA1c and FPG clinical measures; medication |
| Hypertension | Self-reported diagnosis and treatment; SBP and DBP clinical measures; family history | Self-reported diagnosis; clinical SBP and DBP measures; diet questionnaire | Self-reported diagnosis; SBP, DBP and MAP clinical measures; comprehensive medication questions | Self-reported diagnosis; SBP, DBP and MAP clinical measures; self-reported medications |
| Dyslipidaemia | Clinical measures for Total, HDL, IDL, LDL, Apolipoproteins A and B, and triglycerides; diet questionnaire | Self-reported diagnosis; diet questionnaire; Total cholesterol, HDL, Apolipoproteins A and B clinical measures (No triglycerides) | Total cholesterol and HDL clinical measure; high vs. low TC and HDL from clinical; medication | Clinical measures for Total, HDL, LDL, and triglycerides |
| Obesity | BMI, fat mass, fat percentage, waist circumference and WHR clinical measures | Body fat %, impedance waist circumference, WHR, BMI clinical measures | Self-reported; BMI, WHR, WC clinical measures (pre-grouped) | BMI, waist circumference and WHR clinical measures |
| Family History | Angina/MI and stroke questionnaire for parents/siblings; nothing for PAD | Asks if parents died from MI or stroke (no non-fatal, sibling or anything on PAD) | NO VIABLE VARIABLES | Cause of death of parents |
| Physical Activity | METs; breakdown by activity based on self-reports | Self-reported frequency of activity/inactivity | Self-reported time spent (pre-grouped) | Self-reported frequency |
| Alcohol Consumption | Self-reported Yes/No, units per week  | Self-reported Yes/No, frequency, units | Self-reported Yes/No, frequency, units | Maximum consumed; Yes/No |
| Inflammatory Markers | CRP and IL-6 clinical measures | CRP clinical measure | NO VIABLE VARIABLES | CRP clinical measure |
| Hyper-viscosity/hyper-coagulable state  | Factor VII, Fibrinogen, plasma viscosity, von Willebrand’s factor clinical measures | Fibrinogen, prothrombin time, haematocrit | Self-reported clotting disorder; self-reported anti-platelet medication | Hemoglobin and fibrinogen clinical measures; self-reported clotting disorder |
| Hyperchromocysteinaemia | NO VIABLE VARIABLES | NO VIABLE VARIABLES | NO VIABLE VARIABLES | NO VIABLE VARIABLES |
| Chronic Renal Insufficiency | Diet questionnaire; clinical EGFR measurements | Creatinine clinical measure; Diet questionnaire | NO VIABLE VARIABLES | Self-reported kidney trouble |
| Socioeconomic Status | IMD 2004 and 2007; Townsend Index |  | IMD score |  |

Table 8 shows coverage of outcome variables from the potential data sources we examined.

Table 8: coverage of outcome variables from potential data sources

| Outcome | Whitehall II | Airwave | HSfE 2013 | ELSA |
| --- | --- | --- | --- | --- |
| Coronary Heart Disease | Self-reported angina; Self-reported CABG operation; Self-reported heart disease risk; Self-reported hospital admission; Self-reported ischemia; Self-reported nitrates; ECG (results and Minnesota codes)Clinical records of hospitalisationSelf-reported angiogram/angioplasty??Self-reported chest pain??Self-reported other heart troubles??Clinical heart rate variability measures?? | Self-reported diagnosis of angina/MISelf-reported CABG operationClinical ECG interpretationECG (results and Minnesota codes)Self-reported chest pain??Self-reported other heart tests/operation??Self-reported angiogram/angioplasty?? | Self-reported MI/angina diagnosis (longstanding illness)Self-reported heart/circulatory disease (non-specific)??Self-reported medication | Self-reported angina |
| Cerebrovascular Disease | Self-reported doctor diagnosisSelf-reported slurred speech, visual symptoms, weakness | Self-reported stroke /TIA diagnosis | Self-reported stroke (longstanding illness) | Self-reported stroke |
| Peripheral Arterial Disease | Self-reported diagnosis of ICSelf-reported leg painSelf-reported angiogram/angioplasty?? | NO VIABLE OUTCOME MEASURES | NO VIABLE OUTCOME MEASURES | Self-reported IC diagnosisSelf-reported leg pain whilst walking |

While all of the surveys contained variables for most of the risk factors, only Whitehall II and ELSA had means of assessing PAD as an outcome. Ultimately, it was decided to use Whitehall II as the data source for our model, as the PAD-related questions allowed us to more definitively classify individuals with this disease. Furthermore, Whitehall II has a wider range of variables for specific CVD risk factors, particularly with the clinical records and wider range of biomarker data.

## Definition of Outcomes

For this model we included both haemorrhagic and ischemic strokes as well as transient ischemic attacks (TIA) in defining the outcome of stroke. We developed a diagnostic algorithm based on literature review and expert advice. Figure 2 shows the diagnostic algorithm for ischemic stroke/TIA and Figure 3 shows the algorithm for haemorrhagic stroke. The potential corresponding variables available in Whitehall II are shown in Table X. Based on the data available, we defined stroke as follows:

1. Self-reported doctor diagnosis of stroke or TIA
2. Clinical record of stroke or TIA
3. Mortality from stroke

The chosen variables are indicated in Table X. Variables pertaining to the symptoms of stroke/TIA (e.g. visual symptom, sudden weakness) were not included as these were not specific enough to warrant a definitive diagnosis and would likely lead to an overestimate of the prevalence. As with CHD, diagnoses were accumulated over each phase of data collection.

## Risk factor variables

### Age

All participants provided their year of birth at baseline. We used this variable, ‘yob\_c’, to calculate the age of each participant as of the most recent phase of data collection in 2009. We then used this to generate the variable ‘**age\_gp**’, grouping individuals into age categories in 5 year bands: 55-59 years, 60-64 years, 65-69 years, 70-74 years, and 75-79 years.

### Sex

The variable ‘**sex**’ was used to classify individuals as male or female.

### Ethnicity

Whitehall II only classifies individuals’ ethnicity as either white or non-white. This is coded in the variable ‘ethn\_ds’, which was relabelled as **‘ethnicity**’.

### Hypertension

There are several possible ways of classifying individuals with hypertension in Whitehall II. Data collected from questionnaires contains information on self-reported doctor-diagnosed hypertension. In addition there is a derived variable for anti-hypertensive medication, based on self-reported prescribed medication. However, in order to prevent biased estimates from inaccurate self-reports, we chose to use the clinical measurements of systolic and diastolic blood pressure, which may identify undiagnosed cases of hypertension. Table 9 shows the Whitehall II variables related to hypertension. We combined each of these to create the binary variable ‘hypertension’ for use in regression analysis. shows the variables related to hypertension.

* ‘hypertension’ coded as 1 (Hypertension) if any of the following criteria were met:
1. Participant responded positively to question of hypertension diagnosis during any phase of data collection
2. Participant responded positively to being prescribed an anti-hypertensive drug during any phase of data collection
3. Systolic blood pressure reading was greater than 140 mmHg during any phase of data collection
4. Diastolic blood pressure reading was greater than 90 mmHg during any phase of data collection
* ‘hypertension’ coded as 0 (No hypertension) if none of the above criteria were met

Table 9: Whitehall II variables related to hypertension

| Variable Label | Phase 1 | Phase 2 | Phase 3 | Phase 4 | Phase 5 | Phase 6 |  | Phase 8 | Phase 9 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| High blood pressure – Ever told | bpup | zbpup |  | vhbp | thbp |  |  |  |  |
| Drug subclass – antihypertensives | antihyp | zantihyp | xantihyp | vantihyp | tantihyp | qantihyp |  | mantihyp |  |
|  |  |  |  |  |  |  |  |  |  |

### Dyslipidaemia

Whitehall II does not contain any questions for self-reported diagnosed dyslipidaemia. There is however a derived variable classifying self-reported medication into lipid lowering drugs as well as statins specifically. In addition, blood tests for total cholesterol, HDL and LDL cholesterol, and triglycerides are available from clinical assessments in phases 1, 3, 5, 7 and 9. Table 10 shows the Whitehall II variables related to dyslipidaemia. We combined these variables to generate the binary variable ‘**dyslipidaemia’** as follows:

* ‘dyslipidaemia’ coded as 1 (Dyslipidaemia) if any of the following criteria met:
1. Participant responded positively to being prescribed a lipid-lowering drug or statin during any phase of data collection
2. Blood cholesterol ≥6.2 mmol/L in any phase of data collection
3. HDL cholesterol ≤ 1 mmol/L in any phase of data collection
4. LDL cholesterol ≥ 4.1 mmol/L in any phase of data collection
5. Triglycerides ≥ 1.7 mmol/L in any phase of data collection
* ‘dyslipidaemia’ coded as 0 (No dyslipidaemia) if none of the above criteria were met

Table 10: Whitehall II variables related to dyslipidaemia

| Variable Label | Phase 1 | Phase 3 | Phase 4 | Phase 5 | Phase 6 | Phase 7 | Phase 8 | Phase 9 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Drug subclass: lipid lowering | drug3 |  | vlipdrg | tlipdrg | qlipdrg | mlipdrg | klipdrg | jlipdrg |
| Drug subclass: statins |  |  |  |  |  |  | kstatins |  |
| Blood cholesterol | blchol | xblchol |  | tblchol |  | mblchol |  | jblchol |
| High density lipoprotein | hdl | xhdl |  | thdl |  | mhdl |  | jhdl |
| Low Density Lipoprotein, mm |  | xldl |  | tldl |  | mldl |  | jldl |
| Triglyceride | trig | xtrig |  | ttrig |  | mtrig |  | jtrig |

### Family History (CHD)

Whitehall II variables related to family history of coronary heart disease come from questionnaire data, based on questions about whether parents or siblings have suffered from heart attack or angina. In addition, questions regarding parents’ cause of death contains an option for heart attacks. Table 11 shows Whitehall II variables related to family history of CHD.

