# Osteoarthritis prevalence models for small populations

## Technical document produced for Arthritis Research UK

Samanta Adomaviciute, Michael Soljak, Hilary Watt, Kim Foley, Roger Newson
Department of Primary Care & Public Health
School of Public Health

October 2018

Imperial College London

# Contents

# Osteoarthritis prevalence models for small populations: Technical document produced for Arthritis Research UK

## 1. Background

Osteoarthritis (OA) is a major public health issue due to its high prevalence, costs, pain and disability.[1] Over 8.5 million people in the UK have OA, while 71% of them (6 million) are in constant pain.[2 3] Common OA symptoms are soreness, pain, stiffness, swelling, tenderness, aching and/or discomfort.[4] OA can affect any joint, but the most commonly affected are the knee, hip and hand.[2]

In a 1995 study the age- and sex-standardized incidence rate for hand OA was 100/100,000 person-years (95% CI 86-115), for hip OA 88/100,000 person-years (95% CI 75-101), and for knee OA 240/100,000 person-years (95% CI 218-262).[5] Causes and risk factors are multifactorial, but about half are explained by genetic factors and half by environmental factors. Knowledge of risk factors for onset of OA is helpful as most of them are modifiable, such as obesity. A recent study analyzed long-term trends in knee OA prevalence in the United States using cadaver-derived skeletons of people aged ≥50 y whose BMI at death was documented, and who lived during the early industrial era (1800s to early 1900s; n = 1,581) and the modern postindustrial era (late 1900s to early 2000s; n = 819).[6] Knee OA among individuals estimated to be ≥50 y old was also assessed in archeologically derived skeletons of prehistoric hunter-gatherers and early farmers. Overall, knee OA prevalence was found to be 16% among the postindustrial sample but only 6% and 8% among the early industrial and prehistoric samples, respectively.

The total cost of OA to the UK economy is estimated to be 1% of annual gross national product with 36 million lost working days between 1999-2000.[1 2] Musculoskeletal (MSK) disease is the fourth highest area of NHS spending after mental health disorders (1st), circulation problems (2nd) and cancers/tumours (3rd).[2 7] NHS spending on MSK diseases has steadily increased from £3.14 billion in 2003/4 to £5.34 billion in 2012-13. [7]

Independence is affected as people with OA are not able to have a full and active work and family life: 80% of people with OA have limitations of movement, while 25% cannot perform their main daily activities.[1] Most physical disability in the UK is caused by arthritis as almost 70% of people with this disease experience constant chronic pain. [1 2] Knee OA is one of the most frequent causes of pain and disability [8].

This public health burden is expected to increase as the population ages, obesity becomes more prevalent, fewer people are physically active, and the age of diagnosis drops.[1 2]. A worrying estimate that there will be 17 million people in the UK living with OA by 2030.[2] The scale of the problem highlights the importance of OA and the need for solutions including appropriate health services and care pathways, and the presence of clear and robust policy recommendations.[2] We review OA risk factors here and present new evidence about them from our analysis. This confirms the importance of reducing modifiable risk factors to prevent OA from developing or deteriorating. New Guidelines from the National Institute for Clinical & Public Health Excellence were published in February 2014.[9] A recent network meta-analysis of paracetamol, diclofenac, ibuprofen, naproxen, celecoxib, intra-articular (IA) corticosteroids, IA hyaluronic acid, oral placebo, and IA placebo using a Bayesian random-effects model showed all treatments except paracetamol resulted in clinically significant improvement from baseline pain.[10] So it is important to ensure that OA cases are identified and appropriate treatments offered.

Commonly OA diagnosis is delayed leading to more severe disease at presentation. In 2011 the time from presenting with first OA symptoms to diagnosis was on average 2.8 years compared to a shorter period (18 months) in 2003.[2] Diagnosis time varies geographically – it takes on average three years to be diagnosed with OA in Scotland, 1.5 years in Northern Ireland and two years in Wales and England. It is essential to promote better self-management for OA, increase earlier diagnosis, and better integrate health and social care.

## 1.1. OA Risk Factors

A literature search was conducted using the MedLine database, and was supplemented with key references provided by Arthritis Research UK. Table 1 shows the risk factors we identified, together with the references from which the risk factors were sourced.

**Table 1: osteoarthritis risk factor table provided by Arthritis Research UK**

| Risk factor | References |
|---|---|
| Age | [11-13] |
| Gender | [11 14] |
| Obesity/BMI | [11 15-24] |
| Previous injury/trauma/joint malignment | [11 25-31] |
| Smoking | [11 32] |
| Occupation | [11 28 33-36] |
| Physical activity/sport | [11 37-40] |
| Socioeconomic | [11 41] |
| Functional/social/leisure limitations | [42] |
| Hormone replacement therapy | [43] |

### 1.1.1 Age

Most of the studies analysed in the 2010 systematic review and meta-analysis by Blagojevic et al showed that increasing age was a risk factor, but it was not possible to identify the pooled odds ratio (OR) due to differences in age-group categorisation and differing age ranges.[44] The Chingford study found an association between age and incident osteophytes (OR 2.41, 95% CI 1.11-5.24).[45] The incidence of knee hand, hip and knee OA increases dramatically with age, especially after age 50. For example, knee OA incidence rates among women aged 20-29 was virtually zero person-years, while for women aged 70-79 it was 1,082/100,000.[4] Men had lower incidence rates of knee OA in old age: in the 20-29 age group it was 5/100,000 person-years, while in 70-79 age category it was 839/100,000.

### 1.1.2 Sex

The Chingford study showed that 2-3% of women will develop OA every year.[45] The 2010 systematic review found nine studies which reported that females are more likely to develop knee problems, with pooled ORs of 1.84 (95% CI: 1.32-2.55).[44] Another meta-analysis of 12 studies (n=22,359) showed a relative risk (RR) reduction for knee OA for males with a pooled RR of 0.63 (95% CI 0.53-0.75).[46] The same study found 13 studies (n=30,762) in which a risk reduction for males was observed for hand OA, with a pooled RR of 0.81 (95% CI: 0.73-0.90). On the other hand, there was no significant sex difference for hip OA (pooled RR 1.18, 95%CI: 0.91-1.52) [46].

### 1.1.3 Obesity/BMI

A meta-analysis showed that being overweight and obese were risk factors for knee OA, with variable effect sizes.[44] The random-effects pooled OR for being obese compared to normal weight subjects was 2.63 (95% CIs: 2.28-3.05). The importance of identifying modifiable risk factors of OA is illustrated by the finding that in the presence of high body mass index (BMI) and subsequent weight loss, OA risk decreased substantially.[44 47] The Chingford study found a strong association between BMI and incident osteophytes (OR 2.38, 95% CI 1.29-4.39), previously shown as a primary diagnostic feature of knee OA in cross-sectional epidemiological studies.[45] Increasing the load and impact on the cartilage has been suggested as possible mechanism through which obesity causes OA.[48] Evaluation of a prospectively followed cohort of Swedish male construction workers found an increase for both hip and knee OA with increasing BMI. Hip OA RR for overweight individuals was 1.54 (95% CI 1.38-1.72), while for obese individuals it was 2.02 (95% CI 1.68-2.43). Another study showed almost double the RR for knee OA was observed in overweight individuals and obese males 4.82 (95% CI 3.65-6.38), respectively.[24]

### 1.1.4 Previous injuries

The 2010 meta-analysis included 14 studies which showed that previous knee injury was a risk factor, yielding the random-effects pooled OR of 3.86 (95% CI: 2.61-5.70).[44] Follow-up of 22 years in a large, nationally representative, population sample of Finns showed that individuals had a 5-fold higher risk of developing knee OA if they had experienced any previous knee injury compared to individuals without an injury.[48]

### 1.1.5 Smoking

In contrast to the generally consistent findings for other risk factors, the 2010 meta-analysis found results from 18 studies of smoking varied from having no effect to having a protective effect, with a pooled OR of 0.84 (95% CI: 0.74-0.95).[44] Smoking is a risk factor for a number of diseases such as cancer, diabetes and cardiovascular problems. On the other hand it is reported to have negative associations with conditions such as ulcerative colitis, Alzheimer's disease and Parkinson's disease. A more recently published (2011) meta-analysis of 48 studies (537,730 participants) showed an overall negative association between smoking and OA (OR=0.87; 95%CI 0.80 to 0.94).[49] One reason for the observed finding could be that smokers are less physically active and that they tend to be leaner. It is important to note that the negative association was only observed for current smokers, which lessens the plausibility of a causal relationship with OA. Moreover, a stratified analysis by study design shows that only case-control studies have a significant association, and smoking becomes neutral in cohort and cross-sectional studies, implying the possibility of the association being a false negative.[49] In the Finnish cohort study it was found that current smokers had half of the risk developing knee OA compared to people that never smoked, with an OR of 0.5 (95% CI: 0.3-1.0; adjusted for age and sex).[48] However, there is no biological mechanism to explain this finding.

### 1.1.6 Occupational activities

Studies analysing general occupational physical workload and joint stress produced mixed results. It was found that sitting jobs (>2 hours/day) had a protective effect, while jobs with excessive kneeling, squatting, climbing steps, standing (>2 hours/day) and lifting increased OA risk.[44 50] An analysis of 518 patients listed for surgical knee treatment identified elevated risk levels for people in occupations involving kneeling or squatting (OR=1.9; 95% CI: 1.3-2.8), walking more than 2 miles per day (OR=1.9; 95%CI: 1.4-2.8) and regularly lifting heavy (>25 kg) weights (OR=1.7; 95% CI: 1.2-2.6). [50] The Finnish cohort study showed that people with severe physical stress at work had up to an 18-fold higher OA risk compared to people with very low physical activity (OR adjusted for age and sex was 11.5, 95% CI: 2.9-45.8; OR adjusted for all covariates – 18.3, 95% CI: 4.2-79.4).[48] Heavy physical stress might cause and sustain micro- and macro-scale joint tissue damage leading to secondary knee OA.

### 1.1.7 Physical activity

Although the expert consensus view is that "normal" (non-professional) physical activity does not cause OA, the effects of various specific types of exercise on the development of OA are unclear and there are varying published findings.[51-53] An analysis of the Framingham Heart Study participants found an association between the number of hours per day spent in heavy physical activity and the risk of incident radiographic knee OA in the elderly. For example, the OR for more than four hours of heavy physical activity per day compared with no heavy physical activity was 7.0 (95% CI: 2.4-20).[8] 'Heavy' physical activities were considered lifting or carrying objects greater than five pounds, gardening with heavy tools, chopping wood and other strenuous sports or recreation.[8] In contrast, there was no such association with 'Moderate' (lifting/carrying light objects, sweeping/mopping/vacuuming) and 'Light' (standing, ironing, leisurely walking) activities.[8] The longitudinal Framingham Offspring cohort study did not find any relation between recreational walking, jogging, or other self-reported activity and knee OA development.[51] However, regular and more intense exercise was associated with increased OA risk. [44] Spector et al (1996) conducted a retrospective cohort study of female ex-elite athletes and found that women having weight-bearing sports activity had 2-3 times higher risk of radiologic OA of the knees and hips compared with controls.[52]

On the other hand, some studies have shown that moderate exercise is beneficial in improving function in knee/hip OA and minimising pain once it has been diagnosed.[53] The main feature of structural changes taking place in the OA joint is cartilage loss.[53] A four-month randomized controlled trial in patients at risk of OA examined the biochemical properties of cartilage tissue to answer the question whether exercise worsens OA.[53] The finding concluded that moderate, supervised exercise improves knee-cartilage glycosaminoglycan content (building blocks of proteoglycans and are essential for the important viscoelastic properties of cartilage) in patients at risk of OA.[53] Therefore, based on this study's findings, exercise may have potential benefits in preventing the development of knee OA, as improvements in pain and function were observed in the study participants. Exercise is recommended as a core treatment for people with OA in the 2014 NICE guidelines.

### 1.1.8 Socioeconomic inequality

Relatively little has been published about associations with deprivation or social class. The General Lifestyle Survey 2009 showed that, compared to social class I, people in social class V have 60% higher prevalence of long term conditions and 30% higher severity of conditions, though this varies significantly by condition.[54] In HSfE 2011 there was significant variation in the distribution of Chronic Pain Grades by equivalised household income.[55] In one study micro-level data were pooled from non-standardized national health surveys conducted in eight European countries in the 1990s to find any associations between chronic diseases and education.[56] Analysis showed that OA was more prevalent in the lower education group compared to higher education group (OR=1.34; 95%CI: 1.21-1.49). Our analysis presented below shows that, after adjustment, education, socioeconomic class and obesity are all independent risk factors for hip OA.

### 1.1.9 Hormone replacement therapy

Hormone replacement therapy (HRT) is positively associated with reduction of menopausal symptoms, heart disease and cerebrovascular disease. In the Chingford study 1,003 women aged 45-64 (mean 54.2) were asked details of HRT use.[43] A significant protective effect for knee OA, as defined by osteophytes, was observed for current HRT users (n=72) compared to never users with OR of 0.31 (95% CI: 0.11-0.93). A smaller non-significant risk reduction was seen for ever users of HRT (n=129) for knee OA compared to ever users (OR=0.80, 95% CI: 0.43-1.49). Moreover, no protective association was observed for HRT ex-users.

**Table** 2 displays pooled and/or adjusted odds ratios for the risk factors or osteoarthritis, compiled from the ARUK table and our own literature search.

**Table 2: Osteoarthritis risk factors with their pooled or adjusted odds ratios**

| Risk factor | Type of Odds Ratio | Odds Ratio | 95% CI | Effect on Outcome |
|---|---|---|---|---|
| Age | Adjusted [45] | 2.41 | 1.11-5.24 | Risk Factor |

| Risk factor | Type of Odds Ratio | Odds Ratio | 95% CI | Effect on Outcome |
|---|---|---|---|---|
| Female sex | Pooled (1) | 1.84 | 1.32–2.55 | Risk Factor |
| | Adjusted [48] | 1.7 | 1.0-3.1 | Risk Factor |
| Obesity | Pooled (1) | 2.63 | 2.28–3.05 | Risk Factor |
| | Adjusted [45] | 2.38 | 1.29-4.39 | Risk Factor |
| Body mass index | | | | |
| **<18.5** | Adjusted (31) | 2.1 | 1.1–4.2 | Risk Factor |
| **18.5–24.9** | Adjusted (31) | 1 | | |
| **25.0–29.9** | Adjusted (31) | 1.1 | 0.8–1.5 | NS |
| | Adjusted [48] | 1.7 | 1.0-2.8 | NS |
| | Adjusted [50] | 3.2 | 2.2-4.7 | Risk Factor |
| **≥30** | Adjusted (31) | 1.3 | 1.0–1.8 | Risk Factor |
| | Adjusted [48] | 7.0 | 3.5-14.10 | Risk Factor |
| | Adjusted [50] | 8.3 | 5.2-13.4 | Risk Factor |
| Musculoskeletal injury | Adjusted (32) | 5.0 | 1.9, 13.3 | Risk Factor |
| | Adjusted (sex, age) [48] | 4.7 | 1.4-15.5 | Risk Factor |
| Injury | | | | |
| **No** | | 1 | | |
| **Yes** | Adjusted (31) | 5 | 1.9 - 13.3 | NS |
| Smoking | | | | |
| **Never smoked** | | 1 | | |
| **Ex-smoker** | Adjusted (31) | 1 | 0.4, 2.4 | NS |
| | Adjusted (sex, age) [48] | 0.8 | 0.4-1.4 | NS |
| | Pooled [49] | 1.02 | 0.91-1.14 | NS |
| **Smoker** | Adjusted (31) | 0.9 | 0.4, 2.3 | NS |
| | Adjusted (sex, age) [48] | 0.5 | 0.3-1.0 | NS |
| | Pooled [49] | 0.87 | 0.80-0.94 | Protective? |
| Generic work characteristics | | | | |
| **Work status** | | | | |
| **Not working** | | 1 | | |
| **Working** | Adjusted (31) | 1.3 | 1.0–1.7 | Risk Factor |
| Annual household income | | | | |
| **<$20,000** | | 1 | | |
| **≥$20,000** | Adjusted (31) | 0.7 | 0.5–1.0 | NS |
| **Unknown** | Adjusted (31) | 2.1 | 1.2–3.7 | Risk Factor |
| Received benefits in last year | | | | |
| **No** | | 1 | | |
| **Yes** | Adjusted (31) | 1.4 | 1.0–2.0 | Risk Factor |
| Occupational activity | | | | |
| **Lifting ≥10kg >10times/week** | Adjusted [50] | 1.7 | 1.2-2.4 | Risk Factor |
| **Lifting≥25kg >10times/week** | Adjusted [50] | 1.7 | 1.2-2.6 | Risk Factor |
| **Lifting≥50kg >10times/week** | Adjusted [50] | 1.4 | 0.9-2.2 | NS |
| **Sitting >2h/day in total** | Adjusted [50] | 0.7 | 0.5-1.0 | NS |
| **Standing or walking >2h/day in total** | Adjusted [50] | 1.5 | 0.8-2.9 | NS |