Table 11: Whitehall II variables related to family history

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Label | Phase 1 | Phase 2 | Phase 7 |
| Father: cause of fathers death (=1 for heart attack) | codf | zcodf | mcodf |
| Mother: mothers cause of death (=1 for heart attack) | codm | zcodm | mcodm |
| Angina – Either parent suffered angina | angpar |  |  |
| MI – either parent suffered a heart attack | hapar |  |  |
| MI – any sibling suffered a heart attack | hasib | zhasib |  |

We generated the variable **‘famhx\_chd’** to indicate a whether a participant had a family history of heart disease:

* ‘famhx\_chd’ coded as 1 (Family history) if either parent or sibling suffered from heart attack or angina OR if either parent reported as having died from a heart attack in any phase of data collection
* ‘famhx\_chd’ coded as 0 (No family history) if no parent or sibling suffered heart attack or angina AND neither parent reported as having died from heart attack

### Smoking

There are a number of variables related to smoking status in Whitehall II, with questions on the number and type of tobacco smoked. We used the derived variable labelled ‘Smoking habits (never/ex/current)’ to classify each participant’s smoking status for our analysis. As with other variables we used the most recent wave for which information was available on each participant, as smoking habits may change over time. Table 12 shows Whitehall II variables related to smoking. Information was available is phases 1, 2, 3, 5, 7, and 9. We generated the variable ‘**smoke\_cat’** as follows:

* ‘smoke\_cat’ coded as 1 (Never Smoker) if participant classified as ‘never-smoker’ during most recent phase for which data is available for that individual
* ‘smoke\_cat’ coded as 2 (Former Smoker) if participant classified as ‘ex-smoker’ during most recent phase for which data is available for that individual
* ‘smoke\_cat’ coded as 3 (Current Smoker) if participant classified as ‘current smoker’ during most recent phase for which data is available for that individual

Table 12: Whitehall II variables related to smoking

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Label | Phase 1 | Phase 2 | Phase 3 | Phase 5 | Phase 7 | Phase 9 |
| Smoking habits (never/ex/current) | esmoke | zesomke | xesmoke | tesmoke | mesmoke | jesmoke |

 |  |  | 1 | ESMOKE | ZESMOKE | XESMOKE |  | TESMOKE |  | MESMOKE |  | JESMOKE |

### Obesity

There are several different ways of measuring obesity based on the clinical assessment. These include BMI, fat percentage, waist circumference and waist-hip ratio. We chose to use BMI as this is the most commonly used measurement of obesity in epidemiological studies on CVD. Using the variable labelled ‘Body mass index’ we categorised participants into five categories based on WHO classification guidelines (ref). Once again, we used the most recent phase for which data was available for each individual. Table 13 shows Whitehall II variables related to obesity. We generated the variable ‘**bmi\_cat’** as follows:

* ‘bmi\_cat’ coded as 1 (Underweight) if BMI <18.5 kg/m2 during most recent phase for which data is available for that individual
* ‘bmi\_cat’ coded as 2 (Normal Weight) if BMI ≥18.5 kg/m2 and ≤25 kg/m2 during most recent phase for which data is available for that individual
* ‘bmi\_cat’ coded as 3 (Overweight) if BMI >25 kg/m2 and ≤30 kg/m2 during most recent phase for which data is available for that individual
* ‘bmi\_cat’ coded as 4 (Obese) if BMI >30 kg/m2 and ≤35 kg/m2 during most recent phase for which data is available for that individual
* ‘bmi\_cat’ coded as 5 (Severely Obese) if BMI >35 during most recent phase for which data is available for that individual

 Table 13: Whitehall II variables related to obesity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Label | Phase 1 | Phase 3 | Phase 5 | Phase 7 | Phase 9 |
| Body mass index | bmi | xbmi | tbmi | mbmi | jbmi |

### Diabetes

There are several potential ways to ascertain whether an individual has diabetes in Whitehall II. In addition to self-reports from questionnaires there are clinical measures of HbA1c, fasting plasma glucose and HOMA-IR. We decided to use the derived variable ‘jdmincum’, which tabulates the total number of diabetes cases in the cohort. The variables ‘jdmincum’ was relabelled ‘**diabetes**’ and used for regression analysis.

### Chronic Kidney Disease

Chronic kidney disease is clinically diagnosed based on EGFR value ≤60 mL/min/1.73m2 (ref). EGFR was first assessed during clinical examination in Whitehall II only in the most recent phase (Phase 9). As such there is a substantial amount of missing data for this risk factor (n=4242, 41% missing). We used the variable ‘jegfr’ to create the binary variable ‘**ckd**’ as follows:

* ‘ckd’ coded as 1 (Chronic kidney disease) if jegfr ≤60 mL/min/1.73m2
* ‘ckd’ coded as 0 (No chronic kidney disease) if jegfr >60 mL/min/1.73m2

### “Hypercoagulable state”/on anti-coagulant

Clinical biomarkers for a number of different clotting factors are available in Whitehall II. However we were unable to find consensus criteria on the use of these in the clinical diagnosis of a hypercoagulable state. We therefore had to rely on self-reports of anti-coagulant medication use to identify individuals in a hypercoagulable state. This method may not be ideal, as individuals who have had heart attacks and stroke are often prescribed anti-coagulant medication as a treatment, creating a question of the direction of causality. Table 14 shows Whitehall II variables related to hypercoagulation We generated the variable ‘**hypercoagulation**’ based on self-reported medication use only. However due to the potential for confounding by reverse causality, we constructed models both including and excluding this variable:

* ‘hypercoagulation’ coded as 1 (Hypercoagulation) if participant reported being prescribed anti-coagulant or anti-platelet drugs during any phase of data collection. There was insufficient time for anticoagulants and antiplatelet agents to have been considered separately
* ‘hypercoagulation’ coded as 0 (No hypercoagulation) if participant never reported use of anti-coagulant or anti-platelet drugs

Table 14: Whitehall II variables related to hypercoagulation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable Label | Phase 1 | Phase 3 | Phase 4 | Phase 5 | Phase 6 | Phase 7 | Phase 8 | Phase 9 |
| Drug subclass: Anticoagulant drugs | drug4 |  |  |  |  |  |  |  |
| Drug subclass: Antiplatelets |  |  | vdrg29 | tdrg29 | qdrg29 | mdrg29 | kdrg29 | jdrg29 |
| Factor VII | fact7 | xfactor7 |  |  |  |  |  |  |
| Fibrinogen | fib | xfibrin |  | tfibrin |  |  |  |  |
| Plasma viscosity (plasma) |  |  |  | tplasvis |  |  |  |  |
| Von Willebrand’s factor |  | xvwf |  |  |  |  |  |  |

### Inflammatory Markers

Clinical biomarkers for CRP and IL-6 are available in Whitehall II. We restricted our analysis to CRP as this has a more defined cut-off for defining normal levels. Blood tests for CRP were performed in phases 3, 5 and 7, with at least one measurement available for 8467 (82%) individuals. We used a value of CRP≥3 g/L to define high levels of CRP. The variable **‘high\_crp**’ was generated as follows:

* ‘high\_crp’ coded as 1 (High CRP) if CRP ≥3 g/L in any phase of data collection
* ‘high\_crp’ coded as 0 (Normal CRP) if CRP < 3 g/L for all phases of data collection

However as we had no immediately available local data on anti-coagulant use we have not considered this variable further. We do have a database of practice-level prescribing data but there was insufficient time to use it.

### Socioeconomic status/deprivation

Whitehall II contains several indices of deprivation, including the Indices of Multiple Deprivation 2004 and 2007 (IMD 2004 and 2007) as well as the Townsend Index. While the Townsend Index has fewer missing values in the dataset, we chose IMD 2007 as this is more commonly used. Participants were categorised into quintiles based on their IMD 2007 score, and the variable **‘ses’** was created containing this information.

### Physical Activity

We intended to include physical activity in the modelling but abandoned this because the categories used in Whitehall II are too different from those in the local physical activity data provided by the Sport England *Active People* survey.

## Statistical Analyses

### Whitehall II missing data

When performing logistic regression with multiple variables, Stata performs list-wise deletion by default, so that any individual missing a value for any of the outcome or risk factor variables is excluded from the analysis. This can result in potential biases and a loss of statistical power to assess associations. While most of the risk factors have very few missing values, there are several with very poor coverage in the dataset e.g. many participants are missing an IMD score. In order to deal with this, we used multiple imputation to impute all missing values for risk factors. Surprisingly, because one of the objectives of Whitehall II was to analyse inequalities, a significant proportion of deprivation scores are missing.

### Whitehall II descriptive analyses

We performed a number of descriptive analyses on the patient-level dataset including demographics, risk factor breakdowns and categories.

### Whitehall II regression modelling

We fitted univariate then multivariate logistic regression models as described in previous publications, to produce odds ratios (ORs) and regression coefficients.[[62](#_ENREF_62)] A range of multivariate regression models were fitted in order to obtain the best performing. We included one additional variable at a time to observe the effects. We performed backwards and forwards stepwise variable selection in the multivariable models. Stata has a *stepwise* command which automates this procedure but we perfomed variable selection manually using likelihood ratios and Wald tests because we had used multiple imputation of missing data.

Deprivation has been shown to be a risk factor for CVD. For example, QRISK2, the individual CVD predictive risk score recommended by NICE, found that deprivation (in the form of Townsend scores) and ethnicity were separate risk factors.[[42](#_ENREF_42)] The variables included in the final model are also determined by the availability of local data to match with the model variables. Hence variable selection has to be a compromise between the best model which can be produced from Whitehall II data and the local variable available.