| Risk factor | Type of Odds Ratio | Odds Ratio | 95% CI | Effect on Outcome |
|---|---|---|---|---|
| **Kneeling >1h/day in total** | Adjusted [50] | 1.8 | 1.2-2.6 | Risk Factor |
| **Squatting >1h/day in total** | Adjusted [50] | 2.3 | 1.3-4.1 | Risk Factor |
| **Getting up from kneeling or squatting >30times/day** | Adjusted [50] | 1.7 | 1.2-2.6 | Risk Factor |
| **Driving for >4h/day in total** | Adjusted [50] | 0.9 | 0.6-1.5 | NS |
| **Walking >2miles/day in total** | Adjusted [50] | 1.9 | 1.4-2.8 | Risk Factor |
| **Climbing a ladder or flight of stairs >30times/day** | Adjusted [50] | 1.5 | 1.0-2.3 | NS |
| **Kneeling/squatting or heavy lifting** | Adjusted [50] | 1.7 | 1.1-2.7 | Risk Factor |
| **Heavy lifting but no kneeling/squatting** | Adjusted [50] | 1.5 | 0.9-2.4 | NS |
| **Both kneeling/squatting and heavy lifting** | Adjusted [50] | 3.0 | 1.7-5.4 | Risk Factor |
| Physical work load | | | | |
| **Light sedentary** | | 1 | | |
| **Other sedentary** | Adjusted (32) | 1.1 | 0.1- 10.0 | NS |
| **Light standing/movements** | Adjusted (31) | 1.2 | 0.4- 3.4 | NS |
| **(1 h per day)** | Adjusted [8] | 1.7 | 0.7-4.5 | NS |
| **Fairly light or medium heavy** | Adjusted (31) | 3.1 | 1.2- 8.0 | Risk Factor |
| **(1 h per day)** | Adjusted [8] | 1.3 | 0.6-2.7 | NS |
| **Heavy manual** | Adjusted (31) | 6.7 | 2.3- 19.5 | Risk Factor |
| **(1 h per day)** | Adjusted [8] | 2.2 | 1.2-4.2 | Risk Factor |
| Heavy manual labour | Adjusted (32) | 6.7 | 2.3 – 19.5 | Risk Factor |
| Physical activity | | | | |
| **Recommended** | | 1 | | |
| **Insufficient** | Adjusted (31) | 0.9 | 0.7–1.2 | NS |
| **Inactive** | Adjusted (31) | 1.2 | 0.9–1.6 | NS |
| Leisure time physical activity | | | | |
| **Little physical exercise** | | 1 | | |
| **Irregular physical exercise** | Adjusted (31) | 1.2 | 0.5, 2.9 | NS |
| | Adjusted (sex, age) [48] | 0.8 | 0.5-1.3 | NS |
| **Regular physical exercise** | Adjusted (31) | 1.1 | 0.4, 2.8 | NS |
| | Adjusted (sex, age) [48] | 0.5 | 0.3-1.0 | NS |
| Race/ethnicity | | | | |
| **Non-Hispanic white** | | 1 | | |
| **Non-Hispanic black** | Adjusted (31) | 1.6 | 1.2–2.3 | Risk Factor |
| **Hispanic** | Adjusted (31) | 1.8 | 1.2–2.6 | Risk Factor |
| **Non-Hispanic other** | Adjusted (31) | 1.4 | 0.8–2.4 | NS |
| Education | | | | |
| **High school or less** | | 1 | | |
| **High school graduate** | Adjusted (31) | 0.9 | 0.6–1.3 | NS |
| **Some college** | Adjusted (31) | 0.8 | 0.6–1.1 | NS |
| **At least college** | Adjusted (31) | 0.6 | 0.4–0.8 | Protective |
| **High education** | Adjusted [56] | 1 | | |
| **Low education** | Adjusted [56] | 1.34 | 1.21-1.49 | Risk Factor |

| Risk factor | Type of Odds Ratio | Odds Ratio | 95% CI | Effect on Outcome |
|---|---|---|---|---|
| Marital/cohabitating status | | | | |
| **Never married** | | 1 | | |
| **Married/common law** | Adjusted (31) | 1.1 | 0.8–1.5 | NS |
| **Divorced/separated/widowed** | Adjusted (31) | 1.1 | 0.8–1.6 | NS |
| Functional/social/leisure limitations | | | | |
| **No** | | 1 | | |
| **Yes** | Adjusted (31) | 1.8 | 1.4–2.3 | Risk Factor |
| Chronic conditions, no. | | | | |
| **0** | | 1 | | |
| **1** | Adjusted (31) | 1.2 | 0.9–1.7 | NS |
| **2** | Adjusted (31) | 1.1 | 0.8–1.7 | NS |
| **3** | Adjusted (31) | 1.1 | 0.7–1.7 | NS |
| **4** | Adjusted (31) | 1 | 0.6–1.7 | NS |
| **5** | Adjusted (31) | 0.9 | 0.5–1.6 | NS |
| **≥6** | Adjusted (31) | 0.8 | 0.5–1.3 | NS |
| Neck or back pain | | | | |
| **No** | | 1 | | |
| **Yes** | Adjusted (31) | 1.5 | 1.2–1.9 | Risk Factor |
| Anxiety/depression | | | | |
| **No** | | 1 | | |
| **Yes** | Adjusted (31) | 1.1 | 0.8–1.4 | NS |
| Recurring pain | | | | |
| **No** | | 1 | | |
| **Yes** | Adjusted (31) | 1.7 | 1.3–2.2 | Risk Factor |
| Self-rated health in general | | | | |
| **Fair/poor** | | 1 | | |
| **Good/very good/excellent** | Adjusted (31) | 0.7 | 0.5–0.9 | Protective |
| No. office visits to any doctor in past year | | | | |
| **0–7** | | 1 | | |
| **≥8** | Adjusted (31) | 1.4 | 1.1–1.8 | Risk Factor |
| Alcohol intake, g/week | | | | |
| **0** | | 1 | | |
| **1–49** | Adjusted (31) | 1.1 | 0.5- 2.4 | NS |
| **50–249** | Adjusted (31) | | | |
| **>250** | Adjusted (31) | 2.2 | 0.6- 7.7 | NS |
| Hormone Replacement Therapy (HRT) | | | | |
| **Never** | | | | |
| **Current** | Adjusted [43] | 0.31 | 0.11-0.93 | Protective |
| **Ever** | Adjusted [43] | 0.80 | 0.43-1.49 | NS |

## 1.2 Objectives

The overall aim of the project is to develop small population estimates of OA prevalence and need for related healthcare, and to relate this to actual and expected activity and costs.[1] Specific objectives are:

[1] For specific information relating to activity levels and costs of hip and knee replacements in your local area, please contact data@arthritisresearchuk.org.

- To develop from nationally (England) representative survey data prevalence model for OA
- To apply these to English general practice and Middle Layer Super Output Area (MLSOA) populations
- To project these estimates to 2021-22 using population age and other risk factor projections
- To include within these where possible categorical estimates of severity, impairment and need for healthcare

# 2 Methods

## 2.1 Data source

The English Longitudinal Study of Ageing (ELSA) data was used to develop hip and knee OA models. ELSA is a large multicentre and multidisciplinary study of people aged 50 and over and their partners, living in private households in England.[57] ELSA was chosen as the basis of our model since OA is rare under 50 years of age.[5] Additionally, the survey uses patient-reported doctor diagnosed disease criteria; it includes questions concerning limitations with activities of daily living; and it allows us to differentiate disease severity. ELSA was chosen over 2011 Health Survey for England (HSfE) and General Lifestyle Survey (GLS) as the latter surveys could not differentiate well between different MSK diseases.

The sample for the survey was designed to be representative of the English population. The sampling frame was drawn from households that had previously responded to the HSfE in 1998, 1999 or 2001. The HSfE dataset thus established what is known as "ELSA Wave 0", and included information regarding housing/accommodation, education, employment, income, food and drink consumption (including consumption of fruit and vegetables), smoking, physical activity, biomedical measurements (such as blood pressure, BMI), cardiovascular disease (and associated risk factors) and health of ethnic minority groups, among others.

The sample population was approached in the field during 2002-2003, and these respondents thus comprise the baseline “ELSA Wave 1” study group. The core ELSA questionnaire was administered by computer-aided personal interviewing (CAPI), together with a paper self-completion questionnaire. Respondents were then re-approached every two years (with a nurse visit offered at alternate interviews at Waves 2 and 4).[58] **Table 3** below demonstrates why ELSA was chosen as the data source.

**Table 3: Outcome measures and survey datasets**

| | ELSA | HSfE | GLS |
|---|---|---|---|
| **Arthritis** | ✓ | ✓ | ✓ |
| **Osteoarthritis** | ✓ | x | x |
| **Rheumatoid Arthritis** | ✓ | x | x |
| **Other Arthritis** | ✓ | x | x |
| **Back pain** | ✓ | ✓ (2011) | x |

Arthritis is recorded in all three datasets as a single question. ELSA is more specific than HSfE or GLS, allowing a specific type of arthritis to be recorded, which was an important reason for selecting it. The recording of arthritis in ELSA and HSfE are displayed in **Table 4**. In ELSA Wave 4 the **hedbdar** variable records ‘whether confirms arthritis diagnosis’. If arthritis is indicated in the ELSA questionnaire, the Interviewer asks: “May I check, which type or types of arthritis have...

1. Osteoarthritis?
2. Rheumatoid arthritis?
3. Some other kind of arthritis?”

ELSA also details what was diagnosed in the previous wave (Wave 3), so we can determine if the arthritis was previously diagnosed or was newly identified in Wave 4. Each patient has a unique identifier so it is possible to link these data across years for a more complete picture of previous diagnoses. Although the same data have been collected in Wave 1 and Wave 4, it is recorded slightly differently. Each variable allows the recording of an arthritis type with values of OA, rheumatoid arthritis (RA) and other arthritis. In Wave 4 the arthritis type is divided into separate variables.

Primary/secondary arthritis types cannot be determined. The arthritis variable names are listed in **Table 4**.

**Table 4: ELSA data variables for arthritis type**

| Type | Wave1 2002/2003 Variable | Value | Wave4 2008/9 Variable | Value | Wave5 2010/11 Variable | Value |
|---|---|---|---|---|---|---|
| Osteoarthritis | heart1 or<br>heart2 or<br>heart3 | OA<br>OA<br>OA | Heartoa | mentioned | heartoa | Mentioned |
| Rheumatoid Arthritis | heart1 or<br>heart2 or<br>heart3 | RA<br>RA<br>RA | Heartra | mentioned | heartra | Mentioned |
| Other Arthritis | heart1 or<br>heart2 or<br>heart3 | Other<br>Other | Heartot | mentioned | heartot | Mentioned |

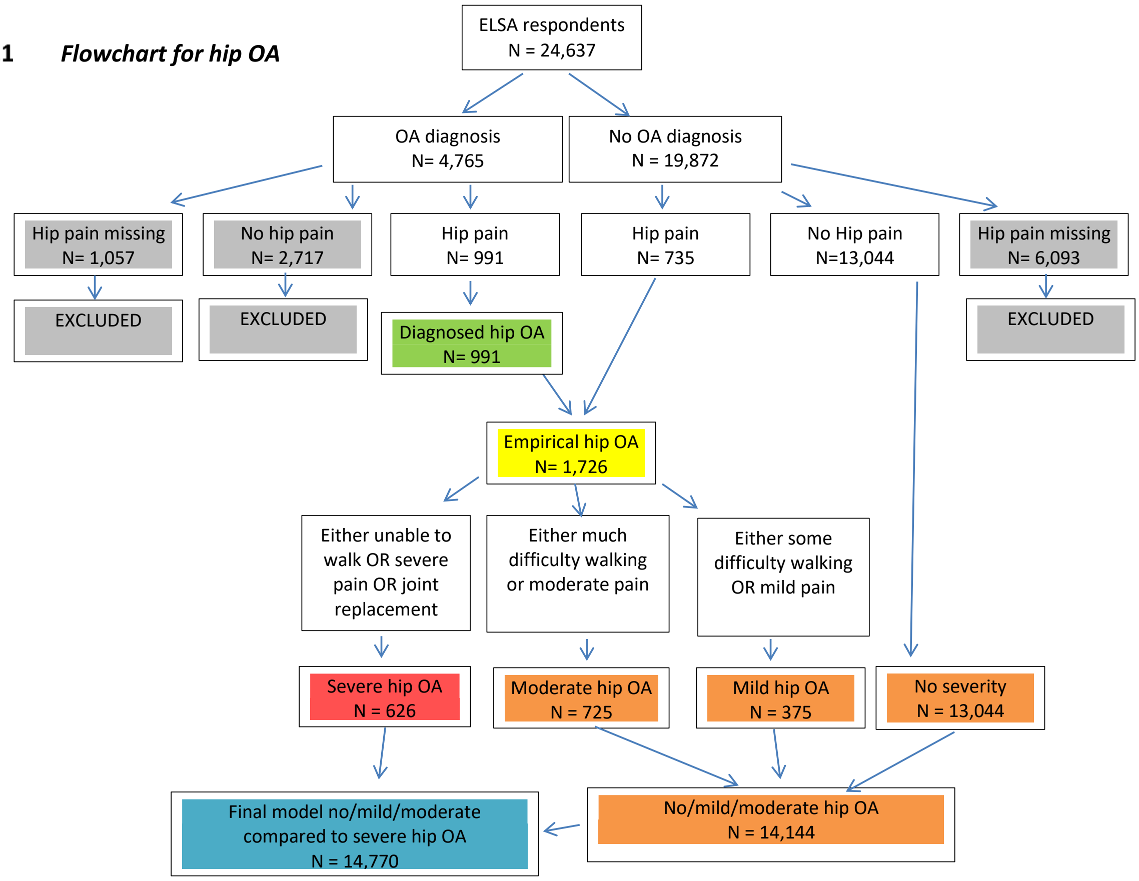
The two OA models described here use ELSA data from Wave 0 to Wave 5, with data linked across years using the unique patient identifier. This logistic analysis is therefore cross-sectional at ELSA Wave 5, with most individual information taken from Wave 5. However, if the information is missing at Wave 5, it is taken from next previous Wave (going from 5 to 0 sequentially). This method is used because some individuals were present only in some Waves, and some of them were enrolled in the study at different Waves. This method allows us to maximise our sample size.  A limitation of ELSA is that, as the name implies, it includes no data on under 50 year olds. Therefore hip and knee OA prevalence is not modelled in younger age groups- we chose 45-64 as the youngest age group because of the availability of local data for this age group, to which we applied ELSA data for 50-64.

## 2.2 Outcome definitions

There were 24,637 respondents in the merged dataset after combining ELSA Waves 0 to 5. A total of 19,872 (80.66%) out of 24,637 do not have OA diagnosis, while 4,765 (19.34%) do have an OA diagnosis. Those recorded as having hip pain are also classified as **diagnosed hip OA** cases, as there is evidence that about 95% of those with hip pain will have OA:[59] 735 (3.70%) of 19,872 do not report an OA diagnosis but indicated the presence of hip pain. Therefore, the **empirical hip OA** case definition (n=1,726) includes diagnosed hip OA cases (n=991) as well as respondents that recorded hip pain (n=735). Respondents that do not have hip pain information are excluded from analysis (n=7,150) as well as respondents who indicated presence of non-site specific OA but do not have hip pain (n=2,717), unless they had a joint replacement (see Figure 1 below, also  2: ELSA data flowchartsFlowchart for hip OA).

**Figure 1: flowchart for hip OA**

### 1.1 *Flowchart for hip OA*



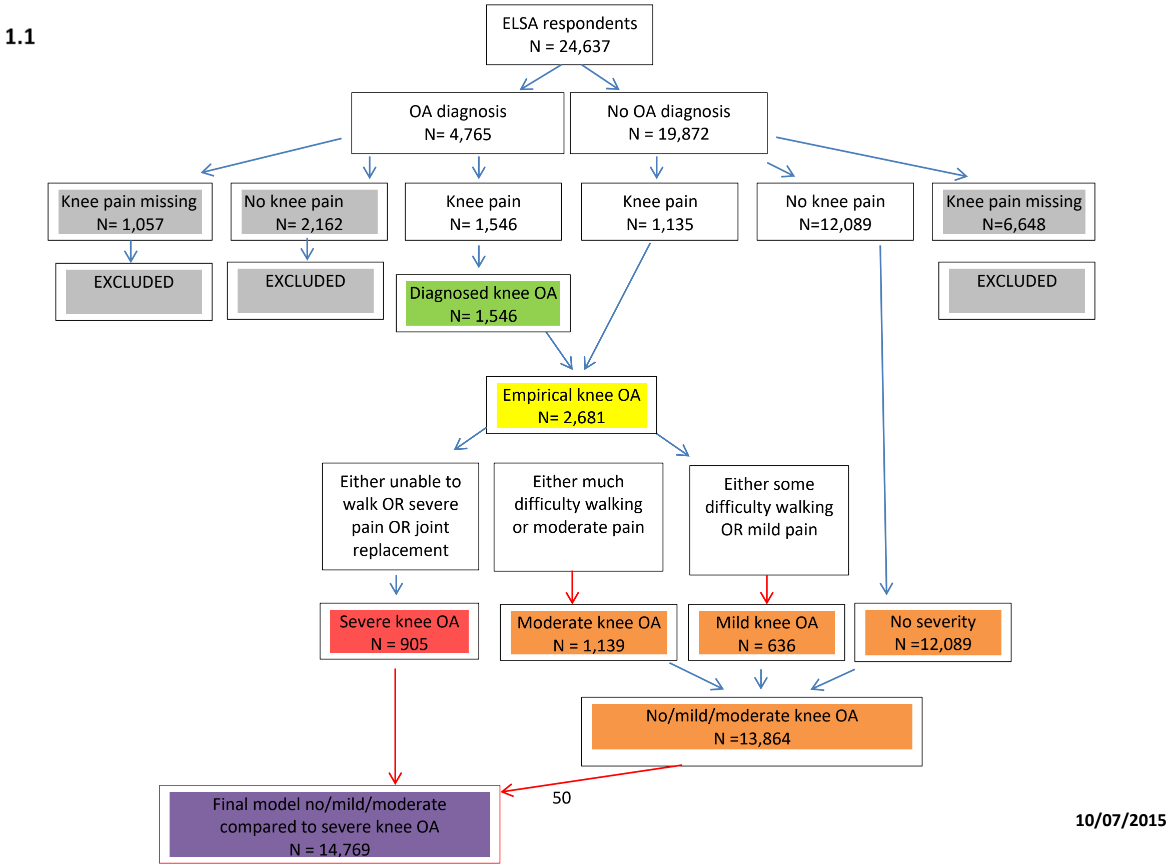
**Of 4,765 OA diagnoses recorded as having knee pain, 1,546 (32.44%) are classified as** diagnosed knee OA **cases. In addition, 1,135 (5.72%) out of 19,872 do not record an OA diagnosis, but indicated the presence of knee pain. Hence our agreed** empirical knee OA **case definition (n=2,681) included diagnosed knee OA cases (n=1,546) as well as respondents that recorded knee pain (1,135). Respondents that do not have knee pain information are excluded from analysis, unless they have had knee replacement (n=7,705) as well as respondents who indicated presence of non-site specific OA but do not have knee pain (n=2,162) (see**

Figure 2, also Appendix 2: ELSA data flowcharts, Section 9.2 Flowchart for knee OA).

Hip and knee OA cases were grouped into three severity categories (mild, moderate, severe) based on the pain severity question (4 groups: none, mild, moderate and severe) and walking interference question (4 groups: no difficulty, some difficulty, much difficulty and inability to walk). Respondents were assigned to a specific group by either having pain or mobility interference. Any respondents that had joint replacements due to arthritis were classified as being in the high severity category. All controls were assigned to the 'No severity' category. Informants were deemed to have the 'severe' form of OA if their answers included any one of the following statements:

- severe pain most of the time;
- unable to walk ¼ mile unaided;
- had previously undergone hip or knee replacement due to OA.

**Figure 2: flowchart for knee OA**


1.1
ELSA respondents N = 24,637
OA diagnosis N= 4,765
No OA diagnosis N = 19,872
Knee pain missing N= 1,057
No knee pain N= 2,162
Knee pain N= 1,546
Knee pain N= 1,135
No knee pain N=12,089
Knee pain missing N=6,648
EXCLUDED
EXCLUDED
Diagnosed knee OA N= 1,546
EXCLUDED
Empirical knee OA N= 2,681
Either unable to walk OR severe pain OR joint replacement
Either much difficulty walking or moderate pain
Either some difficulty walking OR mild pain
Severe knee OA N = 905
Moderate knee OA N = 1,139
Mild knee OA N = 636
No severity N =12,089
No/mild/moderate knee OA N =13,864
Final model no/mild/moderate compared to severe knee OA N = 14,769
50
10/07/2015

Out of 1,726 respondents with hip OA, 626 (36.27%) have severe hip OA, while 905 (33.76%) have severe knee OA out of 2,681 respondents with knee OA. Models predicting severe hip or knee OA cases from the whole population (combination of no severity, mild and moderate) have good prediction and their results are presented in this document.

## 2.3 Risk factors in the model

**Table 5** shows the list of risk factors identified in the literature review and the proportion of the missing data for each of them in ELSA. Because of the relatively low levels of missing data we did not use methods such as multiple imputation to replace it. Those variables with significant data missing had such high levels of it that multiple imputation could not be used. Stata software drops observations with missing data from the analysis.

**Table 5 Risk factors and their missing proportion in ELSA**

| Risk Factors Considered in ELSA | Inclusion / Exclusion Criteria | Missing data in ELSA | Missing data for hip OA (ESLA) | Missing data for knee OA (ELSA) |
|---|---|---|---|---|
| Age | Include | 0% | 0% | 0% |
| Sex | Include | 0% | 0% | 0% |
| Ethnicity | Include | 0.25% | 0% | 0% |
| Education | Include | 0.69% | 0.25% | 0.34% |

| | | | | |
|---|---|---|---|---|
| Socioeconomic status | Include | 22.03% | 1.85% | 1.57% |
| Obesity/BMI | Include | 16.82% | 13.27% | 13.02% |
| Physical activity (leisure) | Include | 17.45% | 0% | 0% |
| Smoking | Include | 33.08% | 0.06% | 0% |
| Member at sports clubs, gyms etc. | Include | 32.76% | 3.19% | 3.32% |
| Physical activity (work) | Exclude | 31.12% | 35.46% | 32.94% |
| Housework/gardening activity level | Exclude | 50.36% | 59.97% | 58.93% |
| Lifting at work | Exclude | 86.22% | 87.95% | 87.36% |

Variables, definitions and categories included in the regression modelling are shown in **Table 6**.

**Table 6: variables included in regression models**

| Variable | Variable definition | Variable categories |
|---|---|---|
| **Sex** | Sex | Male/Female |
| **Age group** | Age 45+ in year bands | Missing, 45-64, 65-74, over 75 |
| **Ethnicity** | Ethnic origin of individual | Missing, White, non-white |
| **Socioeconomic factors** | Nine groups | Missing, Higher managerial and professional occupation, Lower managerial and professional occupation, Intermediate occupations, Small employers and own account workers, Lower supervisory and technical occupations, Semi-routine occupations, Routine occupations, Never worked and long term unemployed *(Other was excluded as it coded for missing, incomplete data)* |
| **Education** | Obtained education | Missing, NVQ4/NVQ5/Degree or equivalent, Higher education below degree, NVQ3/GCE A level equivalent, NVQ2/GCE O level equivalent, NVQ1/CSE other grade equivalent, Foreign/other, No qualification |
| **BMI**[2] | BMI grouped into four categories | Missing, <18.5 - underweight, 18.5-24 – normal range, 25-29 – overweight, >30 – obese |
| **Leisure physical activity** | Levels of physical activity grouped into four categories | Missing, Sedentary, Low, Moderate, High |
| **Smoking status** | Cigarette Smoking Status grouped into three categories | Missing, Current cigarette smoker, Ex-regular cigarette smoker, Never regular cigarette smoker |
| **Membership at gym/sports club** | Presence or absence of the gym/sports club membership | Missing, No, Yes |

## 2.4 Regression modelling

Baseline characteristics and analyses were performed using Stata v11. All variables are recoded to drop negative values for estimation purposes (in ELSA various non-response categories are assigned negative values). The methodology applied for model fitting is logistic regression. A full explanation of this method is available from textbooks such as *Medical Statistics*.[60] The US Centers for Disease Control have recently begun a programme of small population prevalence estimates for which they used multilevel logistic regression because some of the survey data was reported at different geographic levels.[61] However this is not the case with ELSA data.

[2] BMI grouping was changed to match the format found in Active People Survey (three 'obese' categories were merged into one 'obese' if BMI was more than 30.

The choice of variables for original inclusion in the merged dataset included all those known to be hip or knee OA risk factors. The choice of ELSA variables for inclusion in the new prevalence model was based on the data availability in ELSA for the selected risk factors. Explanatory variables obtained from the ELSA dataset and included in the final model were age, sex, ethnicity, socioeconomic status, education, BMI categories, leisure physical activity, smoking status and membership in the gym/sports club. Questions about housework activity and lifting at work were excluded from the models as they had high levels of missing data- 50.36% and 86.22%, respectively. Work physical activity was excluded from the models as it had 31.12% missing data and there was no local area data.

"Complete" models (using all available variables) are selected by reverse stepwise selection using likelihood ratio and Wald tests i.e. the model is fitted using all available exposure variables, omitting each in turn and recording p values.[62] The variable with the highest p value is omitted and the tests are repeated. Since the models are to be used for prediction rather than hypothesis testing, a p value of 0.05 is used. We call these models "complete" because risk factors can only be applied to local data to produce estimates if it exists at local level. Comparing the performance of a "local" model with a "complete" model is a good method of internal validation.

For categorical variables the effects are estimated relative to the reference category. Stata uses the first category as reference (baseline OR). The four derived separate models can be used to derive the prevalence ratios for total and severe hip or knee OA for subjects with various combinations of risk factors in relation to baseline. The prevalence in each age, sex, socioeconomic status, BMI, physical activity level, (plus smoking status, education and presence of gym membership – these are additions for knee OA) were derived from the odds, using the formula:

Prevalence = Odds/(1 + Odds).

$$Severe/Total\ OA = \alpha + \beta_1 \times RF_1 + \beta_2 \times RF_2 + \cdots + \beta_n \times RF_n$$

- $RF$ is individual Risk Factor
- The outcome is 1 for reported total/severe hip/knee OA, 0 otherwise
- Logistic regression is used since the outcome variable is binary

The initial output consists of two tables each for hip and knee OA: one with the estimated regression coefficients, corresponding p-values and 95% confidence intervals, and another with the odds ratios, corresponding p-values and 95% confidence intervals. A positive sign of the estimated coefficient is associated with an increase in the odds of the outcome, and a negative sign is associated with a decrease in the odds. Once ORs are obtained, internal validation is carried out using areas under receiver operating characteristics (ROC) curves and prediction probabilities, and the modelled ORs are then used to derive the predicted probability that the specified survey informant has the disease, based on their risk factors. These predictions are called fitted values. The difference between the fitted and the observed values are called residuals. These can then be tabulated against the observed presence of the disease to assess misclassification by each model in a 2x2 table.

## 2.5 Regression model internal validation

Ideally the best prediction should result from utilising the most information in the regression model. We performed an initial validation of the local model. Sensitivity and specificity are calculated using the area under the ROC curve using Stata 11.[63] ROC analysis (also known as c-statistic) is a useful tool for evaluating the performance of diagnostic tests and more generally for evaluating the accuracy of a statistical model e.g., logistic regression, linear discriminant analysis that classifies subjects into one of two categories, diseased or non-diseased, as in this model.[64 65] Its function as a simple graphical tool for displaying the accuracy of a medical diagnostic test is one of the most well-known applications of ROC curve analysis.

A ROC curve is a plot of sensitivity on the y axis against (1-specificity) on the x axis for varying values of the threshold t. The 45° diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding to random chance. The ROC curve for the gold standard is the line connecting (0,0) to (0,1) and (0,1) to (1,1). Generally, ROC curves lie between these two extremes. The area under the ROC curve is a summary measure that essentially averages diagnostic accuracy across the spectrum of test values, and is an overall summary of diagnostic accuracy. Area under ROC curve equals 0.5 when the curve corresponds to random chance, and 1.0 for perfect accuracy. On rare occasions, the estimated area under the curve is less than 0.5, indicating that the test does worse than chance. We compared the performance of different models using Hanley's methods to calculate CIs.[66]

Another method of assessing performance is to use the regression model to predict the response for each subjectWe tabulated predicted against the observed presence of hip or knee OA to assess "misclassification" by each model.

## 2.6 External validation

We used Clinical Practice Research Datalink (CPRD) data to fit a logistic regression model for patients who had knee or hip OA. The details of the validation is shown in section 5.

## 2.7 Data issues

Data was missing for the ELSA outcome variable: 40.05% of respondents did not answer the pain question (whether hip or knee pain was absent/present), so we did not know whether or not they had OA. Unless this data is missing at random, the dataset may not give a definitive answer as to the prevalence of OA, nor necessarily to the age/sex trend in prevalence. Rather than relying on the inaccurate prevalence in this dataset, or losing all the data from these respondents from the analysis, we substituted hip and knee OA prevalence by age and sex from a cohort with more complete ascertainment of OA status.[67] We did this by performing inverse probability weighting of our results (*pweight* option in Stata), which is appropriate where estimated probabilities reflect the probability that a person with these characteristics is included in our results (this could reflect differential probabilities that outcome measures will be missing in different age groups).[68] We assumed that all OA cases have equal probability, and then adjusted the probability of the controls by age and sex group, so that weighted prevalence by age and sex groups equalled those reported in the Keele report.[67] The implicit assumption here is that probability of being included in the model is related only to age and sex, and is not related to the value of other risk factors assessed (education, ethnicity, BMI, socioeconomic factors, smoking status, leisure and membership at sports club).

## 2.8 Synthetic estimation: application of the model to small population data

Derived ORs 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, ONS socio-economic Class (NSSeC), which is used as a proxy for occupational group, smoking, ethnicity, education, obesity, long-term limiting illness (LLI) and fractures. The local model uses locally available data. If data is not available at MLSOA level but is available at LA level, LA values are applied to MLSOA populations, using the median generated at LA level (except for work physical activity, which was excluded from the model as it is not available at any level). An alternative to doing this is to use a multi-level model.[61]

The “local” model includes only those variables that are available at local population level i.e. age, sex, socioeconomic status, BMI, leisure physical activity, smoking status, education[3] and gym membership.[4] 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:

$$\text{Predicted log odds of prevalence} = b_0 + b_1x_{1i} + b_2x_{2\,i} + b_3x_{3\,i} + b_4x_{4\,I}$$

$$\text{where } b_0 = \text{regression constant, } b_1, b_2, b_3, b_4 = \text{other regression coefficients}$$

$$x_{1\,i}, x_{2\,i}, x_{3\,i}, x_{4\,i} = \text{value of risk factors for individual } \boldsymbol{i}$$

(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:

$$\text{Predicted log odds of prevalence in community of } n \text{ individuals:}$$

$$= 1/n \sum_{i=1}^{n} (b_0 + b_1x_{1i} + b_2x_{2\,i} + b_3x_{3\,i} + b_4x_{4\,i})$$

$$= b_0 + b_1p_1 + b_2p_2 + b_3p_3 + b_{4p}p_4$$

where $p_1, p_2, p_3, p_4$=proportion of individuals in the community with characteristic $x_1, x_2, x_3, x_4$. (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:

$$\text{prevalence} = \exp(\text{log odds})/[1+\exp(\text{log odds})]$$

$$=\exp(b_0 + b_1x_{1i} + b_2x_{2\,i} + b_3x_{3\,i} + b_4x_{4\,i})\,/[1+ \exp(b_0 + b_1x_{1i} + b_2x_{2\,i} + b_3x_{3\,i} + b_4x_{4\,i})]$$

$$\text{Predicted prevalence in community of } n \text{ individuals:}$$

$$= 1/n \sum_{i=1}^{n}\{\exp(b_0+b_1x_{1i}+b_2x_{2\,i}+b_3x_{3\,i}+b_4x_{4\,i})/[1+ \exp(b_0 +b_1x_{1i}+b_2x_{2\,i}+b_3x_{3\,i}+b_4x_{4\,i})]\}$$

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.

[3] Only for severe/total knee and total hip OA models

[4] Only for severe knee and severe hip OA models

In order to find a synthetic estimate of prevalence, we need to know the distribution 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 can assume 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. 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, where 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. 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), are distributed equally amongst 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, weighted 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. We used both methods for the MSK calculator as a means of validating the synthetic estimation step.  The Stata code we developed and used is included in Section 10.1 Synthetic estimates in Appendix 3: synthetic estimation using Stata software.

Table 7 shows the local data used in the overall OA model as an example. Where data was only available at LA level we used that result for constituent MSOAs. We will explore the possibility of obtaining more specific estimates for lower geographies in Phase 2 of the project.[69 70]  For example, Sport England has produced synthetic estimates of these variables for sub-LA geographies, but we had not been able to obtain this data prior to the publication of this document, and were not able to obtain it subsequently.

**Table 7: local data used in the overall OA model**

| Risk factor data | At MLSOA | At LA | Source | Year |
|---|---|---|---|---|
| **Age and sex** | Present | Calculated | Office for National Statistics | Mid-2012 |

| Risk factor data | At MLSOA | At LA | Source | Year |
|---|---|---|---|---|
| **Socioeconomic status** | Present | Present | Census | 2011 |
| **BMI** | Absent | Present | Sport England Active People Survey | Mid-January 2012 to mid-January 2013 |
| **Smoking status** | Present | Calculated | Integrated Household Survey | 2011 |
| **Education data** | Present | Calculated | Census | 2011 |
| **Physical activity** | Absent | Present | Sport England Active People Survey | Mid-January 2012 to mid-January 2013 |
| **Gym membership** | Absent | Present | Sport England Active People Survey | October 2012 to October 2013 |

## 2.9 Calculating confidence intervals for prevalence estimates using boot-strap procedures

As of the time of revising the Technical Document in October 2018, CIs for the local estimates have been produced but are not yet on the MSK Calculator website for reasons related to the website itself. However we have included here the methods used for their calculation to demonstrate what has been done and what can be anticipated in the near future.

The philosophy underlying the boot-strap 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.

Boot strap samples are taken from our initial populations (the subsets of the ELSA population that has complete data on appropriate risk factors). The first person to be included in our new boot strap data set is chosen at random from our starting (ELSA) dataset, with each person being equally likely to be chosen. Then the second person to be included in this boot strap data set is chosen at random in the same way, again with each person being equally likely to be chosen.

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.

More detail about the bootstrap methods and the Stata code we developed and used is included in Section 10.2 Calculating confidence intervals for prevalence estimates using bootstrap procedures, in Appendix 3: synthetic estimation using Stata software.

# 3 Results

## 3.1 Incidence and prevalence

New and prevalent hip OA and knee OA cases at each ELSA Wave are presented in Table 8.

**Table 8: ELSA incidence and prevalence at each Wave**

| Category | Wave 1 | Wave 2 | Wave 3 | Wave 4 | Wave 5 |
|---|---|---|---|---|---|
| New hip OA cases | 0 | 6 | 176 | 354 | 7 |
| Hip OA prevalence | 1,183 (8.01%) | 1,189 (8.05%) | 1,365 (9.24%) | 1,719 (11.64%) | 1,726 (11.69%) |
| New knee OA cases | 0 | 7 | 272 | 549 | 12 |
| Knee OA prevalence | 1,841 (12.46%) | 1,848 (12.51%) | 2,120 (14.35%) | 2,669 (18.07%) | 2,681 (18.15%) |

As noted previously, survey respondents who were identified as having hip or knee OA in a given wave, but did not report or identified as having hip or knee OA in the previous survey were classified as 'new' and then added to the cohort, and missing data for our outcome variables was replaced. The pattern of missing OA data varies by age, and only minor differences are observed between sex groups (**Table 9**). In males there is less missing data (26.67%) in the 45-64 years age group compared to 27.56% of missing data in the group of 65-74, while respondents over 75 years have 45.56% missing outcome data. Younger (45-64 age group) females have 29.69% missing OA outcome data compared to 21.53% in the 65-74 group and 48.14% of missing data for female respondents over 75 i.e. missing data increases as respondents age.

**Table 9: prevalence stratified by age and sex**

| | Female | | | Male | | | Both sexes | | |
|---|---|---|---|---|---|---|---|---|---|
| **Age group** | **45-64** | **65-74** | **75+** | **45-64** | **65-74** | **75+** | **45-64** | **65-74** | **+75** |
| Hip OA ELSA[5] | 478 (13.28%) | 326 (15.42%) | 320 (13.90%) | 236 (8.83%) | 208 (10.50%) | 154 (8.00%) | 714 (11.21%) | 534 (13.04) | 474 (11.22%) |
| Hip OA/ 10,000 ARUK **[67]** | 685 (7%) | 1,252 (13%) | 1,628 (16%) | 376 (4%) | 832 (8%) | 1,127 (11%) | NA | NA | NA |
| Knee OA ELSA[6] | 654 (18.17%) | 484 (22.89%) | 465 (20.20%) | 472 (17.05%) | 320 (16.15%) | 283 (14.71%) | 1,126 (17.68%) | 804 (19.63%) | 748 (17.70%) |
| Knee OA/10,000 ARUK **[67]** | 1,542 (15%) | 2,290 (23%) | 2,545 (25%) | 1,414 (14%) | 1,970 (20%) | 2,313 (23%) | NA | NA | NA |
| Missing OA data in ELSA[7] | 29.69% | 21.53% | 48.14% | 26.67% | 27.56% | 45.56% | 28.39% | 24.13% | 47.03% |

[5] Total number is 1,722 as 4 respondents do not have age record

[6] Total number is 2,678 as 3 respondents do not have age record

[7] The percentage shows the proportion of missing data in specific age and sex category compared to the total number of missing data in that sex group

## 3.2 Baseline characteristics

Table 10 shows baseline characteristics for cases and controls.