We have over time increased the number of variables used in the local models as more local data has become available. However as more variables are added we need to take account of the joint effects of multiple risk factors, i.e. it assumes they operate independently. Estimation of the joint effects of multiple risk factors is complex for several reasons. In particular, some of the effects of more distal risk factors are mediated through intermediate factors.

Finally, there can be collinearity between exposure to various risk factors, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Collinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

### Interactions

There is an interaction between the effects of two exposures if the effect of one exposure varies according to the level of the other exposure.[[63](#_ENREF_63)] For example, there might be an interaction between the risk factors of education level and social class. An alternative term for interaction is effect modification. In this example, we can think of this as educational level modifying the effect of social class. The most flexible approach to examine interactions is to use regression models, but when using Mantel-Haenszel methods to control for confounding an alternative is to use a χ2 test for effect modification, commonly called a test of heterogeneity. Interaction, effect modification and heterogeneity are three different ways of describing the same thing. We compared log likelihoods between models excluding and including the interaction parameters to test the null hypothesis that there is no interaction between selected variables.

### Internal validation

We fitted a range of multivariate logistic regression models in order to obtain the best performing. We included one additional variable at a time to observe the effects. In order to obtain the most parsimonious models we then applied stepwise backward and forward variable selection using the *stepwise* command in Stata. Finally, we internally validated the models by generating receiver operating characteristic (ROC) curves, by using the ***predict*** regression post-estimation command to generate for each CPRD patient the probability of having back pain using the derived odds ratios (ORs), and by using these probabilities to examine sensitivity and specificity.

All statistical analysis was carried out in StataSE14.

## Local prevalence estimates

Derived ORs (or rather, regression coefficients) are used to estimate prevalence in small population subgroups. Local population breakdowns for each risk factor are used, where these are available. ICL has a wide range of small population risk factor prevalence breakdowns, including age, sex, deprivation, smoking, ethnicity, cardiovascular diseases and other disease conditions. The local model uses locally available data.

The “local” model includes only those variables that are available at local population level i.e. age, sex, socioeconomic status, BMI, smoking status, depression and other disease conditions. The steps in applying the prevalence estimates are as follows and in the equations below:

• Use the regression coefficients to generate log odds (since they are from a logistic regression model) for each risk factor subcategory

• Generate a similar table of odds by exponentiation

• Generate a similar table of prevalence in each risk factor subcategory using the epidemiologic formula

• Produce a matching table of small population subcategories. If there are no corresponding local data with a sufficiently granular breakdown e.g. ethnicity by age by sex, this requires deciding how each risk factor should be attributed across other risk factor categories, with evenly as the default. For example, we used the national age/sex/ethnicity breakdown from the Census and age/smoking breakdowns from the HSfE to attribute this data at small population levels. The actual breakdown will be somewhat different and needs to be borne in mind as another source of potential error.

• Multiply the population cells by the corresponding prevalence to estimate the number of people in each cell with the disease

In mathematical notation:

Predicted log odds of prevalence = *b0* + *b1x1i* +  *b2x2 i* + *b3x3 i* +  *b4x4 I*

where *b0* = regression constant, *b1, b2,  b3, b4*= other regression coefficients

*x 1 i, x2 i, x3 i, x4 i* = value of risk factors for individual ***i***

(NB since all the variables are binary variables, x =1 if specified risk factor is present, x=0 if it is absent). Predicted log odds of prevalence for a community of n individuals is derived by averaging over the values for all individuals included in the community:

Predicted log odds of prevalence in community of n individuals:

= 1/*n* ∑i=1n (*b0* + *b1x1i* +  *b2x2 i* + *b3x3 i* +  *b4x4 i)*

= *b0* + *b1p1* +  *b2p2* + *b3p3* +  *b4pp4*

where p1 , p2, p3, p4=proportion of individuals in the community with characteristic x1 , x2 , x3 , x4 . (i.e. proportion with x.=1 rather than x.=0 as in the remainder).

The predicted prevalence for an individual is derived from their predictive log odds using:

prevalence = exp(log odds)/[1+exp(log odds)]

= *exp*(*b0* + *b1x1i* +  *b2x2 i* + *b3x3 i* +  *b4x4 i)/[1+* *exp*(*b0* + *b1x1i* +  *b2x2 i* + *b3x3 i* +  *b4x4 i)]*

Predicted prevalence in community of n individuals:

*= 1/n ∑i=1n[exp(b0 +b1x1i +b2x2 i +b3x3 i +b4x4 i)/[1+ exp(b0 +b1x1i +b2x2 i +b3x3 i +b4x4 i)]]*

Unfortunately, the equation above does not simplify to a linear combination of the predictor variables (in the way the mean log odds does). The average/overall prevalence is not the same as the prevalence for a person with “average” risk factors. So, for instance, it cannot be found by taking exp(log odds)/[1+ exp(log odds)] of the average log odds. There is no linear relationship with the regression coefficients, and with proportions of population with specified risk factors.

In order to find a synthetic estimate of prevalence, ideally we need to know the joint distributions of the included risk factors in the relevant population (the population on which are synthetic estimates are required). Ideally, we would know how many people in the population have each specific combination of risk factors. In practice, it might be good enough to know the distribution of some risk factors individually, rather than in combination. For instance, we might know what proportion of the population are smokers, and what proportion are ex-smokers, but not how many smokers we have by age and sex. In this situation, we have assumed that the same proportion of all ages and both genders are smokers and ex-smokers. Even if this is not exactly correct, then the synthetic estimate of prevalence may still be a reasonably accurate estimate (assuming that the smoking distribution does not vary too much by age, sex and other included risk factors). This is considered a good enough approach, and the best possible based on the information currently available in many cases.

In practice, we know the population distributions by age and sex, therefore we do not need to make the assumption that the proportion of males is the same for each age group. We use the more precise method of using the actual proportions of males in each age group. From the ELSA longitudinal survey we also know that older people/ older females in particular are generally less educated (on the basis of qualifications held). Therefore we apply the proportions with any educational qualifications according to age and sex group.

For other risk factors, we do not know whether these risk factors are more or less common in males than in females, nor according to age group, nor educational status i.e. we do not know their distributions in combination with any of the other risk factors included in the model. Therefore we make the assumption that the distribution of all other risk factors (apart from afore-mentioned age, sex and educational status), is equal across all other risk factors. This makes the calculations somewhat easier, even though this assumption might make for slightly less accurate estimates, the loss of accuracy is not thought to be great.

In order to find the estimated prevalence for each population, it is necessary to calculate the synthetic prevalence of risk factors for each possible combination of risk factor (as included in the chosen disease-specific logistic regression model). The estimated prevalence for a population is then the weighted average of the prevalence estimates for each combination of risk factors, according to the estimated number of people with each risk factor combination in the population (the population on which synthetic estimates are sought). These calculations can be carried out in Excel (using VBA code to link prevalence and risk factor spreadsheets with formulae in a workbook) or in Stata software to produce confidence intervals as well as the estimates.

We have developed two methods for producing small population estimates and associated CIs in Stata software. One uses a bootstrapping method to produce repeated samples (Method 1), the other (Method 2) uses inverse probability weights. Both methods produce CIs for the estimates, which are derived from the variance in the logistic model, not the local populations. It would have been useful to compare the results of both methods, but because of the short timeframe for this project we only used Method 2: Logistic regression and inverse probability weights.

### Method 1: bootstrapping procedure to produce repeated samples

The detailed methods of the Stata code we developed and used is included in Annex 1: synthetic estimation using Stata. In summary, within Stata, a new set of variables is created, one for each combination of these risk factors pertinent to the logistic regression model for the chosen disease. With our dataset set up in this way, we can now use Stata’s “predict” command to give us the predicted log odds. Then we find the weighted average of these, averaged across all possible combinations of risk factors, using the weights calculated as above (stored in variable named xyz). The weighted average can be found using the “collapse” command as follows, which results in one line of data per practice or MLSOA (using the population identifier as the by variable) in Stata.

We calculated in Stata CIs for prevalence estimates using a “bootstrap” procedure. There is uncertainty in these synthetic estimates of prevalence based on the imprecision not in the more usual sample of people from the population (since the estimates are not a sample but are externally applied), but in the estimated coefficients from the logistic regression equations. A bootstrap procedure can be used to construct confidence intervals on these synthetic estimates of prevalence, based on the imprecision in these logistic regression coefficients.

The philosophy underlying the bootstrap procedure is to consider that the people included in the data set used to derive the logistic regression equation represent the whole population of possible people. However, the whole population is effectively considered to contain thousands of copies of each of these people. Bootstrap samples are taken randomly from our initial populations (the subsets of the CPRD population that has complete data on appropriate risk factors). Logistic regression of the same risk factors can then be applied to this boot strap sample, i.e. we rerun the logistic regression that gave us our chosen predictive model. However, we get slightly different regression coefficients, because of the modified sample. Prevalence estimates are then derived for each combination of risk factors, based on these new regression equations.

This process is repeated 1,000 times, to find 1,000 different boot strap samples, by random sampling processes, and to then fit logistic regression equations on each. The prevalence estimates are calculated for each combination of risk factors, for each of these 1,000 boot strap samples. For each small population, a synthetic estimate is calculated for each boot strap sample, by appropriately weighting the prevalence estimates on each combination of risk factors (with the same weights as described above which reflect the anticipated prevalence of each combination of risk factors in the population). From these 1,000 synthetic estimates of prevalence of each population, a 95% confidence interval is calculated as the 2.5th to 97.5th centiles. Given that the estimates are distributed normally, these are taken to be mean +/- 1.96 SD (taking mean and SD of the 1,000 boot strap synthetic prevalence estimates for each specified region).

### Method 2: Logistic regression and inverse probability weights

Inverse probability weighting methods are used to standardise from a sampled population to a target population. They are usually defined as a function of a panel of one or more sampling-probability predictor variables. For each combination of the predictor variables, the sampling probability weight is the ratio of the frequency of that combination in the target population to the frequency of that combination in the sampled population. Inverse probability weighting is therefore a generalization of direct standardization. In Stata, it is implemented by using a *pweight* qualifier on an estimation command. This normally implies the use of a Huber variance formula to generate the confidence limits.