**Table 10: baseline characteristics for cases and controls**

| | Hip OA cases | Knee OA cases | Controls (definite) | Controls (short) | Controls (extended) |
|---|---|---|---|---|---|
| **Total number of respondents** | 1,726 | 2,681 | 11,323 | 3,259 | 19,406 |
| **Age** | | | | | |
| 45-50 | 42 | 64 | | 91 | 1,391 |
| 45-64 | 902 (52.47%) | 1,424 (53.11%) | 5,608 (49.53%) | 1,630 (50.02%) | 9,562 (49.27%) |
| 65-74 | 507 (29.49%) | 746 (27.83%) | 3,041 (26.86%) | 911 (27.95%) | 4,826 (24.87%) |
| Over 75 | 310 (18.03%) | 501 (18.69%) | 2,550 (22.52%) | 696 (21.36%) | 4,539 (23.39%) |
| Missing | 7 (0.41%) | 10 (0.37%) | 124 (1.10%) | 22 (0.68%) | 479 (2.47%) |
| **Gender** | | | | | |
| Female | 1,126 (65.24%) | 1,604 (59.83%) | 6,006 (53.04%) | 1,866 (57.26%) | 10,564 (54.44%) |
| Male | 600 (34.76%) | 1,077 (40.17%) | 5,317 (46.96%) | 1,393 (42.74%) | 8,842 (45.56%) |
| **Ethnicity** | | | | | |
| White | 1,670 (96.76%) | 2,574 (96.01%) | 10,919 (96.43%)) | 3,136 (96.23%) | 18,543 (95.55%) |
| Non-white | 56 (3.24%) | 107 (3.99%) | 393 (3.47%) | 123 (3.77%) | 804 (4.14%) |
| Missing | 0 | 0 | 11 (0.10%) | 0 | 59 (0.30%) |
| **Education** | | | | | |
| NVQ4/NVQ5/Degree or equivalent | 210 (12.20%) | 330 (12.31%) | 2,034 (17.96%) | 480 (14.73%) | 2,514 (12.95%) |
| Higher education below degree | 221 (12.84%) | 343 (12.79%) | 1,527 (13.49%) | 440 (13.50%) | 2,098 (10.81%) |
| NVQ3/GCE A level equivalent | 143 (8.31%) | 207 (7.72%) | 889 (7.85) | 258 (7.92%) | 1,296 (6.68%) |
| NVQ2/GCE O level equivalent | 319 (18.48%) | 465 (17.34%) | 2,082 (18.39%) | 633 (19.42%) | 3,114 (16.05%) |
| NVQ1/CSE other grade equivalent | 79 (4.58%) | 131 (4.89%) | 497 (4.39%) | 133 (4.08%) | 894 (4.61%) |
| Foreign/other | 129 (7.47%) | 214 (7.98%) | 801 (7.07%) | 255 (7.82%) | 1,368 (7.05%) |
| No qualification | 620 (35.92%) | 982 (36.63%) | 3,411 (30.12%) | 1,048 (32.16%) | 7,973 (41.09%) |
| Missing | 5 (0.29%) | 9 (0.34%) | 82 (0.72%) | 12 (0.37%) | 149 (0.77%) |
| **Socioeconomic status** | | | | | |
| Higher managerial and professional occ. | 114 (6.60%) | 169 (6.30%) | 1,170 (10.33%) | 266 (8.16%) | 1,298 (6.69%) |

|  | Hip OA cases | Knee OA cases | Controls (definite) | Controls (short) | Controls (extended) |
|---|---|---|---|---|---|
| Lower managerial and professional occ. | 342 (19.81%) | 549 (20.48%) | 2,549 (22.51%) | 689 (21.14%) | 2,932 (15.11%) |
| Intermediate occupations | 210 (12.17%) | 324 (12.09%) | 1,515 (13.38%) | 464 (14.24%) | 1,865 (9.61%) |
| Small employers and own account workers | 206 (11.94%) | 274 (10.22%) | 1,259 (11.12%) | 346 (10.62%) | 1,551 (7.99%) |
| Lower supervisory and technical occ. | 178 (10.31%) | 311 (11.60%) | 1,089 (9.62%) | 330 (10.13%) | 1,473 (7.59%) |
| Semi-routine occ. | 341 (19.76%) | 517 (19.28%) | 1,868 (16.50%) | 591 (18.13%) | 2,471 (12.73%) |
| Routine occ. | 281 (16.28%) | 459 (17.12%) | 1,523 (13.45%) | 497 (15.25%) | 2,252 (11.60%) |
| Never worked and long term unemployed | 22 (1.27%) | 36 (1.34%) | 112 (0.99%) | 32 (0.98%) | 229 (1.18%) |
| Other | 3 (0.17%) | 2 (0.07%) | 12 (0.11%) | 5 (0.15%) | 26 (0.13%) |
| Missing | 29 (1.68%) | 40 (1.49%) | 226 (2.00%) | 39 (1.20%) | 5,309 (27.36%) |
| **BMI** |  |  |  |  |  |
| <18.5 underweight | 9 (0.52%) | 15 (0.56%) | 158 (1.40%) | 37 (1.14%) | 428 (2.21%) |
| 18.5 – 24 normal | 302 (17.50%) | 396 (14.77%) | 2,888 (25.51%) | 796 (24.42%) | 4,721 (24.33%) |
| 25 – 29 overweight | 585 (33.89%) | 875 (32.64%) | 4,191 (37.01%) | 1,153 (35.38%) | 6,909 (35.60%) |
| 30 – 34 obese | 498 (28.85%) | 875 (32.64%) | 2,171 (19.17%) | 772 (23.69%) | 3,270 (16.85%) |
| 35 – 39 obese | 71 (4.11%) | 115 (4.19%) | 176 (1.55%) | 45 (1.38%) | 504 (2.60%) |
| >40 | 32 (1.85%) | 56 (2.09%) | 63 (0.56%) | 23 (0.71%) | 167 (0.86%) |
| Missing | 229 (13.27%) | 349 (13.02%) | 1,676 (14.80%) | 433 (13.29%) | 3,407 (17.56%) |
| **Physical activity (leisure)** |  |  |  |  |  |
| Sedentary | 184 (10.66%) | 292 (10.89%) | 726 (6.41%) | 305 (9.36%) | 1,109 (5.71%) |
| Low | 593 (34.36%) | 885 (33.01%) | 2,096 (18.51%) | 933 (28.63%) | 2,281 (11.75%) |
| Moderate | 735 (42.58%) | 1,171 (43.68%) | 5,630 (49.72%) | 1,463 (44.89%) | 9,437 (48.63%) |
| High | 214 (12.40%) | 333 (12.42%) | 2,144 (18.93%) | 558 (17.12%) | 2,525 (13.01%) |
| Missing | 0 | 0 | 727 (6.42%) | 0 | 4,054 (20.89%) |
| **Smoking** |  |  |  |  |  |
| Current smoker | 225 (13.04%) | 289 (10.78%) | 1,389 (12.27%) | 476 (14.61%) | 1,407 (7.25%) |

| | Hip OA cases | Knee OA cases | Controls (definite) | Controls (short) | Controls (extended) |
|---|---|---|---|---|---|
| Ex-smoker | 938 (54.35%) | 1,445 (53.90%) | 5,607 (49.52%) | 1,628 (49.95%) | 5,670 (29.22%) |
| Never smoked | 562 (32.56%) | 947 (35.32%) | 4,323 (38.18%) | 1,153 (35.38%) | 4,226 (21.78%) |
| Missing | 1 (0.06%) | 0 | 4 (0.04%) | 2 (0.06%) | 8,103 (41.76%) |
| **Physical activity (work)[8]** | | | | | |
| Sedentary occupation | 649 (37.60%) | 1,007 (37.56%) | 4,817 (42.54%) | 1,310 (40.20%) | 8,609 (44.36%) |
| Standing occupation | 245 (14.19%) | 407 (15.18%) | 1,932 (17.06%) | 527 (16.17%) | 2,519 (12.98%) |
| Physical work | 193 (11.18%) | 326 (12.16%) | 1,527 (13.49%) | 437 (13.41%) | 2,029 (10.46%) |
| Heavy manual work | 27 (1.56%) | 58 (2.16%) | 236 (2.08%) | 74 (2.27%) | 236 (1.22%) |
| Missing | 612 (35.46%) | 883 (32.94%) | 2,811 (24.83%) | 911 (27.95%) | 6,013 (30.99%) |
| **Member at sports clubs, gym etc.** | | | | | |
| No | 1,394 (80.76%) | 2,182 (81.39%) | 8,307 (73.36%) | 2,570 (78.86%) | 9,296 (47.90%) |
| Yes | 277 (16.05%) | 410 (15.29%) | 2,298 (20.29%) | 561 (17.21%) | 2,344 (12.08%) |
| Missing | 55 (3.19%) | 89 (3.32%) | 718 (6.34%) | 128 (3.93%) | 7,766 (40.02%) |
| **Housework/gardening activity level[7]** | | | | | |
| Inactive | 189 (10.95%) | 305 (11.38%) | 1,023 (9.03%) | 305 (9.36%) | 3,057 (15.75%) |
| Light (some non-heavy activity) | 80 (4.63%) | 136 (5.07%) | 704 (6.22%) | 178 (5.46%) | 1,268 (6.53%) |
| Moderate (heavy activity) | 422 (24.45%) | 660 (24.62%) | 3,117 (27.53%) | 842 (25.84%) | 5,595 (28.83%) |
| Missing | 1,035 (59.97%) | 1,580 (58.93%) | 6,479 (57.22%) | 1,934 (59.34%) | 9,487 (48.89%) |
| **Lifting at work[7]** | | | | | |
| Lifting heavy loads | 20 (1.16%) | 29 (1.08%) | 118 (1.04%) | 38 (1.17%) | 214 (1.10%) |
| Lifting and carrying heavy loads | 72 (4.17%) | 138 (5.15%) | 512 (4.52%) | 152 (4.46%) | 881 (4.54%) |
| Not lifting | 116 (6.72%) | 172 (6.42%) | 1,014 (8.96%) | 226 (6.93%) | 1,642 (8.46%) |
| Missing | 1,518 (87.95%) | 2,342 (87.36%) | 9,679 (85.48%) | 2,843 (87.24%) | 16,669 (85.90%) |

[8] Physical activity types other than leisure were dropped from the final models but is shown here for comparison purposes.

## 3.3 Severe hip OA

### 3.3.1 Univariate and multivariate analyses for severe hip OA

**Table 11** shows univariate and multivariate logistic regression results for individual risk factors to predict severe hip OA cases. All logistic regression analyses for no hip OA versus severe hip OA were weighted using England's population distribution by age and sex. Comparing the unadjusted results, we can conclude that age, sex, education, socioeconomic class and obesity (when BMI >30) are significant risk factors for our sample population. Physical leisure activity and membership at sports club/gym are protective factors. In the multivariate analysis, however, education, SeC, obesity and gym membership were not significant risk factors. The R squared at 9% is low, suggesting that other relevant risk factors have not been captured.

**Table 11 Univariate and multivariate analyses results for severe hip OA**

| | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | OR | 95%CI | p-value | OR | 95%CI | p-value |
| **Age** | | | | | | |
| 45-64 (Reference) | 1 | | | 1 | | |
| 65-74 | 2.44 | [2.02-2.95] | <0.001 | 1.91 | [1.53-2.38] | <001 |
| Over 75 | 3.41 | [2.79-4.17] | <0.001 | 2.00 | [1.54-2.59] | <001 |
| **Sex** | | | | | | |
| Male (Reference) | 1 | | | 1 | | |
| Female | 1.58 | [1.34-1.87] | <0.001 | 1.42 | [1.15-1.75] | 0.001 |
| **Ethnicity** | | | | | | |
| White (Reference) | 1 | | | 1 | | |
| Non-white | 1.05 | [0.7-1.58] | 0.807 | 1.15 | [0.66-1.98] | 0.623 |
| **Education** | | | | | | |
| NVQ4/NVQ5/Degree or equivalent (Reference) | 1 | | | 1 | | |
| Higher education below degree | 1.62 | [1.12-2.34] | 0.010 | 1.11 | [0.74-1.67] | 0.609 |
| NVQ3/GCE A Level equivalent | 1.43 | [0.92-2.2] | 0.108 | 1.02 | [0.63-1.65] | 0.937 |
| NVQ2/GCE O Level equivalent | 2.11 | [1.52-2.95] | <0.001 | 1.33 | [0.91-1.94] | 0.140 |
| NVQ1/CSE other grade equivalent | 3.00 | [1.91-4.72] | <0.001 | 1.53 | [0.9-2.61] | 0.119 |
| Foreign/other | 2.70 | [1.81-4.03] | <0.001 | 1.27 | [0.8-2.02] | 0.307 |
| No qualification | 3.70 | [2.74-5.01] | <0.001 | 1.41 | [0.95-2.09] | 0.085 |
| **Socioeconomic class** | | | | | | |
| Higher managerial and professional occupations (Reference) | 1 | | | 1 | | |
| Lower managerial and professional occupations | 1.49 | [0.99-2.24] | 0.057 | 1.13 | [0.72-1.76] | 0.593 |
| Intermediate occupations | 2.05 | [1.34-3.15] | 0.001 | 0.99 | [0.61-1.63] | 0.981 |
| Small employers and own account workers | 2.04 | [1.33-3.15] | 0.001 | 1.41 | [0.87-2.28] | 0.160 |
| Lower supervisory and technical occupations | 2.41 | [1.56-3.72] | <0.001 | 1.40 | [0.85-2.3] | 0.185 |
| Semi-routine occupations | 2.49 | [1.66-3.72] | <0.001 | 1.15 | [0.72-1.85] | 0.556 |
| Routine occupations | 2.88 | [1.92-4.32] | <0.001 | 1.16 | [0.71-1.9] | 0.546 |
| Never worked and long term unemployed | 3.70 | [1.77-7.77] | 0.001 | 0.98 | [0.39-2.49] | 0.967 |

| | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | OR | 95%CI | p-value | OR | 95%CI | p-value |
| **BMI**[9] | | | | | | |
| Underweight <18.4 (Reference) | 1 | | | 1 | | |
| Normal weight 18.5-24 | 2.26 | [0.55-9.23] | 0.258 | 2.49 | [0.62-10.06] | 0.201 |
| Overweight 25-29 | 2.87 | [0.71-11.66] | 0.141 | 3.15 | [0.79-12.62] | 0.105 |
| Obese >30 | 5.07 | [1.25-20.58] | 0.023 | 4.54 | [1.14-18.12] | 0.032 |
| **Physical activity** | | | | | | |
| Sedentary (Reference) | 1 | | | 1 | | |
| Low | 0.63 | [0.5-0.8] | <0.001 | 0.73 | [0.56-0.97] | 0.028 |
| Moderate | 0.18 | [0.14-0.23] | <0.001 | 0.27 | [0.2-0.37] | <001 |
| High | 0.08 | [0.06-0.12] | <0.001 | 0.16 | [0.1-0.25] | <001 |
| **Smoking status** | | | | | | |
| Current smoker (Reference) | 1 | | | 1 | | |
| Ex-smoker | 1.16 | [0.91-1.5] | 0.234 | 1.28 | [0.96-1.71] | 0.097 |
| Never smoked | 0.90 | [0.69-1.17] | 0.439 | 1.05 | [0.77-1.42] | 0.767 |
| **Membership at gym** | | | | | | |
| **No** | 1 | | | 1 | | |
| **Yes** | 0.41 | [0.32-0.54] | <0.001 | 0.77 | [0.57-1.04] | 0.087 |
| **Pseudo R Squared**[10] **0.0952** | | | | | | |

### 3.3.2 Variable selection models for multivariate analyses for severe hip OA

This section shows how models were compared to arrive at the final multivariable model in Table 12. Backward and forward variable selection models were fitted using the stepwise function in Stata, with forward selection set as pr(.05) and backward – pr(.05) separately). Automatic stepwise backward model was chosen as the final method and its selected variables for hip OA are shown in Table 12. All logistic regression analyses for (no hip OA versus severe hip OA were weighted using England's population distribution by age and sex).

**Table 12: final severe hip OA model**

| | Auto stepwise backward (logistic) | | | **Auto stepwise backward (logit)** | | |
|---|---|---|---|---|---|---|
| | OR | 95% CI | p-value | Coefficient | 95% CI | p-value |
| **Age** | | | | | | |
| 45-64 (Reference) | 1 | | | 0 | | |
| 65-74 | 2.01 | [1.62-2.5] | <0.001 | 0.70 | [0.48-0.92] | <0.001 |
| Over 75 | 2.12 | [1.66-2.71] | <0.001 | 0.75 | [0.51-1] | <0.001 |
| **Sex** | | | | | | |
| Male (Reference) | 1 | | | 0 | | |
| Female | 1.39 | [1.15-1.68] | 0.001 | 0.33 | [0.14-0.52] | 0.001 |
| **BMI** | | | | | | |

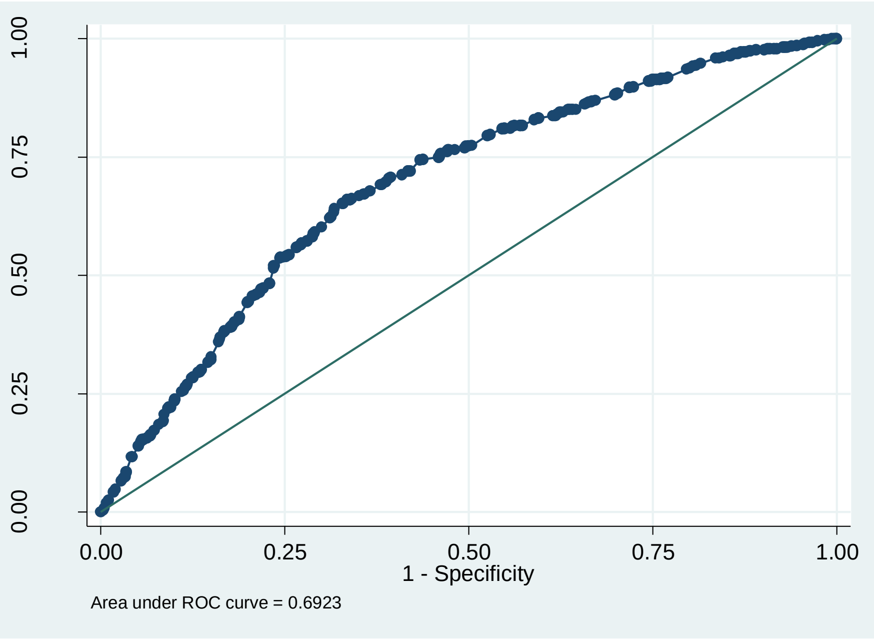
[9] BMI grouping was changed to match the format found in Active People Survey (three 'obese' categories were merged into one 'obese' if BMI was more than 30.

[10] R squared measures the proportion of variance explained by the model. This figure is relatively low, suggesting that many risk factors have not been included

| | Auto stepwise backward (logistic) | | | **Auto stepwise backward (logit)** | | |
|---|---|---|---|---|---|---|
| | OR | 95% CI | p-value | Coefficient | 95% CI | p-value |
| Underweight <18.4 (Reference) | 1 | | | 0 | | |
| Normal weight 18.5-24 | 1 | | | 0 | | |
| Overweight 25-29 | 1.33 | [1.03-1.7] | 0.028 | 0.28 | [0.03-0.53] | 0.028 |
| Obese >30 | 1.94 | [1.52-2.49] | <0.001 | 0.66 | [0.42-0.91] | <0.001 |
| **Physical activity** | | | | | | |
| Sedentary (Reference) | 1 | | | 0 | | |
| Low | 0.72 | [0.55-0.95] | 0.019 | -0.33 | [-0.6--0.05] | 0.019 |
| Moderate | 0.26 | [0.2-0.35] | <0.001 | -1.34 | [-1.63--1.05] | <0.001 |
| High | 0.15 | [0.1-0.24] | <0.001 | -1.89 | [-2.35--1.44] | <0.001 |
| **Smoking status** | | | | | | |
| Current smoker (Reference) | 1 | | | 0 | | |
| Ex-smoker | 1.24 | [1.03-1.49] | 0.023 | 0.21 | [0.03-0.4] | 0.023 |
| Never smoked | 1 | | | 0 | | |
| **Membership at gym** | | | | | | |
| No (Reference) | 1 | | | 0 | | |
| Yes | 0.72 | [0.54-0.98] | 0.034 | -0.32 | [-0.62--0.02] | 0.034 |
| Constant | NA | NA | NA | -3.42 | [-3.81--3.04] | <0.001 |
| **Pseudo R2 = 0.0911** | | | | | | |

This model predicts reasonably well as area under the ROC curve is 0.69±0.01 (95% CI 0.67-0.71) (See Figure 3).