In a population case-control study, our sampled population is an exhaustive list of disease cases, plus a random sample of controls without the disease, with a known sampling fraction. The sampling probability weights are inversely proportional to the sampling fraction for each sub-population. For cases, the sampling probability weight is 1. And, for controls, the sampling probability weight is the reciprocal of the sampling fraction. (So, if the sampling fraction is 1/8, then controls are weighted upwards by a factor of 8.) These sampling-probability weights are used in logistic regression models. Predicted disease probabilities from these models will then be unbiased, if the model is correctly specified.

Similarly to Method 1 we estimated population parameters for logistic regression models. The risk factors in the model fell into two classes, namely always-present risk factors and sometimes-missing risk factors. The always-present risk factors were gender (Male or Female), age group (18-44, 45-64, 65-74 and 75+), ethnicity (White, Mixed, Black, Asian or Other, imputed to White if not known). The sometimes-missing risk factors were practice index of multiple deprivation (IMD) quintile (1, 2, 3, 4 or 5), smoking status (Non-smoker, Ex-smoker or Smoker), alcohol units per week category (None, (0,14], (14,42] or >42), and body mass index in kilos/square metre (BMI) category ((0, 18.5], (18.5,25], (25,30] or >30).

We fitted the logistic regression model, using Huber variances and sampling-probability weights. The parameters were a baseline odds for each of the 2x4=8 combinations of gender and age group, an odds ratio for each ethnicity except White, an odds ratio for each IMD quintile except the first, an odds ratio for each smoking status except Non-smoker, an odds ratio for each alcohol consumption category except Zero units, and an odds ratio for each BMI category except (18.5,25] kilos per square metre. The sampling-probability weights used were equal to the products of two sets of component sampling-probability weights. The first set of component weights standardised by case status from the case-control study sample to the denominator population from which the cases and controls were sampled, and were equal to 1 for RA cases (assumed to be sampled exhaustively from the cases in the CPRD denominator population), and equal in the controls to the reciprocal of the sampling fraction of the controls as a fraction of the non-cases in the CPRD denominator population (equal to 27.211693).

We also use inverse probability weights to correct for missing values as an easy-to-use alternative to multiple imputation. We then define the inverse probability weights using a completeness-propensity score. We have a panel of variables *V1…VK* that are always present (such as age and gender), and a panel of variables *U1…UJ* that are sometimes missing. Let C (for completeness) be the binary indicator variable indicating that all the variables *U1…UJ* are present. We then use a logistic regression model, regressing C with respect to the always-complete variables *V1…VK*. The completeness-propensity score is defined as the predicted completeness probability for each individual, under that regression model. The inverse-probability weight, for each individual with a complete set of data *U1…UJ*, is then the reciprocal of that individual’s completeness-propensity score. Therefore, individuals with a high probability of having complete data (like elderly females) are weighted downwards. And individuals with a low probability of completeness (like young males) are weighted upwards. These inverse-probability weights can then be used in further regression models, such as a logistic regression model to predict disease.

Therefore, the second set of component weights were computed to standardise the sample of cases and controls with all risk factors present to the total sample of cases and controls by gender, age group and ethnicity, and were derived as inverse probabilities of presence of the full set of risk factors (completeness) from a logistic regression model with completeness as the outcome, fitted to the cases and controls, using the first set of sampling-probability weights to standardise by case status, and whose parameters were a baseline odds for each of the 8 combinations of gender and age group and an odds ratio for each non-white ethnic category. The product weights therefore were computed to standardise the odds and odds ratios from the sample of cases and controls with all risk factors present (272,369 subjects out of a total of 101,870 cases and 440,293 sampled controls) to the total denominator population of subjects aged at or above 18 years, with or without RA, on their birthdays in 2015 (13,864,783 subjects). We also fitted logistic regression models of RA status with respect to the 8 combinations of gender and age only, using only the first set of sampling probability weights to standardise by RA status, in order to estimate odds (and thereby prevalence) of RA for each combination of gender and age group in the CPRD population at large.

Having estimated the regression model parameters, we used these for out-of-sample prediction of RA prevalence, using the *margprev* add-on Stata package [[64](#_ENREF_64) [65](#_ENREF_65)]. These predicted prevalence estimates were for the sub-populations of patients for 7,692 practices, for 204 clinical care groups (CCGs), and for 6,755 MSOAs, for which information was available on the marginal frequencies of the seven risk factors in the model. We computed estimated prevalence assuming that, within each sub-population, the seven risk factors were mutually statistically independent, implying that we could give each possible combination of the seven risk factors a sampling-probability weight proportional to the product of the proportions of subjects with each of the appropriate risk-factor values. Therefore, for each subpopulation, we had 2x4x5x5x3x4x4=9600 combinations of risk factor values, with proportions of subjects calculated assuming statistical independence, and estimated the expected subpopulation prevalence of RA accordingly. The assumption of statistical independence of risk factors is probably not literally true, but might be expected to give prevalence estimates that are not vastly in error if the effects of the risk factors are not too non-additive. We have not internally or externally validated this method yet.

We have used method 2, logistic regression and inverse probability weights for these models because of the large number of variables in most of the models. This required us to produce Stata datasets of local risk factor data which have one observation for every permutation of all the risk factors for every practice, which generated very large files (up to 60 GB). We were able to process these using Stata/MP, the fastest and largest version of Stata. On dual-core chips, Stata/MP runs 40% faster overall and 72% faster on time-consuming estimation commands. It can handle a maximum number of 32,767 variables and 20 billion observations. Some of the datasets we used included over one billion observations. Processing was carried out on a multicore server. It would not have been possible to run the bootstrapping procedure to produce repeated samples which requires fitting a logistic model 1,000 times for each practice.

## Validation of local estimates

### Internal validation

In addition to the internal and external validation of the regression models, the local estimates can also be validated by aggregating them to the lowest geography available in the raw data and comparing them, a form of internal validation. This can be done in, for example HSfE which has a Regional variable. However Whitehall II does not- nearly all the participants were from London, so this was not possible. As noted above, we have over time increased the number of variables used in the local models as more local data has become available. However as more variables are added we need to take account of the joint effects of multiple risk factors, i.e. it assumes they operate independently. Estimation of the joint effects of multiple risk factors is complex for several reasons. In particular, some of the effects of more distal risk factors are mediated through intermediate factors. We have acknowledged this by creating specific joint distributions for variables where this is known e.g. age and educational level, as older age groups are less likely to have tertiary education.

### External validation

Because of the short timeframe for this project we have not had time to externally validate the local estimates using other similar datasets. However there are Quality & Outcomes Framework (QOF) disease registers[[66](#_ENREF_66)] for all the models we produced here. We have experience in comparing QOF-registered prevalence and estimated prevalence right down to practice level using spatial analyses.[[67](#_ENREF_67)] The local estimates can also be validated against the corresponding QOF register for each geography using Bland-Altman plots. This method uses graphical methods to investigate the assumptions of the method and also gives confidence intervals.[[68](#_ENREF_68)] It aims to quantify the agreement between and clinical importance of two methods of clinical measurement using the differences between observations made using the two methods on the same subjects. The 95% limits of agreement, estimated by mean difference 1.96 standard deviation of the differences, provide an interval within which 95% of differences between measurements by the two methods are expected to lie. The second method is based on errors-in-variables regression in a classical (X,Y) plot and focuses on confidence intervals, whereby two methods are considered equivalent when providing similar measures notwithstanding the random measurement errors.[[69](#_ENREF_69)] A recent update reconciles these two methodologies and shows their similarities and differences using both real data and simulations.[[70](#_ENREF_70)]

# Results

## Stroke prevalence from Whitehall II Data

### Missing data

Note that there 67% missing data for deprivation in Whitehall II data using Index of Multiple Deprivation (IMD). This is surprising since this is normally attributed using postcode of informants address, which the Whitehall II team must have. Table 1Shows missing data for key risk factor variables.

Table 15: missing data for key risk factor variables

|  |  |
| --- | --- |
| Risk Factor | Missing Values (%) |
| Age Group | 0 (0.0%) |
| Sex | 0 (0.0%) |
| Ethnicity | 0 (0.0%) |
| Hypertension |  |
| Dyslipidaemia |  |
| Family History (CHD) | 33 (0.3%) |
| Family History (Stroke) | 35 (0.3%) |
| Smoking | 12 (0.1%) |
| Obesity | 5 (0.04%) |
| Diabetes | 0 (0.0%) |
| Chronic Kidney Disease | 4242 (41.2%) |
| Hypercoagulation | 353 (3.4%) |
| Inflammatory markers | 1841 (17.9%) |
| Physical Activity | 3181 (6.7%) |
| Socioeconomic status | 6910 (67.0%) |

### Baseline characteristics of Whitehall II Respondents

Table 16 shows the baseline characteristics of the stroke dataset of Whitehall II participants