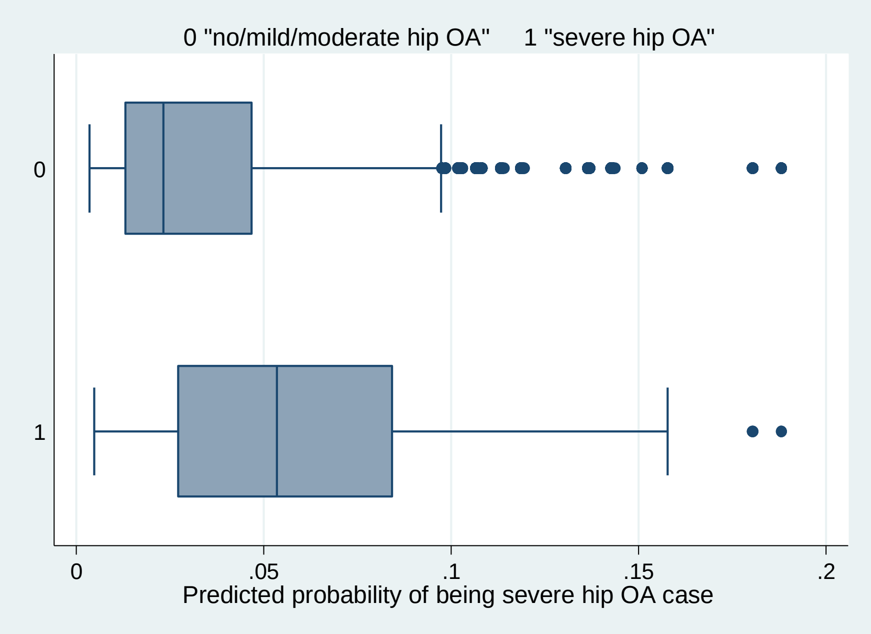
**Figure 3 ROC curve for final severe hip OA model**



### 3.3.3 Internal validation: non/mild/moderate hip OA versus severe hip OA

An ELSA manual backward model was used to predict the probability of an individual being a severe hip OA case in ELSA data set (no/mild/moderate hip OA compared to severe hip OA). In Figure 4 the two box plots show the predicted probability of people with severe hip OA caseness among the non/mild/moderate hip OA and severe hip OA groups. We can see that informants with severe hip OA have higher predicted probability of severe caseness than those with non/mild/moderate hip OA. Since we have a binary response model, we can choose a cut-off point on the predicted probability to separate the predicted severe hip OA cases (with higher predicted probability) from the predicted non/mild/moderate hip OA cases (with lower predicted probability). We can tell from the box plots that no matter which cut-off point we choose, there will always be mis-classified people. Either the non/mild/moderate hip OA people being classified as predicted severe hip OA cases, or severe hip OA people being classified as predicted non/mild/moderate hip OA cases. Therefore, we use sensitivity and specificity plots to help with this decision.

**Figure 4: severe hip OA predicted probability**

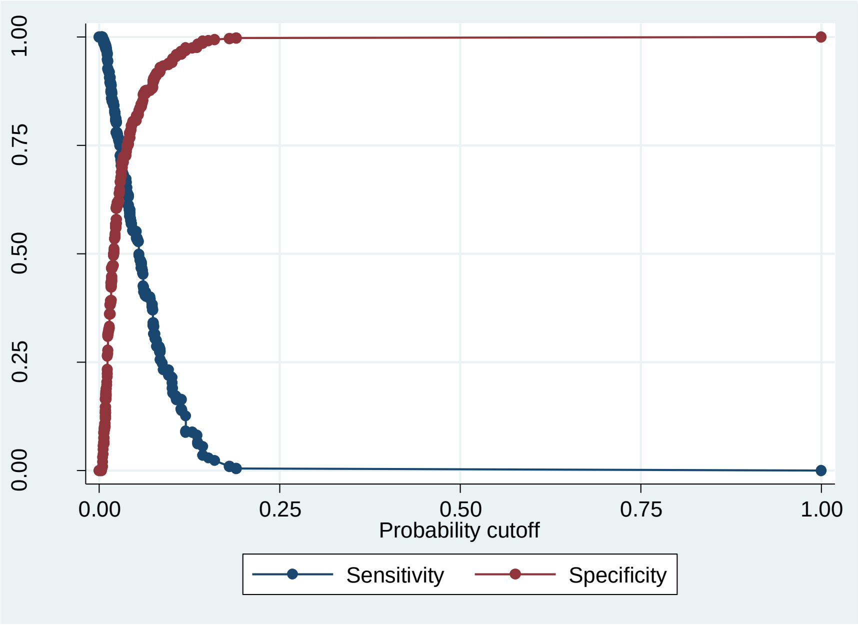


The sensitivity/specificity versus probability cut-off plot shows us the corresponding sensitivity and specificity in each possible probability cut-off point (See **Figure 5**). Higher sensitivity would usually yield low specificity and vice versa, the rule of thumb is to choose a cut-off probability to maximize both. If the cut-off probability is chosen where sensitivity and specificity lines cross - at 0.03, the sensitivity and specificity both reach around 72% and 68%, respectively. Applying this cut-off probability to our data, the following table shows the comparison between predicted and true cases of severe hip OA in ELSA **(Table 13).** Confirming 72% (449 out of 626) correctly classified as severe hip OA cases, and 68% (9,549 out of 14,144) correctly classified non/mild/moderate hip OA.

**Table 13 Predicted severe hip OA caseness comparison with different cut-off probability**

| Probability cut-off | 0 | 0.025 | 0.03 | 0.035 | 0.04 | 0.05 | 0.06 | 0.1 | 0.25 |
|---|---|---|---|---|---|---|---|---|---|
| Sensitivity (%) | 100 | 77.14 | 71.70 | 67.19 | 63.51 | 55.16 | 46.17 | 21.37 | 0 |
| Specificity (%) | 0 | 61.79 | 67.51 | 72.44 | 75.09 | 80.80 | 84.83 | 94.09 | 100 |
| True positive | 626 | 483 | 449 | 421 | 398 | 345 | 289 | 134 | 0 |
| False positive | 14,144 | 5,404 | 4,595 | 3,898 | 3,523 | 2,716 | 2,146 | 836 | 0 |
| True negative | 0 | 8,740 | 9,549 | 10,246 | 10,621 | 11,428 | 11,998 | 13,308 | 14,144 |
| False negative | 0 | 143 | 177 | 205 | 228 | 281 | 337 | 492 | 626 |

**Figure 5 Sensitivity/Specificity vs. Probability Cut-Off**



## 3.4 Severe knee OA

### 3.4.1 Univariate and multivariate analyses for severe knee OA

Table 14 shows univariate and multivariate logistic regression results for severe knee OA cases. All logistic regression analyses for no knee OA versus severe knee OA were weighted using England's population distribution by age and sex. Comparing the unadjusted results, we can conclude that age, sex, ethnicity, education, socioeconomic class, obesity (>30 BMI) and ex-smoking status are significant risk factors for our sample population. In contrast, physical leisure and work activity, and membership at sports club/gym are protective factors.  Higher risk for severe OA was also associated with age, socioeconomic status (lower supervisory/technical, semi-routine and routine occupations) and obesity. Significant protective factors were the same in both univariate and multivariate analyses.

**Table 14: Univariate and multivariate analyses results for severe knee OA**

| | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | OR | 95%CI | p-value | OR | 95%CI | p-value |
| **Age** | | | | | | |
| 45-64 (Reference) | 1 | | | 1 | | |
| 65-74 | 1.93 | [1.65-2.27] | <0.001 | 1.45 | [1.2-1.75] | <0.001 |
| Over 75 | 2.28 | [1.93-2.71] | <0.001 | 1.34 | [1.08-1.66] | 0.009 |
| **Gender** | | | | | | |
| Male (Reference) | 1 | | | 1 | | |
| Female | 1.20 | [1.05-1.38] | 0.009 | 1.06 | [0.89-1.26] | 0.526 |
| **Ethnicity** | | | | | | |
| White (Reference) | 1 | | | 1 | | |
| Non-white | 1.54 | [1.13-2.1] | 0.007 | 1.43 | [0.93-2.19] | 0.102 |
| **Education** | | | | | | |
| NVQ4/NVQ5/Degree or equivalent (Reference) | 1 | | | 1 | | |
| Higher education below degree | 1.54 | [1.13-2.11] | 0.006 | 1.10 | [0.77-1.58] | 0.591 |
| NVQ3/GCE A Level equivalent | 1.50 | [1.05-2.14] | 0.025 | 0.96 | [0.63-1.46] | 0.847 |
| NVQ2/GCE O Level equivalent | 1.84 | [1.38-2.43] | <0.001 | 1.17 | [0.83-1.65] | 0.37 |
| NVQ1/CSE other grade equivalent | 2.49 | [1.68-3.71] | <0.001 | 1.17 | [0.72-1.89] | 0.526 |
| Foreign/other | 2.42 | [1.73-3.39] | <0.001 | 1.22 | [0.81-1.84] | 0.344 |
| No qualification | 3.65 | [2.85-4.67] | <0.001 | 1.53 | [1.08-2.15] | 0.016 |
| **Socioeconomic class** | | | | | | |
| Higher managerial and professional occupations (Reference) | 1 | | | 1 | | |
| Lower managerial and professional occupations | 1.48 | [1.04-2.1] | 0.030 | 1.20 | [0.81-1.78] | 0.352 |
| Intermediate occupations | 1.76 | [1.22-2.55] | 0.003 | 1 | [0.64-1.56] | 0.99 |
| Small employers and own account workers | 1.84 | [1.26-2.69] | 0.002 | 1.34 | [0.87-2.08] | 0.187 |
| Lower supervisory and technical occupations | 2.80 | [1.95-4.04] | <0.001 | 1.67 | [1.08-2.59] | 0.022 |
| Semi-routine occupations | 2.74 | [1.94-3.86] | <0.001 | 1.62 | [1.06-2.47] | 0.026 |
| Routine occupations | 3.37 | [2.39-4.77] | <0.001 | 1.48 | [0.96-2.29] | 0.079 |
| Never worked and long term unemployed | 4.79 | [2.67-8.62] | <0.001 | 1.47 | [0.66-3.27] | 0.349 |
| **BMI**[11] | | | | | | |
| Underweight <18.4 (Reference) | 1 | | | 1 | | |
| Normal weight 18.5-24 | 0.99 | [0.42-2.32] | 0.986 | 1.07 | [0.46-2.48] | 0.88 |
| Overweight 25-29 | 1.54 | [0.67-3.57] | 0.309 | 1.60 | [0.7-3.65] | 0.269 |
| Obese >30 | 3.27 | [1.42-7.53] | 0.005 | 2.80 | [1.23-6.39] | 0.014 |
| **Physical activity** | | | | | | |
| Sedentary (Reference) | 1 | | | 1 | | |
| Low | 0.63 | [0.52-0.77] | <0.001 | 0.70 | [0.55-0.89] | 0.004 |

[11] BMI grouping was changed to match the format found in Active People Survey (three 'obese' categories were merged into one 'obese' if BMI was more than 30.

|  | Univariate analysis |  |  | Multivariate analysis |  |  |
|---|---|---|---|---|---|---|
|  | OR | 95%CI | p-value | OR | 95%CI | p-value |
| Moderate | 0.19 | [0.15-0.23] | <0.001 | 0.26 | [0.2-0.33] | <0.001 |
| High | 0.09 | [0.06-0.12] | <0.001 | 0.15 | [0.1-0.22] | <0.001 |
| **Smoking status** |  |  |  |  |  |  |
| Current smoker (Reference) | 1 |  |  | 1 |  |  |
| Ex-smoker | 1.27 | [1.01-1.58] | 0.037 | 1.43 | [1.11-1.85] | 0.005 |
| Never smoked | 0.93 | [0.73-1.17] | 0.516 | 1.12 | [0.86-1.47] | 0.398 |
| **Membership at gym** |  |  |  |  |  |  |
| No (Reference) | 1 |  |  | 1 |  |  |
| Yes | 0.42 | [0.34-0.52] | <0.001 | 0.77 | [0.6-0.99] | 0.045 |
| **Pseudo R2 =** 0.1116 |  |  |  |  |  |  |

### 3.4.2 Variable selection models for multivariate analyses for severe knee OA

Backward and forward selection models were fitted (using the stepwise function in Stata, with forward selection set as pe(.05) and backward – pr(.05) separately). The automatic stepwise backward model was chosen as the final one and its selected variables for knee OA are shown in Table 15**.** All logistic regression analyses for (no knee OA versus severe knee OA were weighted using England's population distribution by age and sex).

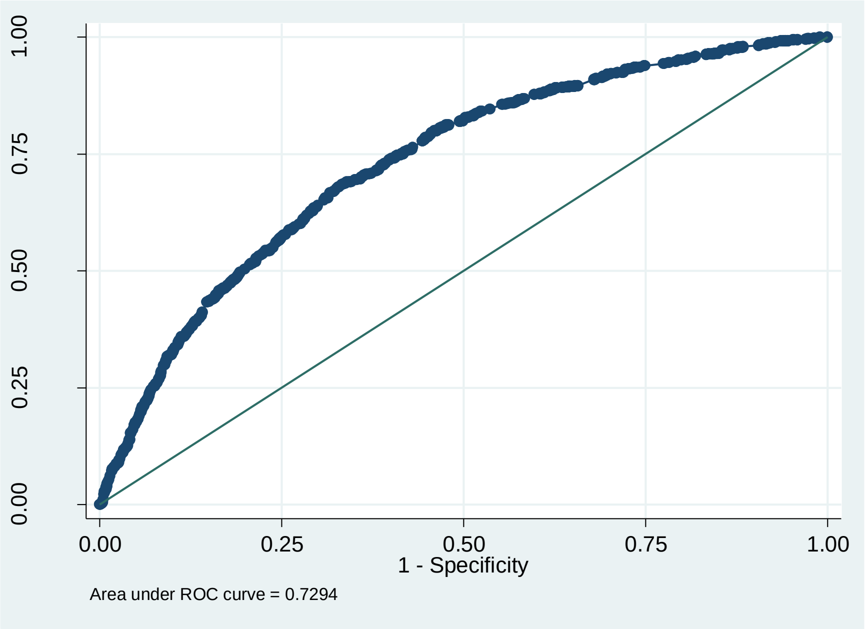
**Table 15: final severe knee OA model**

|  | Auto stepwise backward (logistic) |  |  | Auto stepwise backward (logit) |  |  |
|---|---|---|---|---|---|---|
|  | OR | 95% CI | p-value | Coefficient | 95% CI | p-value |
| **Age** |  |  |  |  |  |  |
| 45-64 (Reference) | 1 |  |  | 0 |  |  |
| 65-74 | 1.47 | [1.22-1.77] | <0.001 | 0.38 | [0.2-0.57] | <0.001 |
| Over 75 | 1.36 | [1.1-1.68] | 0.004 | 0.31 | [0.1-0.52] | 0.004 |
| **Education** |  |  |  |  |  |  |
| NVQ4/NVQ5/Degree or equivalent (Reference) | 1 |  |  | 0 |  |  |
| Higher education below degree | 1 |  |  | 0 |  |  |
| NVQ3/GCE A Level equivalent | 1 |  |  | 0 |  |  |
| NVQ2/GCE O Level equivalent | 1 |  |  | 0 |  |  |
| NVQ1/CSE other grade equivalent | 1 |  |  | 0 |  |  |
| Foreign/other | 1 |  |  | 0 |  |  |
| No qualification | 1.38 | [1.15-1.66] | 0.001 | 0.32 | [0.14-0.5] | 0.001 |
| **Socioeconomic class** |  |  |  |  |  |  |
| Higher managerial and professional occupations (Reference) | 1 |  |  | 0 |  |  |
| Lower managerial and professional occupations | 1 |  |  | 0 |  |  |
| Intermediate occupations | 1 |  |  | 0 |  |  |
| Small employers and own account workers | 1 |  |  | 0 |  |  |

|  | Auto stepwise backward (logistic) |  |  | Auto stepwise backward (logit) |  |  |
|---|---|---|---|---|---|---|
|  | OR | 95% CI | p-value | Coefficient | 95% CI | p-value |
| Lower supervisory and technical occupations | 1.44 | [1.12-1.85] | 0.004 | 0.37 | [0.11-0.62] | 0.004 |
| Semi-routine occupations | 1.43 | [1.16-1.77] | 0.001 | 0.36 | [0.15-0.57] | 0.001 |
| Routine occupations | 1.30 | [1.02-1.65] | 0.031 | 0.26 | [0.02-0.5] | 0.031 |
| Never worked and long term unemployed | 1 |  |  | 0 |  |  |
| **BMI** |  |  |  |  |  |  |
| Underweight <18.4 (Reference) | 1 |  |  | 0 |  |  |
| Normal weight 18.5-24 | 1 |  |  | 0 |  |  |
| Overweight 25-29 | 1.51 | [1.2-1.89] | <0.001 | 0.41 | [0.18-0.64] | <0.001 |
| Obese >30 | 2.66 | [2.13-3.32] | <0.001 | 0.98 | [0.76-1.2] | <0.001 |
| **Physical activity** |  |  |  |  |  |  |
| Sedentary (Reference) | 1 |  |  | 0 |  |  |
| Low | 0.70 | [0.55-0.88] | 0.003 | -0.36 | [-0.6--0.12] | 0.003 |
| Moderate | 0.25 | [0.2-0.32] | <0.001 | -1.38 | [-1.63--1.13] | <0.001 |
| High | 0.14 | [0.1-0.21] | <0.001 | -1.93 | [-2.31--1.55] | <0.001 |
| **Smoking status** |  |  |  |  |  |  |
| Current smoker (Reference) | 1 |  |  | 0 |  |  |
| Ex-smoker | 1.30 | [1.11-1.52] | 0.001 | 0.26 | [0.11-0.42] | 0.001 |
| Never smoked | 1 |  |  | 0 |  |  |
| **Membership at gym** |  |  |  |  |  |  |
| No (Reference) | 1 |  |  | 0 |  |  |
| Yes | 0.77 | [0.6-0.99] | 0.039 | -0.26 | [-0.51--0.01] | 0.039 |
| Constant | NA | NA |  | -2.77 | [-3.09--2.44] | <0.001 |
| **Pseudo R2 = .1097** |  |  |  |  |  |  |

The model predicts moderately well as the area under the ROC curve is 0.73±0.01 (95% CI 0.71-0.75) (See Figure 6).

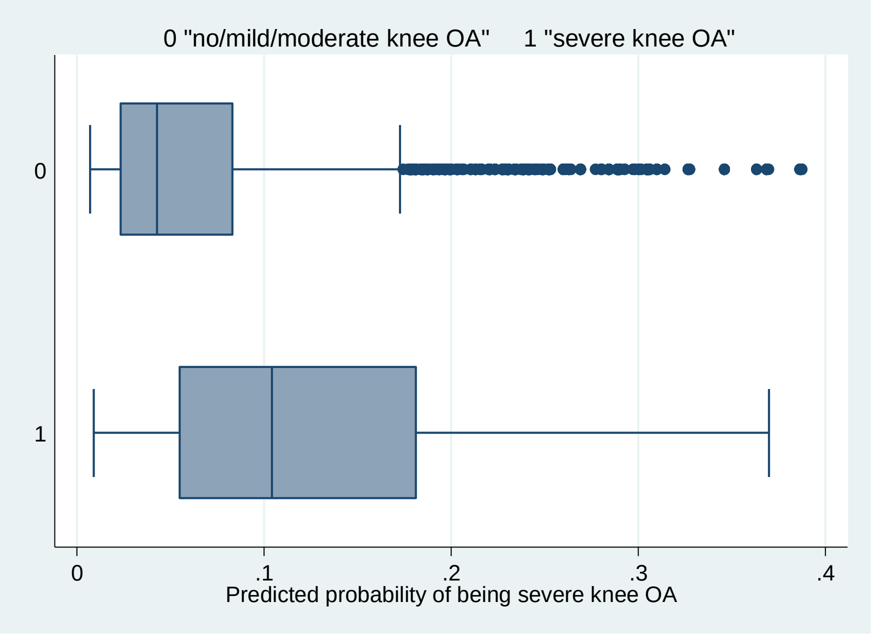
**Figure 6: ROC curve for final severe knee OA model**



### 3.4.3 Non/mild/moderate knee OA versus severe knee OA

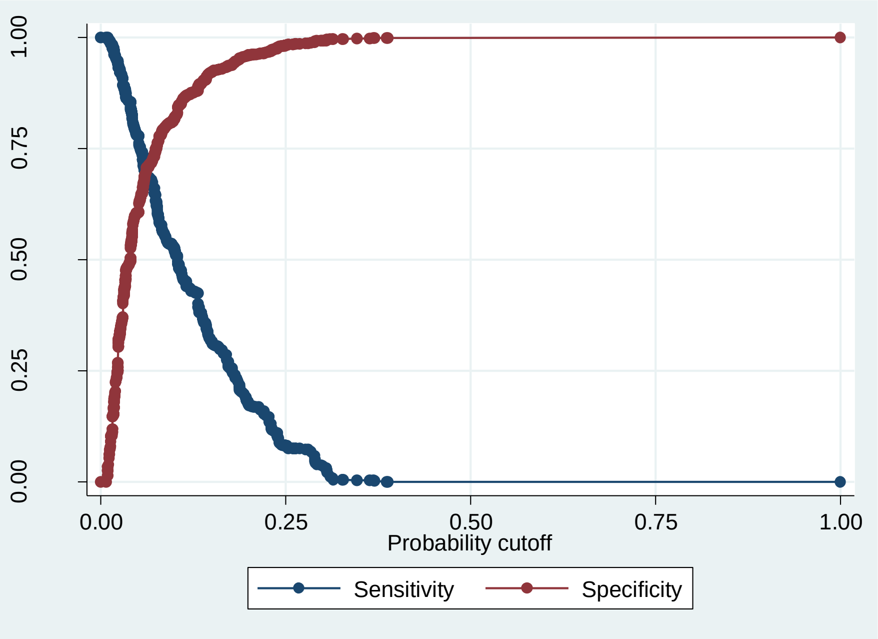
In **Figure 7** the two box plots show the predicted probability of people severe knee OA caseness among the non/mild/moderate knee OA and severe knee OA groups. We can see that people with severe knee OA have higher predicted probability than the non/mild/moderate knee OA people. Since we have a binary response model, we can choose a cut-off point on the predicted probability to separate the predicted severe knee OA cases (with higher predicted probability) from the predicted non/mild/moderate knee OA 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. Either the non/mild/moderate knee OA people being classified as predicted severe knee OA cases, or the severe knee OA people being classified as predicted non/mild/moderate knee OA cases. Therefore, we use sensitivity and specificity plots to help with this decision.

**Figure 7: severe knee OA predicted probability**



The sensitivity/specificity versus probability cut-off plot shows the corresponding sensitivity and specificity in each possible probability cut-off point (See Figure 8). Higher sensitivity would usually yield low specificity and vice versa. A 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. At a cut-off point of predicted probability 0.06, the sensitivity and specificity both reach 69.94% and 68.94%, respectively.

**Figure 8: sensitivity/specificity vs. probability cut-off**



Applying this cut-off probability to our data, the following table shows the comparison between predicted and true cases of severe knee OA in ELSA (Table 16). Confirming 70% (633) correctly classified severe knee OA cases, and 69% (9,556) correctly classified non/mild/moderate knee OA.

**Table 16 Predicted severe knee OA caseness with different cut-off probabilities**

| Probability cut-off | 0 | 0.025 | 0.04 | 0.045 | 0.05 | 0.06 | 0.07 | 0.1 | 0.15 | 0.25 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensitivity (%) | 100 | 92.92 | 85.26 | 80.18 | 77.87 | 69.94 | 67.41 | 51.89 | 31.54 | 8.11 | 0 |
| Specificity (%) | 0 | 32.57 | 49.74 | 58.71 | 60.64 | 68.94 | 72.22 | 82.07 | 92.16 | 98.32 | 100 |
| True positive | 905 | 841 | 772 | 726 | 705 | 633 | 610 | 470 | 285 | 73 | 0 |
| False positive | 13,864 | 9,348 | 6,968 | 8,140 | 8,407 | 4,308 | 3,851 | 2,486 | 1,087 | 233 | 0 |
| True negative | 0 | 4,516 | 6,896 | 5,724 | 5,457 | 9,556 | 10,013 | 11,378 | 12,777 | 13,631 | 13,864 |
| False negative | 0 | 64 | 133 | 179 | 200 | 272 | 295 | 435 | 620 | 832 | 905 |

## 3.5 Total hip OA

### 3.5.1 Univariate and multivariate analyses for total hip OA

Table 17 shows univariate and multivariate analyses for total hip OA.

**Table 17: Univariate and multivariate analyses for total hip OA**

|  | Univariate analysis |  |  | Multivariate analysis |  |  |
|---|---|---|---|---|---|---|
|  | OR | 95%CI | p-value | OR | 95%CI | p-value |
| **Age** |  |  |  |  |  |  |
| 45-64 (Reference) | 1 |  |  | 1 |  |  |
| 65-74 | 1.26 | [1.12-1.42] | <0.001 | 1.08 | [0.94-1.23] | 0.288 |
| Over 75 | 1.10 | [0.96-1.26] | 0.175 | 0.85 | [0.72-1] | 0.053 |
| **Sex** |  |  |  |  |  |  |
| Male (Reference) | 1 |  |  | 1 |  |  |
| Female | 1.68 | [1.5-1.88] | <0.001 | 1.82 | [1.59-2.09] | <0.001 |
| **Ethnicity** |  |  |  |  |  |  |
| White (Reference) | 1 |  |  | 1 |  |  |
| Non-white | 0.77 | [0.57-1.04] | 0.088 | 0.82 | [0.55-1.21] | 0.316 |
| **Education** |  |  |  |  |  |  |
| NVQ4/NVQ5/Degree or equivalent (Reference) | 1 |  |  | 1 |  |  |
| Higher education below degree | 1.34 | [1.09-1.66] | 0.006 | 1.11 | [0.87-1.41] | 0.395 |
| NVQ3/GCE A Level equivalent | 1.55 | [1.22-1.96] | <0.001 | 1.25 | [0.95-1.65] | 0.11 |
| NVQ2/GCE O Level equivalent | 1.42 | [1.17-1.72] | <0.001 | 1.07 | [0.85-1.34] | 0.581 |
| NVQ1/CSE other grade equivalent | 1.58 | [1.18-2.12] | 0.002 | 1.38 | [0.99-1.93] | 0.056 |
| Foreign/other | 1.52 | [1.19-1.94] | 0.001 | 1.11 | [0.84-1.47] | 0.471 |
| No qualification | 1.77 | [1.48-2.1] | <0.001 | 1.20 | [0.95-1.51] | 0.12 |
| **Socioeconomic class** |  |  |  |  |  |  |
| Higher managerial and professional occupations (Reference) | 1 |  |  | 1 |  |  |
| Lower managerial and professional occupations | 1.33 | [1.05-1.69] | 0.019 | 1.07 | [0.82-1.4] | 0.617 |
| Intermediate occupations | 1.39 | [1.08-1.79] | 0.011 | 0.87 | [0.65-1.18] | 0.377 |
| Small employers and own account workers | 1.61 | [1.24-2.08] | <0.001 | 1.34 | [1-1.79] | 0.05 |
| Lower supervisory and technical occupations | 1.56 | [1.2-2.04] | 0.001 | 1.19 | [0.88-1.62] | 0.264 |
| Semi-routine occupations | 1.83 | [1.44-2.32] | <0.001 | 1.16 | [0.87-1.55] | 0.309 |
| Routine occupations | 1.79 | [1.4-2.28] | <0.001 | 1.16 | [0.86-1.56] | 0.324 |
| Never worked and long term unemployed | 1.83 | [1.1-3.05] | 0.021 | 1.05 | [0.57-1.91] | 0.882 |
| **BMI**[12] |  |  |  |  |  |  |
| Underweight <18.4 (Reference) | 1 |  |  | 1 |  |  |
| Normal weight 18.5-24 | 1.65 | [0.81-3.37] | 0.172 | 1.62 | [0.79-3.35] | 0.19 |
| Overweight 25-29 | 2.09 | [1.03-4.25] | 0.042 | 2.18 | [1.06-4.48] | 0.034 |
| Obese >30 | 3.27 | [1.61-6.65] | 0.001 | 2.91 | [1.41-5.99] | 0.004 |
| **Physical activity**[13] |  |  |  |  |  |  |
| Sedentary (Reference) | 1 |  |  | 1 |  |  |
| Low | 1.03 | [0.85-1.24] | 0.782 | 0.97 | [0.78-1.22] | 0.823 |

[12] BMI grouping was changed to match the format found in Active People Survey (three ‘obese’ categories were merged into one ‘obese’ if BMI was more than 30.

[13] Note that work physical activity has been excluded because of lack of data for LAs

| | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | OR | 95%CI | p-value | OR | 95%CI | p-value |
| Moderate | 0.53 | [0.44-0.63] | <0.001 | 0.59 | [0.47-0.73] | <0.001 |
| High | 0.42 | [0.34-0.53] | <0.001 | 0.52 | [0.4-0.68] | <0.001 |
| **Smoking status** | | | | | | |
| Current smoker (Reference) | 1 | | | | | |
| Ex-smoker | 1.03 | [0.87-1.22] | 0.709 | 1.11 | [0.92-1.34] | 0.269 |
| Never smoked | 0.79 | [0.66-0.94] | 0.008 | 0.86 | [0.7-1.05] | 0.134 |
| **Membership at gym** | | | | | | |
| No (Reference) | 1 | | | 1 | | |
| Yes | 0.75 | [0.65-0.87] | <0.001 | 0.90 | [0.76-1.07] | 0.234 |
| **Pseudo R2 = 0.0379** | | | | | | |

### 3.5.2 Variable selection models for multivariate analyses for total hip OA

Backward and forward selection models were fitted (using the stepwise function in Stata, with forward selection set as pe(.05) and backward – pr(.05) separately). The automatic stepwise backward model was chosen as the final one and its selected variables for knee OA are shown in Table 18. All logistic regression analyses for no hip OA versus total hip OA were weighted using England's population distribution by age and sex.

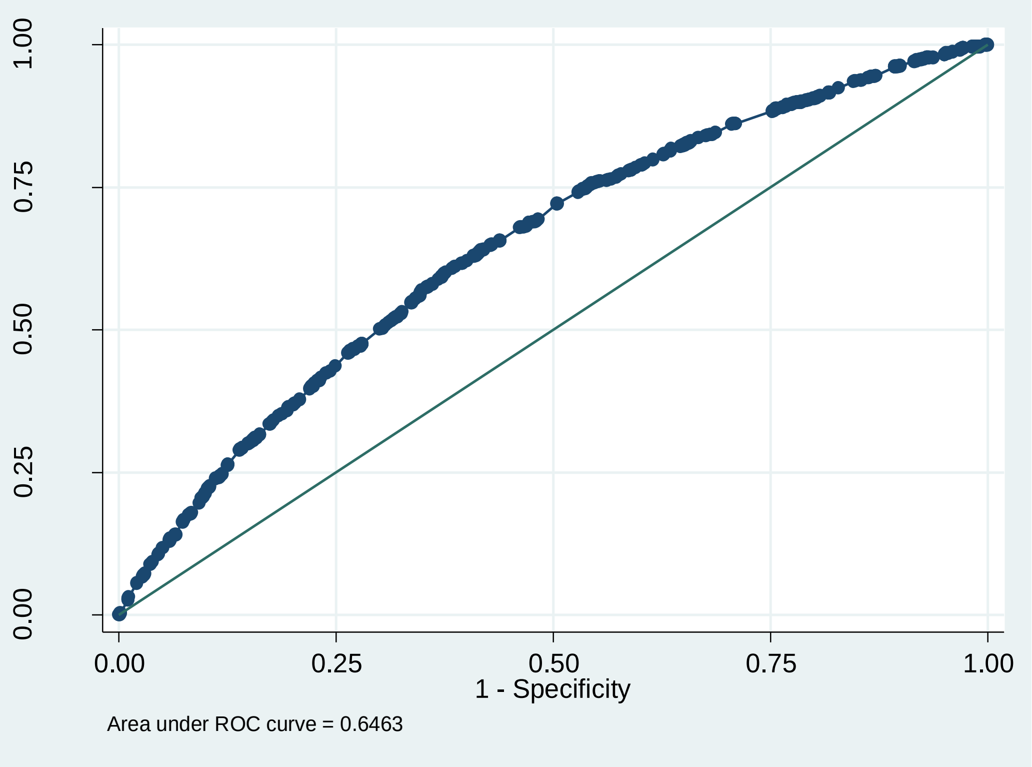
**Table 18: final overall/total hip OA model**

| | **Auto stepwise backward** (logistic) | | | **Auto stepwise backward** (logit) | | |
|---|---|---|---|---|---|---|
| | OR | 95% CIs | p-value | Coeff. | 95% CIs | p-value |
| **Age** | | | | | | |
| 45-64 (Reference) | 1 | | | 0 | | |
| 65-74 | 1 | | | 0 | | |
| Over 75 | 0.85 | [0.73-0.98] | 0.030 | -0.17 | [-0.32--0.02] | 0.030 |
| **Gender** | | | | | | |
| Male (Reference) | 1 | | | 0 | | |
| Female | 1.82 | [1.59-2.07] | <0.001 | 0.60 | [0.46-0.73] | <0.001 |
| **BMI** | | | | | | |
| Underweight <18.4 (Reference) | 1 | | | 0 | | |
| Normal weight 18.5-24 | 1 | | | 0 | | |
| Overweight 25-29 | 1.39 | [1.18-1.63] | <0.001 | 0.33 | [0.17-0.49] | <0.001 |
| Obese >30 | 1.87 | [1.59-2.2] | <0.001 | 0.63 | [0.47-0.79] | <0.001 |
| **Physical activity** | | | | | | |
| Sedentary (Reference) | 1 | | | | | |
| Low | 1 | | | 0 | | |
| Moderate | 0.59 | [0.52-0.68] | <0.001 | -0.52 | [-0.65--0.39] | <0.001 |
| High | 0.50 | [0.41-0.61] | <0.001 | -0.69 | [-0.88--0.5] | <0.001 |
| **Smoking status** | | | | | | |
| Current smoker (Reference) | 1 | | | 0 | | |
| Ex-smoker | 1 | | | 0 | | |

| | Auto stepwise backward (logistic) | | | Auto stepwise backward (logit) | | |
|---|---|---|---|---|---|---|
| | OR | 95% CIs | p-value | Coeff. | 95% CIs | p-value |
| Never smoked | 0.77 | [0.68-0.88] | <0.001 | -0.26 | [-0.38--0.13] | <0.001 |
| **Socioeconomic status** | | | | | | |
| Higher managerial (Reference) | 1 | | | 0 | | |
| Lower managerial | 1 | | | 0 | | |
| Intermediate | 0.79 | [0.66-0.95] | 0.012 | -0.23 | [-0.41--0.05] | 0.012 |
| Small employers | 1.23 | [1.02-1.48] | 0.034 | 0.20 | [0.02-0.39] | 0.034 |
| Lower supervisory | 1 | | | 0 | | |
| Semi-routine | 1 | | | 0 | | |
| Routine | 1 | | | 0 | | |
| Never worked | 1 | | | 0 | | |
| **Education** | | | | | | |
| Qualification present (Reference) | 1 | | | 0 | | |
| No qualification | 1.14 | [1-1.3] | 0.048 | 0.13 | [0-0.26] | 0.048 |
| **Constant** | NA | NA | NA | -2.26 | [-2.46--2.07] | <0.001 |

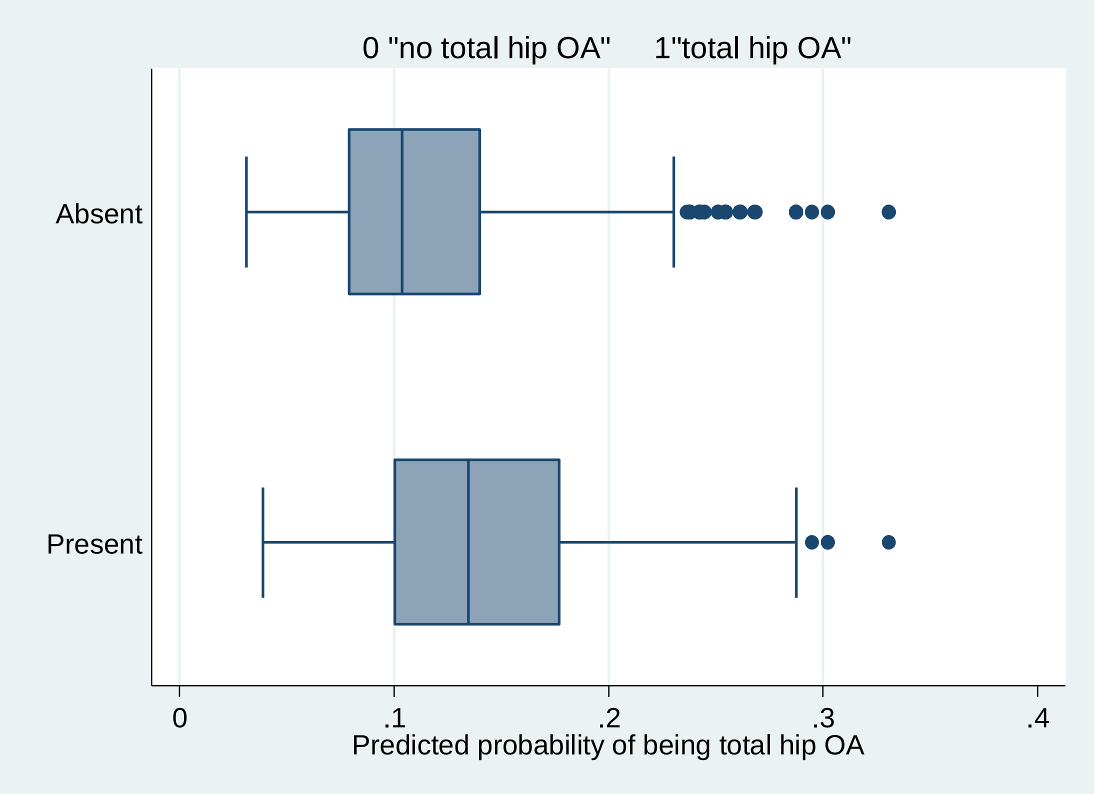
Figure 9 shows that this model does not perform as well as some others as the area under the ROC curve is only 0.65. This needs to be borne in mind in interpreting the synthetic estimates for this model.