Table 16: baseline characteristics of stroke dataset of Whitehall II participants

|  | Stroke cases | No Stroke | Total |
| --- | --- | --- | --- |
| Total Individuals (n) | 4.8% (n=495)  | 95.2% (n=9813) | 100.0% (n=10,308) |
| Age (mean) | 70.5 | 67.2 | 67.3 |
| Age Group |  |  |  |
| 55-59 years | 2.2% (n=22) | 97.8% (n=978) | 9.7% (n=1000) |
| 60-64 years | 2.7% (n=82) | 97.3% (n=2929) | 29.2% (n=3011) |
| 65-69 years | 3.5% (n=82) | 96.5% (n=2254) | 22.7% (n=2336) |
| 70-74 years | 6.4% (n=137) | 93.6% (n=2013) | 20.9% (n=2150) |
| 75-79 years |  9.5% (n=172) | 90.5% (n=1639) | 17.5% (n=1811) |
| Ethnicity |  |  |  |
| White | 4.7% (n=435) | 95.3% (n=8746) | (n=9181) |
| Non-White | 5.3% (n=60) | 94.7% (n=1067) | (n=1127) |
| Sex |  |  |  |
| Male | 5.3% (n=368) | 94.7% (n=6527) | (n=6895) |
| Female | 3.7% (n=127) | 96.3% (n=3286) | (n=3413) |
| Hypertension |  |  |  |
| Yes | 6.6% (n=391) | 93.4% (n=5494) | (n=5885) |
| No  | 2.4% (n=104) | 97.6% (n=4319) | (n=4423) |
| Dyslipidaemia |  |  |  |
| Yes | 5.4% (n=433) | 94.6% (n=7614) | (n=8047) |
| No | 2.7% (n=62) | 97.3% (n=2199) | (n=2261) |
| Family History |  |  |  |
| Yes | 6.5% (n=166) | 93.5% (n=2394) | (n=2560) |
| No | 4.2% (n=329) | 95.8% (n=7419) | (n=7748) |
| Diabetes |  |  |  |
| Yes | 8.3% (n=116) | 91.7% (n=1278) | (n=1394) |
| No | 4.3% (n=379) | 95.7% (n=8535) | (n=8914) |
| Obesity |  |  |  |
| BMI≥30 | 6.7% (n=141) | 93.3% (n=1956) | (n=2097) |
| BMI<30 | 4.3% (n=354) | 95.7% (n=7857) | (n=8211) |
| Smoking |  |  |  |
| Ever smoked | 5.0% (n=277) | 95.0% (n=5291) | (n=5568) |
| Never smoked | 4.6% (n=218) | 95.4% (n=4522) | (n=4740) |
|  |  |  |  |
| Chronic Kidney Disease |  |  |  |
| EGFR≤60 | 7.5% (n=51) | 92.5% (n=629) | (n=680) |
| EGFR>60 | 4.4% (n=239) |  95.6% (n=5147) | (n=5386) |
| Missing | 4.8% (n=205) | 95.2% (n=4037) | (n=4242) |
|  |  |  |  |
| Hypercoagulable state/on anticoagulants |
| Yes | 14.4% (n=322) | 85.6% (n=1919) | (n=2241) |
| No | 2.1% (n=173) | 97.9% (n=7894) | (n=8067) |
| Inflammatory markers |  |  |  |
| CRP≥3.0  | 6.8% (n=173) | 93.2% (n=2387) | (n=2560) |
| CRP<3.0 | 4.2% (n=322) | 95.8% (n=7426) | (n=7748) |
| Physical Activity |  |  |  |
| Low | 5.4% (n=121) | 94.6% (n=2128) | (n=2249) |
| Medium | 5.5% (n=65) | 94.5% (n=1114) | (n=1179) |
| High | 5.6% (n=208) | 94.4% (n=3491) | (n=3699) |
| Missing | 3.2% (n=101) | 96.8% (n=3080) | (n=3181) |
| Socioeconomic status |  |  |  |
| Townsend 1st quintile | 4.9% (n=96) | 95.1% (n=1867) | (n=1963) |
| Townsend 2nd quintile | 4.2% (n=82) | 95.8% (n=1878) | (n=1960) |
| Townsend 3rd quintile | 4.7% (n=92) | 95.3% (n=1863) | (n=1955) |
| Townsend 4th quintile | 5.2% (n=103) | 94.8% (n=1866) | (n=1969) |
| Townsend 5th quintile | 4.8% (n=94) | 95.2% (n=1855) | (n=1949) |
| Missing | 5.5% (n=28) | 94.5% (n=484) | (n=512) |

## Whitehall II CVD definitions, incidence & prevalence

### Whitehall II prevalence

Prevalence of CVD in the Whitehall II data were calculated for CHD, stroke, and PAD based on the diagnostic algorithms defined above. The prevalence of CHD, stroke, and PAD is shown in Table 17, broken down by age and sex.

Table 17: Prevalence of Stroke by age and sex from Whitehall II data

|  |  |  |
| --- | --- | --- |
|  | Females | Males |
| Age Group | **Stroke cases** | **Total number** | **Prevalence** | **Stroke cases** | **Total number** | **Prevalence** |
| 55-59 years | 6 | 284 | 2.1% | 16 | 716 | 2.2% |
| 60-64 years | 18 | 841 | 2.1% | 64 | 2170 | 2.9% |
| 65-69 years | 24 | 763 | 3.1% | 58 | 1573 | 3.7% |
| 70-74 years | 29 | 802 | 3.6% | 108 | 1348 | 8.0% |
| 75-79 years | 50 | 723 | 6.9% | 122 | 1088 | 11.2% |

## Stroke regression modelling

### Univariate logistic analysis

Table 18 shows the results for stroke of a univariate logistic model for individual risk factors and the outcome.

Table 18: univariate logistic model for individual risk factors for stroke

| Risk Factor | Odds Ratio (OR) | 95% CI | p-value |
| --- | --- | --- | --- |
| Age |  |  |  |
| 55-59 years | 1.00 | Reference |  |
| 60-64 years | 1.24 | [0.77-2.00] | 0.368 |
| 65-69 years | 1.62 | [1.00-2.60] | 0.048 |
| 70-74 years | 3.03 | [1.92-4.78] | <0.001 |
| 75-79 years | 4.67 | [2.97-7.32] | <0.001 |
| Sex |  |  |  |
| Male | 1.00 | Reference |  |
| Female | 0.69 | [0.56-0.84] | <0.001 |
| Ethnicity |  |  |  |
| White | 1.00 | Reference |  |
| Non-White | 1.13 | [0.86-1.49] | 0.386 |
| Hypertension |  |  |  |
| Normotensive | 1.00 | Reference |  |
| Pre-hypertensive | 1.05 | [0.85-1.29] | 0.639 |
| Hypertensive | 1.17 | [0.92-1.49] | 0.206 |
| Diabetes |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 2.04 | [1.65-2.54] | <0.001 |
| Smoking |  |  |  |
| Never | 1.00 | Reference |  |
| Former | 1.11 | [0.91-1.34] | 0.295 |
| Current | 1.00 | [0.74-1.34] | 0.974 |
| Obesity |  |  |  |
| BMI < 18.5 | 1.00 | Reference |  |
| BMI 18.5-25 | 2.62 | [0.64—10.71] | 0.178 |
| BMI 25-30 | 3.04 | [0.75-12.39] | 0.121 |
| BMI 30-35 | 4.19 | [1.02-17.25] | 0.047 |
| BMI >30  | 4.51 | [1.07-19.05] | 0.041 |
| Family History |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.56 | [1.29-1.89] | <0.000 |
| Physical Activity |  |  |  |
| Low | 1.00 | Reference |  |
| Medium | 0.87 | [0.67-1.14-2.39] | 0.321 |
| High | 0.69 | [0.56-0.85] | 0.001 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| HDL Cholesterol |  |  |  |
| HDL > 1.55 g/L | 1.00 | Reference |  |
| HDL 1.03-1.55 g/L | 1.34 | [1.10-1.63] | 0.004 |
| HDL < 1.03 g/L | 1.78 | [1.34-2.38] | <0.001 |
| LDL Cholesterol |  |  |  |
| LDL < 3.3 g/L | 1.03 | Reference |  |
| LDL 3.3-4.1 g/L  | 0.44 | [0.34-0.56] | <0.001 |
| LDL > 4.1 | 0.50 | [0.39-0.65] | <0.001 |
| Triglycerides |  |  |  |
| < 1.7 g/L | 1.00 | Reference |  |
| 1.7-5.65 g/L | 1.17 | [0.95-1.45] | 0.143 |
| >5.65 g/L | 0.46 | [0.07-3.14] | 0.425 |
| Chronic Kidney Disease |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.96 | [1.44-2.67] | <0.001 |
| IMD |  |  |  |
| 1st quintile | 1.00 | Reference |  |
| 2nd quintile |  |  |  |
| 3rd quintile |  |  |  |
| 4th quintile |  |  |  |
| 5th quintile |  |  |  |

### Multivariate logistic analysis for stroke

We went through an extensive model fitting process to compare the performance of different models that included CVD cases identified by different risk factors.

Table 19 below show the results of multivariate model fitting for CHD. Note that we have not included hypercoagulable state/on anticoagulants in this model, as the direction of causality may be wrong i.e. people may be on anticoagulants because they have CHD. This is suggested by the high OR of 2.83 [2.47-3.25], <0.001. Note that deprivation (Townsend score) is not significant in this model.