**Figure 9: ROC curve for final total hip OA model**



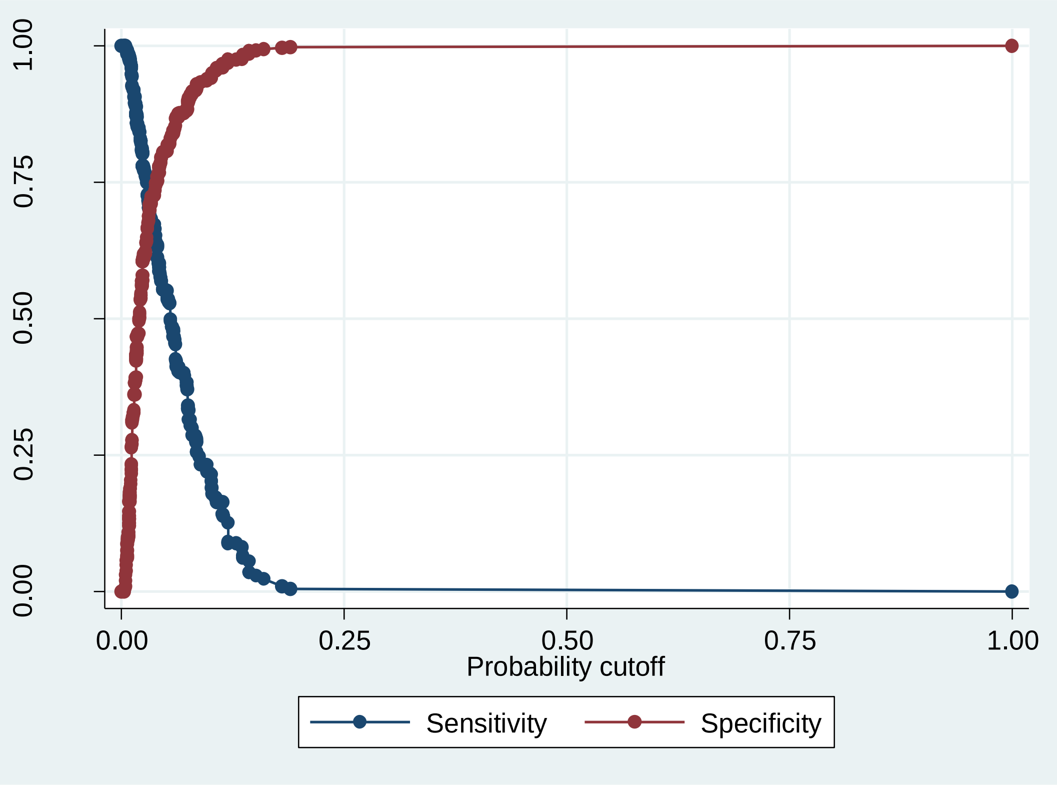
The two box plots show the predicted probability of people with overall hip OA caseness among the versus no hip OA group. We can see that people with overall hip OA have higher predicted probability than the no hip OA respondents. Since we have a binary response model, we can choose a cut-off point using the predicted probability to separate the predicted overall hip OA cases from the predicted no hip OA cases. No matter which cut-off point we choose, there will always be misclassified respondents, and the misclassification is worse in this model which has a lower area under the ROC curve. Therefore, we use sensitivity and specificity plots to help with this decision.

**Figure 10: overall/total hip OA predicted probability**



The sensitivity/specificity versus probability cut-off plot (Figure 11) shows the corresponding sensitivity and specificity in each possible probability cut-off point. We choose the cut-off probability where sensitivity and specificity lines cross. At cut-off point of predicted probability 0.11, the sensitivity is 62% whereas specificity is 60%.

**Figure 11: sensitivity/specificity vs. probability cut-off**



Applying this cut-off probability to our data, the following table shows the comparison between predicted and true cases of hip OA in ELSA (Table 19). Confirming 62% (1,074 out of 1,726) correctly classified hip OA cases, and 60% (7,837 out of 13,044) correctly classified non-hip OA. Table 20 shows the predicted total hip OA caseness with different cut-off probabilities.

**Table 19: actual and predicted total hip OA caseness comparison when cut-off probability is 0.11**

| Compare total hip OA casesness | Reported hip OA | Reported non-hip OA | Total |
|---|---|---|---|
| Predicted hip OA | 1,074 | 5,207 | 6,281 |
| Predicted non-hip OA | 652 | 7,837 | 8,489 |
| Total | 1,726 | 13,044 | 14,770 |

**Table 20: predicted total hip OA caseness with different cut-off probabilities**

| Probability cut-off | 0.025 | 0.04 | 0.05 | 0.07 | 0.1 | 0.11 | 0.12 | 0.15 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| Sensitivity (%) | 100 | 99.59 | 98.06 | 89.22 | 74.02 | 62.25 | 57.70 | 37.99 | 0 |
| Specificity (%) | 0 | 1.28 | 4.94 | 21.11 | 47.71 | 60.08 | 64.54 | 79.16 | 100 |
| True positive | 1,726 | 1,719 | 1,693 | 1,540 | 1,278 | 1,074 | 996 | 656 | 0 |
| False positive | 13,044 | 12,877 | 12,400 | 10,290 | 6,821 | 5,207 | 4,625 | 2,718 | 0 |
| True negative | 0 | 167 | 644 | 2,754 | 6,223 | 7,837 | 8,419 | 10,326 | 13,044 |
| False negative | 0 | 7 | 33 | 186 | 448 | 652 | 730 | 1,070 | 1,726 |

## 3.6 Total knee OA

### 3.6.1 Univariate and multivariate analyses for total knee OA

Table 21 shows univariate and multivariate analyses for total knee OA. As expected fewer risk factors are significant in the multivariable model. However age over 75, lower educational level, lower socioeconomic class, overweight and obesity, lower physical activity and being an ex-smoker remain significant in the multivariable model, while gym membership is protective.

**Table 21: Univariate and multivariate analyses for total knee OA**

| | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | OR | 95%CI | p-value | OR | 95%CI | p-value |
| **Age** | | | | | | |
| 45-64 (Reference) | | | 1 | | | |
| 65-74 | 1.097 | 0.993-1.211 | 0.07 | 0.913 | 0.815-1.023 | 0.119 |
| Over 75 | 0.999 | 0.892-1.119 | 0.982 | 0.723 | 0.629-0.831 | <0.001 |
| **Sex** | | | | | | |
| Male (Reference) | 1 | | | 1 | | |
| Female | 1.212 | 1.108-1.325 | <0.001 | 1.206 | 1.078-1.35 | 0.001 |
| **Ethnicity** | | | | | | |
| White (Reference) | 1 | | | 1 | | |
| Non-white | 1.126 | 0.898-1.411 | 0.306 | 1.139 | 0.839-1.547 | 0.403 |
| **Education** | | | | | | |
| NVQ4/NVQ5/Degree or equivalent (Reference) | 1 | | | 1 | | |
| Higher education below degree | 1.413 | 1.189-1.68 | <0.001 | 1.114 | 0.913-1.36 | 0.288 |
| NVQ3/GCE A Level equivalent | 1.471 | 1.207-1.794 | <0.001 | 1.142 | 0.904-1.442 | 0.265 |
| NVQ2/GCE O Level equivalent | 1.419 | 1.209-1.667 | <0.001 | 1.015 | 0.836-1.232 | 0.882 |
| NVQ1/CSE other grade equivalent | 1.888 | 1.489-2.394 | <0.001 | 1.440 | 1.086-1.911 | 0.011 |
| Foreign/other | 1.763 | 1.444-2.151 | <0.001 | 1.234 | 0.973-1.564 | 0.083 |
| No qualification | 1.978 | 1.715-2.281 | <0.001 | 1.317 | 1.084-1.601 | 0.006 |
| **Socioeconomic class** | | | | | | |
| Higher managerial and professional occupations (Reference) | 1 | | | 1 | | |
| Lower managerial and professional occupations | 1.561 | 1.283-1.9 | <0.001 | 1.356 | 1.084-1.695 | 0.008 |
| Intermediate occupations | 1.541 | 1.248-1.904 | <0.001 | 1.162 | 0.903-1.494 | 0.243 |
| Small employers and own account workers | 1.569 | 1.26-1.953 | <0.001 | 1.240 | 0.963-1.596 | 0.096 |
| Lower supervisory and technical occupations | 2.068 | 1.665-2.568 | <0.001 | 1.592 | 1.233-2.056 | <0.001 |
| Semi-routine occupations | 2.070 | 1.697-2.524 | <0.001 | 1.521 | 1.193-1.939 | 0.001 |
| Routine occupations | 2.317 | 1.892-2.837 | <0.001 | 1.520 | 1.181-1.956 | 0.001 |
| Never worked and long term unemployed | 2.203 | 1.448-3.351 | <0.001 | 1.488 | 0.882-2.511 | 0.136 |

| | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | OR | 95%CI | p-value | OR | 95%CI | p-value |
| **BMI**[14] | | | | | | |
| Underweight <18.4 (Reference) | 1 | | | 1 | | |
| Normal weight 18.5-24 | 1.410 | 0.802-2.478 | 0.233 | 1.368 | 0.783-2.392 | 0.271 |
| Overweight 25-29 | 2.149 | 1.229-3.756 | 0.007 | 2.084 | 1.198-3.625 | 0.009 |
| Obese >30 | 4.355 | 2.492-7.612 | <0.001 | 3.668 | 2.11-6.377 | <0.001 |
| **Physical activity**[15] | | | | | | |
| Sedentary (Reference) | 1 | | | 1 | | |
| Low | 0.954 | 0.812-1.121 | 0.567 | 0.904 | 0.748-1.094 | 0.301 |
| Moderate | 0.482 | 0.413-0.563 | <0.001 | 0.510 | 0.423-0.615 | <0.001 |
| High | 0.373 | 0.31-0.448 | <0.001 | 0.424 | 0.338-0.533 | <0.001 |
| **Smoking status** | | | | | | |
| Current smoker (Reference) | 1 | | | | | |
| Ex-smoker | 1.256 | 1.087-1.451 | 0.002 | 1.348 | 1.143-1.59 | <0.001 |
| Never smoked | 1.030 | 0.887-1.197 | 0.696 | 1.149 | 0.967-1.365 | 0.114 |
| **Membership at gym** | | | | | | |
| No (Reference) | 1 | | | 1 | | |
| Yes | 0.673 | 0.597-0.759 | <0.001 | 0.850 | 0.737-0.98 | 0.026 |
| **Constant** | | | | 0.094 | 0.05-0.177 | <0.001 |
| **Pseudo R2 = 0.0586** | | | | | | |

### *3.6.2* Variable selection models for multivariate analyses for total knee OA

Table 22 shows how models were compared to arrive at the final multivariable overall or total knee OA vs. no OA model. Backward and forward variable selection models were fitted using the stepwise function in Stata, with forward selection set as pr(.05) and backward – pr(.05) separately). Work physical activity was excluded. The automatic stepwise backward model was chosen as the final method and its selected variables for hip OA are shown in Table 22. All logistic regression analyses for no hip OA versus total hip OA were weighted using England's population distribution by age and sex.

**Table 22: final overall/total knee OA model**

| | Auto stepwise backward (logistic) | | | Auto stepwise backward (logit) | | |
|---|---|---|---|---|---|---|
| | OR | 95% CI | p-value | Coeff | 95% CI | p-value |
| **Age** | | | | | | |
| 45-64 (Reference) | 1 | | | 0 | | |
| 65-74 | 1 | | | 0 | | |
| Over 75 | 0.77 | [0.68-0.87] | <0.001 | -0.27 | [-0.39--0.14] | <0.001 |
| **Gender** | | | | | | |
| Male (Reference) | 1 | | | 0 | | |
| Female | 1.24 | [1.11-1.37] | <0.001 | 0.21 | [0.11-0.32] | <0.001 |
| **Education** | | | | | | |

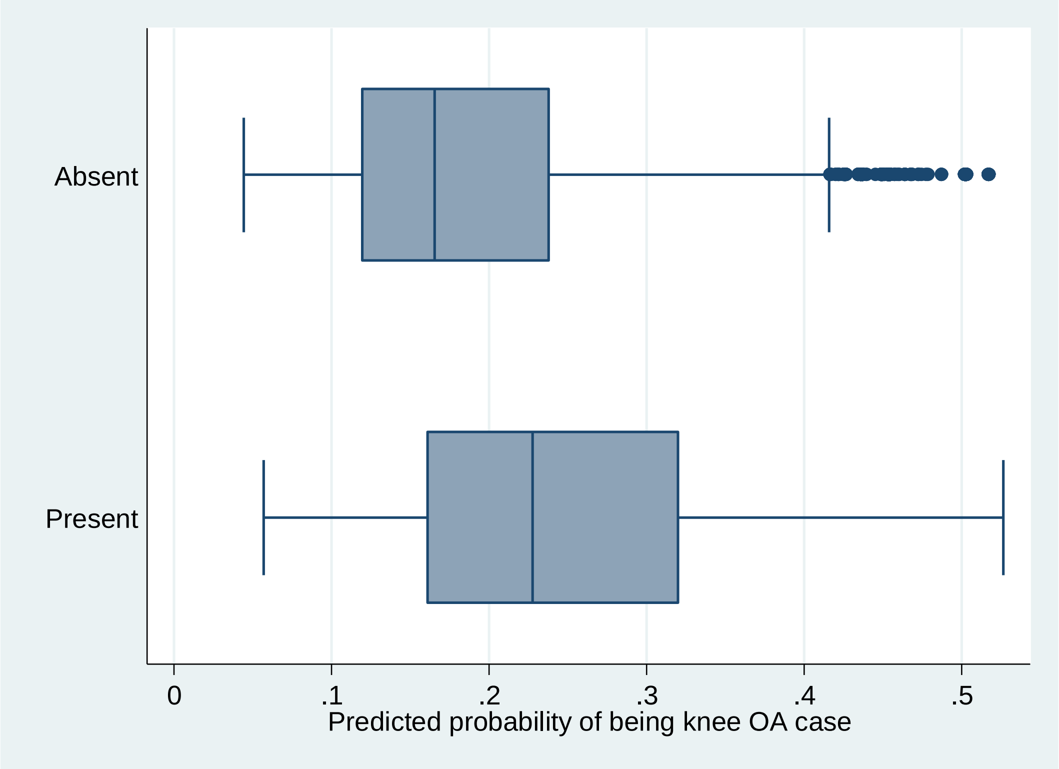
[14] BMI grouping was changed to match the format found in Active People Survey (three 'obese' categories were merged into one 'obese' if BMI was more than 30.