Table 19: full model including all variables logistic regression model for stroke

| Risk Factor | Odds Ratio (OR) | 95% CI | p-value |
| --- | --- | --- | --- |
| Age |  |  |  |
| 55-59 years | 1.00 | Reference |  |
| 60-64 years | 1.60 | [0.77-3.32] | 0.204 |
| 65-69 years | 1.63 | [0.78-3.41] | 0.194 |
| 70-74 years | 2.17 | [1.04-4.50] | 0.037 |
| 75-79 years | 3.17 | [1.53-6.59] | 0.002 |
| Sex |  |  |  |
| Male | 1.00 | Reference |  |
| Female | 0.59 | [0.42-0.85] | 0.004 |
| Ethnicity |  |  |  |
| White | 1.00 | Reference |  |
| Non-White | 1.04 | [0.64-1.67] | 0.880 |
| Hypertension |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.59 | [1.12-2.26] | 0.009 |
| Diabetes |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 0.94 | [0.67-1.31] | 0.703 |
| Dyslipidaemia |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 0.93 | [0.59-1.47] | 0.756 |
| Obesity |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.00 | [0.73-1.38] | 0.978 |
| Smoker |  |  |  |
| Never | 1.00 | Reference |  |
| Ever | 0.93 | [0.71-1.21] | 0.581 |
| Chronic Kidney Disease |  |  |  |
| EGFR>60 | 1.00 | Reference |  |
| EGFR≤60 | 1.18 | [0.82-1.71] | 0.371 |
| Inflammatory Markers |  |  |  |
| CRP<3.0 | 1.00 | Reference |  |
| CRP≥3.00 | 1.23 | [0.92-1.63] | 0.155 |
| Physical Activity |  |  |  |
| Low | 1.00 | Reference |  |
| Medium | 0.99 | [0.65-1.52] | 0.971 |
| High | 0.92 | [0.66-1.28] | 0.612 |
| Family History |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.16 | [0.87-1.54] | 0.319 |
| Socioeconomic Status |  |  |  |
| Townsend 1st quintile | 1.00 | Reference |  |
| Townsend 2nd quintile | 0.88 | [0.59-1.33] | 0.542 |
| Townsend 3rd quintile | 1.14 | [0.77-1.69] | 0.510 |
| Townsend 4th quintile | 1.18 | [0.77-1.80] | 0.441 |
| Townsend 5th quintile | 1.18 | [0.75-1.85] | 0.473 |

We then carried out stepwise forward and backward variable selection using likelihood ratios (LRs) and Wald tests because missing data had been imputed for some variables. Table 20 shows the final logistic regression model for stroke using stepwise forward and backward selection.

Table 20: logistic regression model for stroke (stepwise forward and backward selection)

| Risk Factor | Odds Ratio (OR) | 95% CI | p-value |
| --- | --- | --- | --- |
| Age |  |  |  |
| 70-74 years | 1.86 | [1.17-2.95] | 0.008 |
| 75-79 years | 1.73 | [1.03-2.89] | 0.037 |
| Ethnicity |  |  |  |
| White | 1.00 | Reference |  |
| Non-White | 2.19 | [1.27-3.75] | 0.005 |
| Diabetes |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.85 | [1.21-2.85] | 0.005 |
| Smoker |  |  |  |
| Never | 1.00 | Reference |  |
| Ever | 1.68 | [1.12-2.52] | 0.013 |
| Chronic Kidney Disease |  |  |  |
| EGFR>60 | 1.00 | Reference |  |
| EGFR≤60 | 1.97 | [1.23-3.17] | 0.005 |
| Hypercoagulable State |  |  |  |
| No | 1.00 | Reference |  |
| Yes | 1.84 | [1.22-2.76] | 0.003 |
| Inflammatory Markers |  |  |  |
| CRP<3.0 | 1.00 | Reference |  |
| CRP≥3.00 | 1.78 | [1.20-2.64] | 0.004 |

As noted in the Methods, this final model above did not include a deprivation variable as it dropped out in the stepwise variable selection. However because deprivation is included in UK CVD predictive risk scores we produced a second set of local estimates including deprivation in order to check what effect this had on the estimates, even though it is not significant, so there are two sets of all local estimates, and these files are labelled with or without a deprivation variable. We used IMD in the final model (see Table 21 below) as it is the most commonly used deprivation score at the local level. Note that regression co-efficients are shown here rather than ORS- a positive coefficient means an OR>1 and vice versa.

Table 21: logistic regression model for CHD with deprivation (IMD) score variable (stepwise forward and backward selection except for IMD)

| Parameter | Odds ratio | Lower 95% CI | Upper 95% CI | p value |
| --- | --- | --- | --- | --- |
| Age Group |  |  |  |  |
| 55-59 years | 1 | 1 | 1 | . |
| 60-64 years | 1.249 | 0.774 | 2.013 | 0.362 |
| 65-69 years | 1.659 | 1.027 | 2.678 | 0.038 |
| 70-74 years | 3.066 | 1.934 | 4.861 | 0.000 |
| 75-79 years | 4.713 | 2.983 | 7.446 | 0.000 |
|  |  |  |  |  |
| Male | 1 | 1 | 1 | . |
| Female | 0.548 | 0.441 | 0.680 | 0.000 |
|  |  |  |  |  |
| White | 1 | 1 | 1 | . |
| Non-White | 0.925 | 0.692 | 1.237 | 0.599 |
| Diabetes |  |  |  |  |
| No | 1 | 1 | 1 | . |
| Yes | 1.675 | 1.337 | 2.099 | 0.000 |
| BMI |  |  |  |  |
| BMI <18.5 | 1 | 1 | 1 | . |
| BMI 18.5-25  | 2.342 | 0.571 | 9.604 | 0.237 |
| BMI 25-30 | 2.437 | 0.594 | 9.990 | 0.216 |
| BMI 30-35 | 3.491 | 0.847 | 14.382 | 0.084 |
| Physical Activity |  |  |  |  |
| Low | 1 | 1 | 1 | . |
| Medium | 0.844 | 0.649 | 1.099 | 0.208 |
| High | 0.671 | 0.544 | 0.829 | 0.000 |
| Physical Activity |  |  |  |  |
| Low | 1 | 1 | 1 | . |
| Medium | 1.240 | 0.977 | 1.572 | 0.077 |
| Deprivation |  |  |  |  |
| IMD quintile 1 (least deprived) | 1 | 1 | 1 | . |
| IMD quintile 2 | 0.997 | 0.749 | 1.329 | 0.986 |
| IMD quintile 3 | 1.137 | 0.860 | 1.504 | 0.368 |
| IMD quintile 4 | 0.906 | 0.673 | 1.219 | 0.514 |
| IMD quintile 5 (most deprived) | 1.106 | 0.821 | 1.489 | 0.507 |
| constant | 0.011 | 0.003 | 0.049 | 0.000 |

## Internal validation

### ROC curves

We next examined the receiver operating characteristics (ROC) curves for the various models. The best ROC curve which predicts data perfectly will touch the top-left corner of the plot (area 1.0), and the larger the area under the ROC curve the better the prediction. An area of 0.5 signifies a prediction no better than chance. Unfortunately these models do not perform particulary well- we would hope for an area under the ROC curve of at least 70%. The previous models based on HSfE data achieve areas under the ROC curve of >80%. The results are summarised in Table 22.

Table 22: receiver operating characteristics (ROC) curves for the various CPRD models

| Model description | Model | ROC area | SE | 95% CI |
| --- | --- | --- | --- | --- |
| Logistic regression model for Stroke including all risk factor variables (Full Model) | M2 | 0.6719 | 0.0084 | 0.6555-0.6883 |
| Logistic regression model for Stroke – Forward and backward stepwise selection | M5 | 0.6648 | 0.0084 | 0.6483-0.6812 |

### Probability and sensitivity/specificity analysis

We used the stepwise forward models i.e. M2 and M5 to predict the probability of individual being stroke case in the Whitehall II dataset. Since we have a binary response model, we chose a cut-off point of the predicted probability to separate the predicted stroke cases (with higher predicted probability) from the predicted non- stroke cases (with lower predicted probability). We can tell from the box plots no matter which cut-off point we choose, there will always be mis-classified people.

The sensitivity/specificity versus probability cut-off plot shows us the corresponding sensitivity and specificity in each possible probability cut-off point. Higher sensitivity would usually yield low specificity and vice versa, the rule of thumb is to choose a cut-off probability to maximize both. We choose the cut-off probability where sensitivity and specificity lines cross.

## Local estimates

### Internal validation

We have not carried out a validation by aggregating local estimates because, as noted in the Methods, Whitehall II was not a nationally representative sample.

### External validation of practice estimates against QOF prevalence

The funding for the project does not include an in-depth external validation. For example, this could be carried out by obtaining an extract from a similar dataset e.g. applying the HSfE prevalence models’ equations to Whitehall II data. However another useful external data source is the Quality & Outcomes Framework (QOF) GP-diagnosed stroke prevalence. This can obviously be compared with diagnosed stroke prevalence from the model, taking into account that the Whitehall II definition was derived from the number of patients that reported being told by a nurse or doctor that they had stroke.

We carried out a disagreement analysis between model-estimated and QOF prevalence (%) of diagnosed stroke in practices. We estimated three principal components of disagreement (discordance as measured by Kendall's tau-a, bias as measured by median difference, and calibration as measured by the Theil-Sen median slope). Using the stroke estimates without IMD score in the model, the Kendall's tau-a between model-estimated and QOF prevalence of stroke for 7,500 practices was 0.470 (95% CIs 0.458-0.482)), and p=0.000. Table 23 shows percentile differences between model-estimated and QOF prevalence of diagnosed stroke.

Table 23: percentile differences between model-estimated and QOF prevalence of stroke

| Percent | Percentile | (95% | CI) |
| --- | --- | --- | --- |
| 0 | -14.1 |  | -14.1) |
| 25 | 1.7 | (1.7 | 1.7) |
| 50 | 2.0 | (2.0 | 2.0) |
| 75 | 2.3 | (2.3 | 2.3) |
| 100 | 3.7 | (3.7, | 3.7) |

The best way to display the data is to plot the difference between the measurements by the two methods for each subject against their mean. This plot for the local stroke prevalence (Figure 1) shows explicitly the lack of agreement. The average difference between the estimates is close to 2% higher for estimated prevalence. This is plausible if some patients with stroke have a diagnosis made, but then change GP, or the current GP has told the patient they have stroke but does not record it on the stroke register (they may even be treated with secondary prevention drugs). The percentile slope of model-estimated prevalence with respect to QOF prevalence of diagnosed stroke was 1.043 (95% CI 0.982-1.103). Figure 1 shows a Bland-Altman plot for model-estimated and QOF prevalence of diagnosed stroke, and Figure 2 is a scatter plot of model-estimated and QOF prevalence of diagnosed stroke.

Figure 1: Bland-Altman plot for model-estimated and QOF prevalence of stroke



Figure 2: scatter plot of model-estimated and QOF prevalence of stroke



# References

1. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet 2015(0). URL: <http://www.sciencedirect.com/science/article/pii/S0140673615606924>.

2. Bhatnagar P, Wickramasinghe K, Williams J, et al. The epidemiology of cardiovascular disease in the UK 2014. Heart 2015. URL: <http://heart.bmj.com/content/early/2015/05/06/heartjnl-2015-307516.abstract>.

3. Finegold JA, Asaria P, Francis DP. Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. International journal of cardiology 2013;**168**(2):934-45. URL.

4. Fernández-Friera L, Peñalvo JL, Fernández-Ortiz A, et al. Prevalence, Vascular Distribution and Multi-territorial Extent of Subclinical Atherosclerosis in a Middle-Aged Cohort: The PESA (Progression of Early Subclinical Atherosclerosis) Study. Circulation 2015. URL: <http://circ.ahajournals.org/content/early/2015/04/16/CIRCULATIONAHA.114.014310.abstract>.

5. Bos MJ, Koudstaal PJ, Hofman A, et al. Modifiable etiological factors and the burden of stroke from the Rotterdam study: a population-based cohort study. PLoS Med 2014;**11**(4):e1001634. URL.

6. Eraso LH, Fukaya E, Mohler ER, 3rd, et al. Peripheral arterial disease, prevalence and cumulative risk factor profile analysis. European journal of preventive cardiology 2014;**21**(6):704-11. URL.

7. Fowkes FG, Rudan D, Rudan I, et al. Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: a systematic review and analysis. Lancet (London, England) 2013;**382**(9901):1329-40. URL.

8. O'Donnell MJ, Xavier D, Liu L, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. The Lancet 2010;**376**(9735):112-23. URL: <http://www.sciencedirect.com/science/article/B6T1B-50B8P65-2/2/efe3331b2febdeb0571b6e791df605a3>.

9. Dégano IR, Salomaa V, Veronesi G, et al. Twenty-five-year trends in myocardial infarction attack and mortality rates, and case-fatality, in six European populations. Heart 2015. URL: <http://heart.bmj.com/content/early/2015/04/08/heartjnl-2014-307310.abstract>.

10. Rahimi K, Duncan M, Pitcher A, et al. Mortality from heart failure, acute myocardial infarction and other ischaemic heart disease in England and Oxford: a trend study of multiple-cause-coded death certification. Journal of Epidemiology and Community Health 2015;**69**(10):1000-05. URL: <http://jech.bmj.com/content/69/10/1000.abstract>.

11. Howard DPJ, Banerjee A, Fairhead JF, et al. Population-Based Study of Incidence, Risk Factors, Outcome, and Prognosis of Ischemic Peripheral Arterial Events: Implications for Prevention. Circulation 2015;**132**(19):1805-15. URL: <http://circ.ahajournals.org/content/132/19/1805.abstract>.

12. Peters SA, Huxley RR, Woodward M. Diabetes as a risk factor for stroke in women compared with men: a systematic review and meta-analysis of 64 cohorts, including 775,385 individuals and 12,539 strokes. Lancet (London, England) 2014;**383**(9933):1973-80. URL.

13. Peters SA, Huxley RR, Woodward M. Smoking as a risk factor for stroke in women compared with men: a systematic review and meta-analysis of 81 cohorts, including 3,980,359 individuals and 42,401 strokes. Stroke 2013;**44**(10):2821-8. URL.

14. Mosca L, Barrett-Connor E, Wenger NK. Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes. Circulation 2011;**124**(19):2145-54. URL.

15. Yeh RW, Sidney S, Chandra M, et al. Population Trends in the Incidence and Outcomes of Acute Myocardial Infarction. The New England Journal of Medicine 2010;**362**(23):2155-65. URL: <http://content.nejm.org/cgi/content/abstract/362/23/2155>

<http://content.nejm.org/cgi/content/full/362/23/2155>.

16. McFadden E, Luben R, Wareham N, et al. Social Class, Risk Factors, and Stroke Incidence in Men and Women. A Prospective Study in the European Prospective Investigation Into Cancer in Norfolk Cohort. Stroke 2009;**40**(4):1070-77. URL: <http://stroke.ahajournals.org/cgi/content/abstract/STROKEAHA.108.533414v1>

<http://stroke.ahajournals.org/cgi/content/full/40/4/1070?cookietest=yes>.

17. Appelros P, Stegmayr B, Terent A. Sex Differences in Stroke Epidemiology. A Systematic Review. Stroke 2009:STROKEAHA. URL: <http://stroke.ahajournals.org/cgi/content/abstract/STROKEAHA.108.540781v1>

file://I:\NSFs\Older People\Stroke\References\Appelros P Sex Differences in Stroke Epidemiology. A Systematic Review Stroke 2009 108.540781v1.pdf

<http://stroke.ahajournals.org/cgi/content/abstract/STROKEAHA.108.540781v1?papetoc>.

18. Asplund K, Karvanen J, Giampaoli S, et al. Relative Risks for Stroke by Age, Sex, and Population Based on Follow-Up of 18 European Populations in the MORGAM Project. Stroke 2009;**40**(7):2319-26. URL: <http://stroke.ahajournals.org/cgi/content/abstract/40/7/2319>.

19. Hemingway H, Langenberg C, Damant J, et al. Prevalence of Angina in Women Versus Men: A Systematic Review and Meta-Analysis of International Variations Across 31 Countries. Circulation 2008;**117**(12):1526-36. URL: <http://circ.ahajournals.org/cgi/content/abstract/117/12/1526>.

20. De Bacquer D, De Backer G, Kornitzer M. Prevalences of ECG findings in large population based samples of men and women. Heart 2000;**84**(6):625-33. URL: <http://heart.bmj.com/cgi/content/abstract/84/6/625>.

21. Murphy NF, Simpson CR, MacIntyre K, et al. Prevalence, incidence, primary care burden and medical treatment of angina in Scotland: age, sex and socioeconomic disparities: a population-based study. Heart 2006;**92**(8):1047-54. URL: <http://heart.bmj.com/cgi/content/abstract/92/8/1047>

file://I:\Modelling\Prevalence Modelling\CHD Prevalence Model\References\Murphy N Prevalence , incidence, & primary care burden of angina in Scotland Heart 2006 92 1047.pdf

<http://heart.bmj.com/cgi/citmgr?gca=heartjnl> ;92/8/1047.

22. Chaturvedi N. Ethnic differences in cardiovascular disease. Heart 2003;**89**(6):681-86. URL: <http://heart.bmj.com/cgi/content/full/89/6/681>.

23. Sproston K, Mindell J. Health Survey for England 2004: Health of Ethnic Minorities - Full Report: National Centre for Social Research, Information Centre for health & social care, 2006.

24. Majeed FA, Chaturvedi N, Reading R. Equity in the NHS. Monitoring and promoting equity and secondary care. BMJ 1994;**308**:1426-29. URL.

25. Huang Y, Cai X, Liu C, et al. Prehypertension and the risk of coronary heart disease in Asian and Western populations: a meta-analysis. Journal of the American Heart Association 2015;**4**(2). URL.

26. Heuschmann PU, Grieve AP, Toschke AM, et al. Ethnic Group Disparities in 10-Year Trends in Stroke Incidence and Vascular Risk Factors: The South London Stroke Register (SLSR). Stroke 2008;**39**(8):2204-10. URL: <http://stroke.ahajournals.org/cgi/content/abstract/39/8/2204>

file://I:\NSFs\Older People\Stroke\References\Heuschmann P Ethnic Group Disparities in 10-Year Trends in Stroke Incidence and Vascular Risk Factors Stroke 2008 39 2204.pdf

<http://stroke.ahajournals.org/cgi/content/full/39/8/2204?cookietest=yes>.

27. Smeeton NC, Heuschmann PU, Rudd AG, et al. Incidence of Hemorrhagic Stroke in Black Caribbean, Black African, and White Populations: The South London Stroke Register, 1995 2004. Stroke 2007;**38**(12):3133-38. URL: <http://stroke.ahajournals.org/cgi/content/abstract/38/12/3133>

file://I:\NSFs\Older People\Stroke\References\Smeeton N Incidence of Hemorrhagic Stroke in Black Caribbean , Black African, and White Populations Stroke 2007 38 3133.pdf

<http://stroke.ahajournals.org/cgi/content/full/38/12/3133>.

28. Hajat C, Tilling K, Stewart JA, et al. Ethnic Differences in Risk Factors for Ischemic Stroke: A European Case-Control Study. Stroke 2004;**35**(7):1562-67. URL: <http://stroke.ahajournals.org/cgi/content/abstract/35/7/1562>

file://P:\NSFs\Older People\Stroke\References\Hajat C Ethnic Differences in Risk Factors for Ischemic Stroke Stroke 2004 35 1562.pdf

<http://stroke.ahajournals.org/cgi/content/full/35/7/1562>.

29. Willey JZ, Moon YP, Kahn E, et al. Population attributable risks of hypertension and diabetes for cardiovascular disease and stroke in the northern Manhattan study. Journal of the American Heart Association 2014;**3**(5):e001106. URL.

30. Schnohr P, Marott JL, Kristensen TS, et al. Ranking of psychosocial and traditional risk factors by importance for coronary heart disease: the Copenhagen City Heart Study. Eur Heart J 2015;**36**(22):1385-93. URL.

31. Tolstrup JS, Hvidtfeldt UA, Flachs EM, et al. Smoking and risk of coronary heart disease in younger, middle-aged, and older adults. American journal of public health 2014;**104**(1):96-102. URL.

32. Shields M, Wilkins K. Smoking, smoking cessation and heart disease risk: A 16-year follow-up study. Health reports 2013;**24**(2):12-22. URL.

33. Rautio A, Eliasson M, Stegmayr B. Favorable Trends in the Incidence and Outcome in Stroke in Nondiabetic and Diabetic Subjects: Findings From the Northern Sweden MONICA Stroke Registry in 1985 to 2003. Stroke 2008;**39**(12):3137-44. URL: <http://stroke.ahajournals.org/cgi/content/abstract/39/12/3137>

file://I:\NSFs\Older People\Stroke\References\Rautio A Favorable Trends in the Incidence and Outcome in Stroke in Nondiabetic and Diabetic Subjects 2008 39 3137.pdf

<http://stroke.ahajournals.org/cgi/content/full/39/12/3137>.

34. Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet (London, England) 2004;**364**(9438):937-52. URL.

35. Selvin E, Erlinger TP. Prevalence of and risk factors for peripheral arterial disease in the United States: results from the National Health and Nutrition Examination Survey, 1999-2000. Circulation 2004;**110**(6):738-43. URL.

36. Zhao W, Katzmarzyk PT, Horswell R, et al. HbA1c and coronary heart disease risk among diabetic patients. Diabetes Care 2014;**37**(2):428-35. URL.

37. Mulnier HE, Seaman HE, Raleigh VS, et al. Risk of stroke in people with type 2 diabetes in the UK: a study using the General Practice Research Database. Diabetologia 2006;**49**(12):2859-65. URL.

38. Hinchliffe RJ, Andros G, Apelqvist J, et al. A systematic review of the effectiveness of revascularization of the ulcerated foot in patients with diabetes and peripheral arterial disease. Diabetes/metabolism research and reviews 2012;**28 Suppl 1**:179-217. URL.

39. Ezzati M, Vander Hoorn S, Rodgers A, et al. Estimates of global and regional potential health gains from reducing muliple major risk factors. The Lancet 2003;**362**(9380):271-80. URL: <http://www.sciencedirect.com/science/article/pii/S0140673603139682>.

40. Van Gaal LF, Mertens IL, De Block CE. Mechanisms linking obesity with cardiovascular disease. Nature 2006;**444**(7121):875-80. URL.

41. Winter Y, Rohrmann S, Linseisen J, et al. Contribution of Obesity and Abdominal Fat Mass to Risk of Stroke and Transient Ischemic Attacks. Stroke 2008;**39**(12):3145-51. URL: <http://stroke.ahajournals.org/cgi/content/abstract/39/12/3145>

file://I:\NSFs\Older People\Stroke\References\Winter Y Contribution of Obesity and Abdominal Fat Mass to Risk of Stroke & TIA Stroke 2008 39 3145.pdf

<http://stroke.ahajournals.org/cgi/content/full/39/12/3145>.

42. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008;**336**(7659):1475-82. URL: <http://www.bmj.com/cgi/content/abstract/bmj.39609.449676.25v1>

file://I:\NSFs\CHD\Prevention in High Risk Groups\CVD Risk\References\Hippisley-Cox J Predicting CVD risk in England & Wales- prospective derivation and validation of QRISK2 BMJ 2008 336 1475.pdf

<http://www.bmj.com/cgi/content/full/bmj.39609.449676.25v1>.

43. Anderson KM, Odell PM, Wilson PW, et al. Cardiovascular disease risk profiles. American heart journal 1991;**121**(1):293-98. URL.

44. Joint British Societies. JBS2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. Heart 2005;**91(suppl 5)**:v1-v52. URL: <http://www.library.nhs.uk/GuidelinesFinder/ViewResource.aspx?resID=112476>.

45. Hippisley-Cox J, Coupland C, Robson J, et al. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ 2010;**341**. URL: <http://www.bmj.com/bmj/341/bmj.c6624.full.pdf>.

46. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008;**336**. URL: <http://www.bmj.com/content/336/7659/1475>.

47. Scholes S, Bajekal M, Love H, et al. Persistent socioeconomic inequalities in cardiovascular risk factors in England over 1994-2008: A time-trend analysis of repeated cross-sectional data. BMC Public Health 2012;**12**(1):129. URL: <http://www.biomedcentral.com/1471-2458/12/129>.

48. Bajekal M, Scholes S, Love H, et al. Analysing Recent Socioeconomic Trends in Coronary Heart Disease Mortality in England, 2000–2007: A Population Modelling Study. PLoS Med 2012;**9**(6):e1001237. URL: [http://dx.doi.org/10.1371%2Fjournal.pmed.1001237](http://dx.doi.org/10.1371/journal.pmed.1001237).

49. Arem H, Moore SC, Patel A, et al. Leisure time physical activity and mortality: A detailed pooled analysis of the dose-response relationship. JAMA Internal Medicine 2015;**175**(6):959-67. URL: <http://dx.doi.org/10.1001/jamainternmed.2015.0533>.

50. Wilmot EG, Edwardson CL, Achana FA, et al. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. Diabetologia 2012;**55**(11):2895-905. URL: <http://dx.doi.org/10.1007/s00125-012-2677-z>.

51. Wen CP, Wai JPM, Tsai MK, et al. Minimum amount of physical activity for reduced mortality and extended life expectancy: a prospective cohort study. The Lancet;**378**(9798):1244-53. URL: <http://www.sciencedirect.com/science/article/pii/S0140673611607496>.

52. The Emerging Risk Factors Collaboration. LIpid-related markers and cardiovascular disease prediction. JAMA: The Journal of the American Medical Association 2012;**307**(23):2499-506. URL: <http://dx.doi.org/10.1001/jama.2012.6571>.

53. Gansevoort RT, Correa-Rotter R, Hemmelgarn BR, et al. Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention. Lancet (London, England) 2013;**382**(9889):339-52. URL.

54. Tonelli M, Muntner P, Lloyd A, et al. Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. Lancet (London, England) 2012;**380**(9844):807-14. URL.

55. Toyoda K, Ninomiya T. Stroke and cerebrovascular diseases in patients with chronic kidney disease. The Lancet. Neurology 2014;**13**(8):823-33. URL.

56. Lee M, Saver JL, Chang KH, et al. Low glomerular filtration rate and risk of stroke: meta-analysis. Bmj 2010;**341**:c4249. URL.

57. Kaptoge S, Di Angelantonio E, Pennells L, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. The New England journal of medicine 2012;**367**(14):1310-20. URL.

58. De Caterina R, Scarano M, Marfisi R, et al. Cholesterol-Lowering Interventions and Stroke: Insights From a Meta-Analysis of Randomized Controlled Trials. Journal of the American College of Cardiology 2010;**55**(3):198-211. URL: <http://www.sciencedirect.com/science/article/B6T18-4Y52NY9-5/2/04685d03e19ab862bde6fe8be867cbea>

<http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T18-4Y52NY9-5&_user=217827&_coverDate=01%2F19%2F2010&_alid=1265614074&_rdoc=18&_fmt=high&_orig=search&_cdi=4884&_sort=r&_docanchor=&view=c&_ct=23&_acct=C000011279&_version=1&_urlVersion=0&_userid=217827&md5=094b20fd9c3ce10d52fbb6fb0c0c882e>.

59. Bjorck S, Palaszewski B, Friberg L, et al. Atrial fibrillation, stroke risk, and warfarin therapy revisited: a population-based study. Stroke 2013;**44**(11):3103-8. URL.

60. Gage BF, Waterman AD, Shannon W, et al. Validation of Clinical Classification Schemes for Predicting Stroke: Results From the National Registry of Atrial Fibrillation. JAMA: The Journal of the American Medical Association 2001;**285**(22):2864-70. URL: <http://jama.ama-assn.org/cgi/content/abstract/285/22/2864>

<http://jama.ama-assn.org/cgi/content/full/285/22/2864?maxtoshow=&hits=10&RESULTFORMAT=1&author1=gage&andorexacttitle=and&andorexacttitleabs=and&andorexactfulltext=and&searchid=1&FIRSTINDEX=0&sortspec=relevance&resourcetype=HWCIT>.

61. Newton JN, Briggs ADM, Murray CJL, et al. Changes in health in England, with analysis by English regions and areas of deprivation, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet 2015. URL: <http://www.sciencedirect.com/science/article/pii/S0140673615001956>.

62. Nacul L, Soljak M, Meade T. Model for estimating the population prevalence of chronic obstructive pulmonary disease: cross sectional data from the Health Survey for England. Population Health Metrics 2007;**5**(1):8. URL: <http://www.pophealthmetrics.com/content/5/1/8>.

63. Kirkwood BR, Sterne JAC. Regression modelling. In: K M, ed. Medical Statistics. USA: Blackwell Publishing company, 2003:339-42.

64. Newson RB. Attributable and unattributable risks and fractions and other scenario comparisons. Stata Journal 2013;**13**(4):672-98. URL: <Go to ISI>://WOS:000329680600001.

65. Newson RB. Attributable and unattributable risks and fractions and other scenario comparisons. Stata J 2013;**13**:672-98. URL.

66. NHS Employers. Quality and Outcomes Framework guidance for GMS contract 2013/14. In: Health, ed. Online: NHS Employers, 2013.

67. Soljak M, Samarasundera E, Indulkar T, et al. Variations in cardiovascular disease under-diagnosis in England: national cross-sectional spatial analysis. BMC Cardiovascular Disorders 2011;**11**(1):12. URL: <http://www.biomedcentral.com/1471-2261/11/12>.

68. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999;**8**(2):135-60. URL: <http://smm.sagepub.com/content/8/2/135.abstract>.

69. Ludbrook J. Linear regression analysis for comparing two measurers or methods of measurement: But which regression? Clinical and Experimental Pharmacology and Physiology 2010;**37**(7):692-99. URL: <http://dx.doi.org/10.1111/j.1440-1681.2010.05376.x>.

70. Francq BG, Govaerts B. How to regress and predict in a Bland–Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. Statistics in Medicine 2016;**35**(14):2328-58. URL: <http://dx.doi.org/10.1002/sim.6872>.

# Appendix: additional information