[15] Note that work physical activity has been excluded because of lack of data for LAs

| | Auto stepwise backward (logistic) | | | Auto stepwise backward (logit) | | |
|---|---|---|---|---|---|---|
| | OR | 95% CI | p-value | Coeff | 95% CI | p-value |
| NVQ4/NVQ5/Degree or equivalent (Reference) | 1 | | | 0 | | |
| Higher education below degree | 1 | | | 0 | | |
| NVQ3/GCE A Level equivalent | 1 | | | 0 | | |
| NVQ2/GCE O Level equivalent | 1 | | | 0 | | |
| NVQ1/CSE other grade equivalent | 1 | [1.05-1.7] | 0.018 | 0.29 | [0.05-0.53] | 0.018 |
| Foreign/other | 1 | | | 0 | | |
| No qualification | 1.22 | [1.08-1.37] | 0.002 | 0.20 | [0.07-0.32] | 0.002 |
| **Socioeconomic class** | | | | | | |
| Higher managerial and professional occupations (Reference) | 1 | | | 0 | | |
| Lower managerial and professional occupations | 1 | [1.02-1.35] | 0.023 | 0.16 | [0.02-0.3] | 0.023 |
| Intermediate occupations | 1 | | | 0 | | |
| Small employers and own account workers | 1 | | | 0 | | |
| Lower supervisory and technical occupations | 1.39 | [1.17-1.66] | <0.001 | 0.33 | [0.15-0.51] | <0.001 |
| Semi-routine occupations | 1.32 | [1.14-1.53] | <0.001 | 0.28 | [0.13-0.42] | <0.001 |
| Routine occupations | 1.31 | [1.11-1.55] | 0.001 | 0.27 | [0.11-0.44] | 0.001 |
| Never worked and long term unemployed | 1 | | | 0 | | |
| **BMI** | | | | | | |
| Underweight <18.4 (Reference) | 1 | | | 0 | | |
| Normal weight 18.5-24 | 1 | | | 0 | | |
| Overweight 25-29 | 1.56 | [1.36-1.79] | <0.001 | 0.44 | [0.31-0.58] | <0.001 |
| Obese >30 | 2.75 | [2.4-3.15] | <0.001 | 1.01 | [0.87-1.15] | <0.001 |
| **Physical activity** | | | | | | |
| Sedentary (Reference) | 1 | | | 0 | | |
| Low | 1 | | | 0 | | |
| Moderate | 0.55 | [0.5-0.62] | <0.001 | -0.59 | [-0.7--0.48] | <0.001 |
| High | 0.46 | [0.39-0.55] | <0.001 | -0.77 | [-0.94--0.6] | <0.001 |
| **Smoking status** | | | | | | |
| Current smoker (Reference) | 1 | | | 0 | | |
| Ex-smoker | 1.21 | [1.1-1.34] | <0.001 | 0.19 | [0.09-0.29] | <0.001 |
| Never smoked | 1 | | | 0 | | |
| **Membership at gym** | | | | | | |
| No (Reference) | 1 | | | 0 | | |
| Yes | 0.84 | [0.73-0.97] | 0.019 | -0.17 | [-0.31--0.03] | 0.019 |
| **Constant** | **NA** | **NA** | **NA** | **-1.88** | **[-2.06--1.69]** | **<0.001** |

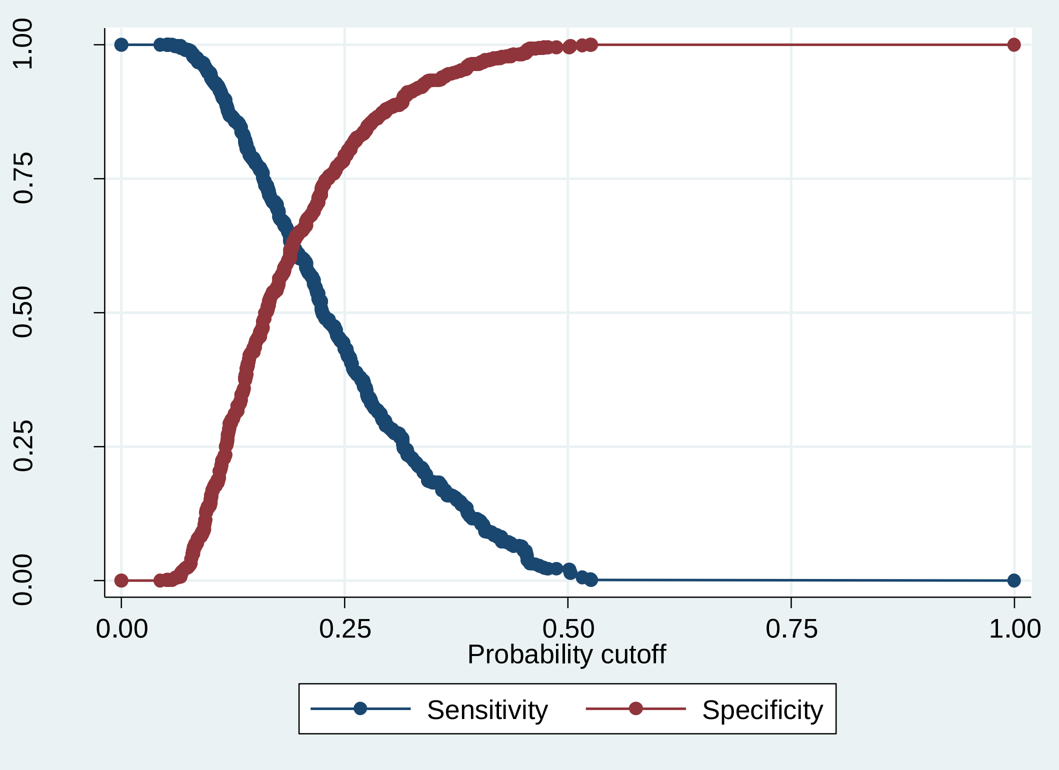
The two box plots in Figure 12 show the predicted probability of people with overall knee OA caseness among the versus no knee OA group. We can see that people with overall knee OA have higher predicted probability than the no knee OA respondents, but there is a large overlap.

**Figure 12: overall/total knee OA predicted probability**



The sensitivity/specificity versus probability cut-off plot (Figure 13) shows the corresponding sensitivity and specificity in each possible probability cut-off point.

**Figure 13: sensitivity/specificity vs. probability cut-off**



We chose the cut-off probability where sensitivity and specificity lines cross. At a cut-off point of predicted probability 0.19, the sensitivity reaches around 63%, whereas specificity reaches around 62%. Applying this cut-off probability to our data, Table 22 and Table 23 show the comparison between predicted and true cases of knee OA in ELSA. Confirming 63.0% (1,692 out of 2,681) correctly classified knee OA cases, and 61.8% (7,475 out of 12,089) correctly classified non-knee OA.

**Table 22 Actual and predicted knee OA caseness comparison when cut-off probability is 0.19**

| Compare knee OA casesness | Reported knee OA | Reported non-knee OA | Total |
|---|---|---|---|
| Predicted knee OA | 1,692 | 4,614 | 6,306 |
| Predicted non-knee OA | 989 | 7,475 | 8,464 |
| Total | 2,681 | 12,089 | 14,770 |

**Table 23 Predicted total knee OA caseness with different cut-off probabilities**

| Probability cut-off | 0.025 | 0.04 | 0.10 | 0.11 | 0.18 | 0.19 | 0.20 | 0.25 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| Sensitivity (%) | 100 | 100 | 94.29 | 91.43 | 67.15 | 63.10 | 60.20 | 43.27 | 1.98 |
| Specificity (%) | 0 | 0 | 14.99 | 20.26 | 57.09 | 61.83 | 65.12 | 79.30 | 99.58 |
| True positive | 2,681 | 2,681 | 2,528 | 2,451 | 1,800 | 1,692 | 1,614 | 1,160 | 53 |
| False positive | 12,089 | 12,089 | 10,277 | 9,640 | 5,187 | 4,614 | 4,217 | 2,502 | 51 |
| True negative | 0 | 0 | 1,812 | 2,449 | 6,902 | 7,475 | 7,872 | 9,587 | 12,038 |
| False negative | 0 | 0 | 153 | 230 | 881 | 989 | 1,067 | 1,521 | 2,628 |

## 3.7 Synthetic estimation

### 3.7.1 Prevalence calculation

As described in Methods section 2.8 and Appendix 3: synthetic estimation using Stata software, the odds ratios and coefficients are on a logarithmic scale, so they were added instead of multiplied. For example, for a female over 75 years of age, who has never worked, with BMI over 30, but with high levels of physical activity, who is an ex-smoker with no educational qualifications the outcome would be calculated (see Figure 39:

Outcome = -2.2649 + (-0.1663) + 0.597079 + 0 + 0.628003 + (-0.68873) + 0 + 0.131858 = -1.76299

ODDS = exp(-1.76299)= 0.1715

Prevalence = 0.1715/(1+0.1715) = 0.1464/14.6% (this shows the prevalence rate of total hip OA for a person with those characteristics).

### 3.7.2 Estimating number of people (population) with different characteristics

Local population data for every risk factor that was used in a model is obtained. For example, for a female over 75 years of age, who has never worked, with a BMI over 30, but with high levels of physical activity, who is an ex-smoker with no educational qualifications the population would be calculated at MSOA level (Hartlepool 001):

Population = Number of females over 75 * proportion of population that never worked * proportion of population whose BMI over 30 * proportion of population that has high physical activity levels * proportion of population that never smoked * proportion of population that do not have education.

This population (0.3221) is then multiplied by the prevalence, or proportion of cases in that population, from the regression model:

Proportion = 0.1464 * 0.3221 = **0.0472** (this shows number of cases of total hip OA at Hartlepool 001 for a person with characteristics described above).

The sum of all values in this table is the number of expected cases of Hip or Knee OA in the selected MSOA/LA or practice/CCG.

### 3.7.3 Estimating number of severe/total hip and knee OA cases

The number of cases is calculated by multiplying the prevalence by the population in each demographic category for the selected MSOA/LA or practice/CCG. For example, for a female over 75 years of age, who has never worked, with BMI over 30, but with high levels of physical activity, who is an ex-smoker with no education, the proportion would be calculated:

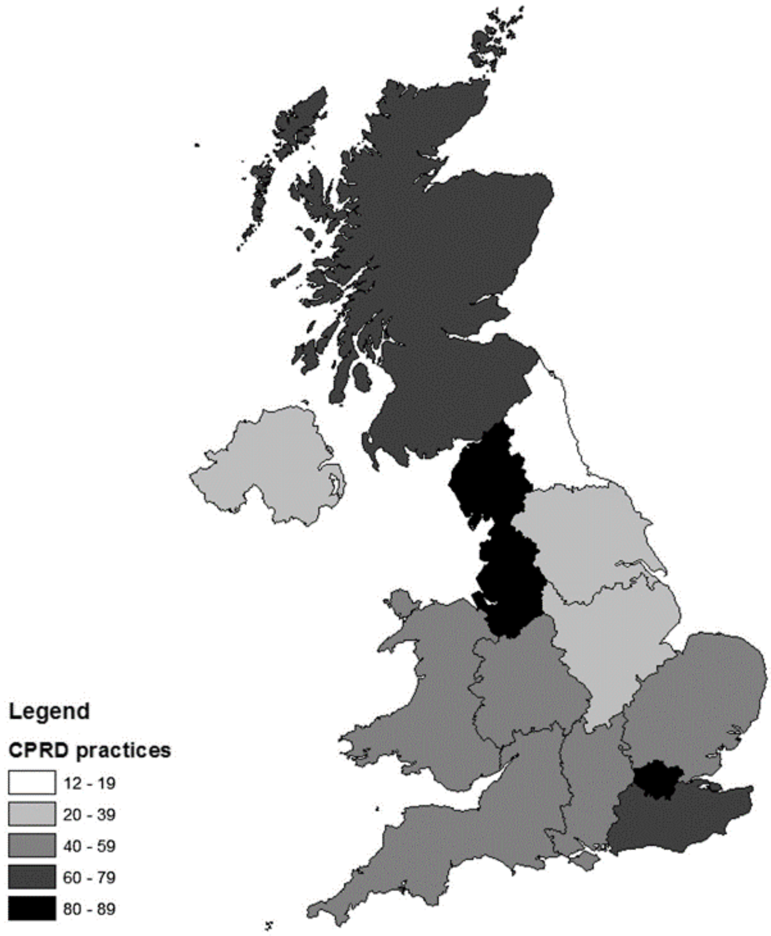
***Proportion*** *= Prevalence * Population*

# 4 External validation - CPRD processing & modelling

## 4.1 Data source & sampling

The Clinical Practice Research Datalink (CPRD) is an ongoing primary care database of longitudinal anonymised electronic health records (EHRs) from general practitioners, with coverage of over 11.3 million patients from 674 practices in the UK. With 4.4 million active (alive, currently registered) patients meeting quality criteria, approximately 6.9% of the UK population are included and patients are broadly representative of the UK general population in terms of age, sex and ethnicity. The distribution of CPRD practices is shown in Figure 1 below.

**Figure 14: distribution of 674 CPRD practices by region in England, and in Wales, Scotland and Northern Ireland**



We used data extracted from the Clinical Practice Research Datalink (http://www.cprd.com/intro.asp) to fit an OA predictive model. We identified cases of OA by a medcodes list. Records with these medcodes were extracted from the CPRD Clinical, Referral and Test databases. The main extraction took place in January 2016. The proportion of the sample were randomly selected from the total CPRD sample.

## 4.2 Missing data

CPRD data source may not include patient's data in terms of all the demographic aspects, such as ethnicity, smoking and BMI. There is some missing data in the above areas, and different methods were used to deal with missing data. Multiple imputation was used to replace missing values for BMI, smoking status, ethnicity group and deprivation. Proportion of missing data at baseline is shown in Table 24.

**Table 24: Proportion of missing data**

| Predictor variables | | Total |
|---|---|---|
| Total number of respondents | | 711,002 |
| Gender | | |
| | **Male** | 263,151 (37.01%) |
| | **Female** | 447,851 (62.99%) |
| | **Missing** | 0% |
| Age group | | |
| | **45-64** | 55,239 (7.77%) |
| | **65-74** | 304,164 (42.78%) |
| | **>75** | 350,612 (49.31%) |
| | **Missing** | 0% |
| Ethnicity | | |
| | **White** | 27,651 (3.89%) |
| | **Mixed** | 992 (0.14%) |
| | **Asian** | 1,168 (0.16%) |
| | **Black** | 558 (0.08%) |
| | **Other** | 424 (0.06%) |
| | **Missing** | 680,209 (95.67%) |
| BMI | | |
| | **Underweight (<18.5)** | 22,503 (3.16%) |
| | **Normal (18.5-25)** | 184,480 (25.95%) |
| | **Overweight (25-30)** | 174,418 (24.53%) |
| | **Obese (>30)** | 102,853 (14.47%) |
| | **Missing** | 226,750 (31.89%) |
| Smoking | | |
| | **Non-smoker** | 325,039 (45.72%) |
| | **Ex-smoker** | 172,709 (24.29%) |
| | **Smoker** | 50,884 (7.16%) |
| | **Missing** | 162,372 (22.84%) |
| Deprivation | | |
| | **1 (least deprived)** | 11,760 (1.65%) |
| | **2** | 12,564 (1.77%) |
| | **3** | 14,150 (1.99%) |
| | **4** | 14,199 (2.00%) |
| | **5 (most deprived)** | 12,357 (1.74%) |
| | **Missing** | 645,972 (90.85%) |

## 4.3 Descriptive characteristics of patients who have had OA

**Table 25** shows the baseline characteristics of patients (both identified OA cases and non-OA cases) that are included in the model. The distributions are relatively similar, despite that there is a greater number of younger individuals in the OA group.

**Table 25: Baseline characteristics of patients involved in the logistic regression model**

| Predictor variables | OA cases | Non-OA cases | Total |
|---|---|---|---|
| **Total number of respondents** | 179,396 | 531,606 | 711,068 |
| **Gender** | | | |
| **Male** | 63,474 (35.38%) | 199,677 (37.56%) | 263,167 (37.01%) |
| **Female** | 115,922 (64.62%) | 331,929 (62.44%) | 447,899(62.99%) |
| **Total** | 179,396 | 531,606 | 711,068 |
| **Age group** | | | |
| **45-64** | 27,891 (15.55%) | 27,348 (5.14%) | 82,302 (11.57%) |
| **65-74** | 61,384 (34.22%) | 242,781 (45.67%) | 298,235 (41.94%) |
| **>75** | 89,134 (49.69%) | 261,478 (49.19%) | 329,544 (46.34%) |
| **Total** | 179,396 | 531,606 | 711,068 |
| **Ethnicity** | | | |
| **White** | 176,254 (98.25%) | 525,573 (98.85%) | 701,828 (98.70%) |
| **Mixed** | 992 (0.55%) | 1,846 (0.35%) | 2,838 (0.40%) |
| **Asian** | 1,168 (0.65%) | 2,119 (0.40%) | 3,287 (0.46%) |
| **Black** | 558 (0.31%) | 1,185 (0.22%) | 1,743 (0.25%) |
| **Other** | 424 (0.24%) | 948 (0.18%) | 1,372 (0.19%) |
| **Total** | | | |
| **BMI** | | | |
| **Underweight (<18.5)** | 6,398 (4.15%) | 16,105 (4.88%) | 22,507 (4.65%) |
| **Normal (18.5-25)** | 54,912 (35.62%) | 129,568 (39.25%) | 184,485 (38.10%) |
| **Overweight (25-30)** | 55,961 (36.30%) | 118,459 (35.88%) | 174,420 (36.02%) |
| **Obese (>30)** | 36,881 (23.93%) | 65,972 (19.98%) | 102,853 (21.24%) |
| **Total** | 154,152 | 330,113 | 484,265 |
| **Smoking** | | | |
| **Non-smoker** | 100,000 (58.48%) | 225,065 (59.59%) | 325,065(59.25%) |
| **Ex-smoker** | 56,887 (33.27%) | 115,823 (30.67%) | 172,710 (31.48%) |
| **Smoker** | 14,114 (8.25%) | 36,771 (9.74%) | 50,886 (9.27%) |
| **Total** | 171,002 | 377,659 | 548,661 |
| **Deprivation** | | | |
| **1 (least deprived)** | 2,785 (16.15%) | 8,975 (18.78%) | 11,760 (18.08%) |
| **2** | 3,388 (19.65%) | 9,176 (19.20%) | 12,564 (19.32%) |
| **3** | 3,688 (21.38%) | 10,463 (21.90%) | 14,151 (21.76%) |
| **4** | 3,900 (22.61%) | 10,300 (21.55%) | 14,200 (21.84%) |
| **5 (most deprived)** | 3,485 (20.21%) | 8,873 (18.57%) | 12,358 (19.00%) |
| **Total** | 17,246 | 47,787 | 65,033 |

## 4.4 Univariate analysis

Table 26 shows the results of univariate models for individual risk factors and the outcome.

**Table 26: Univariate models for individual risk factors**

| Predictor variables | Odds Ratio | P>t | [95% Conf. Interval] |
|---|---|---|---|
| Gender | | | |
| **Male** | 1.00 | | |
| **Female** | 2.114 | <0.001 | [2.083 - 2.146] |
| Age group | | | |
| **45-64** | 5.189 | <0.001 | [5.092 - 5.287] |
| **65-74** | 8.680 | <0.001 | [8.482 - 8.883] |
| **>75** | 5.289 | <0.001 | [5.172 - 5.410] |
| Ethnicity | | | |
| **White** | 1.00 | | |
| **Mixed** | 1.497 | <0.001 | [1.374 - 1.632] |
| **Asian** | 0.965 | 0.223 | [0.910 – 1.022] |
| **Black** | 0.624 | <0.001 | [0.573 - 0.679] |
| **Other** | 0.650 | <0.001 | [0.588 – 0.717] |
| BMI | | | |
| **Underweight (<18.5)** | 1.536 | <0.001 | [1.471 - 1.603] |
| **Normal (18.5-25)** | 1.00 | | |
| **Overweight (25-30)** | 1.297 | <0.001 | [1.272 - 1.322] |
| **Obese (>30)** | 1.720 | <0.001 | [1.685 - 1.756] |
| Smoking | | | |
| **Non-smoker** | 1.00 | | |
| **Ex-smoker** | 1.709 | <0.001 | [1.678 - 1.740] |
| **Smoker** | 0.773 | <0.001 | [0.757 – 0.789] |
| Deprivation | | | |
| **1 (least deprived)** | 1.00 | | |
| **2** | 1.082 | 0.007 | [1.022 - 1.145] |
| **3** | 1.120 | <0.001 | [1.060 - 1.185] |
| **4** | 1.225 | <0.001 | [1.159 - 1.295] |
| **5 (most deprived)** | 1.312 | <0.001 | [1.238 - 1.390] |

## 4.5 Logistic regression model

We have fitted a logistic regression model using CPRD data by including the available risk factor variables as in the ELSA model. However, social economic status, gym membership, physical activity and education data were not available in the CPRD dataset, so we removed these variables from the CPRD OA model.

**Table 27: Logistic Regression model for severe OA from CPRD**

| Predictor variables | Odds Ratio | P>t | [95% Conf. Interval] |
|---|---|---|---|
| **Gender** | | | |
| **Male** | 1.00 | | |
| **Female** | 2.209 | <0.001 | [2.168 – 2.250] |
| **Age group** | | | |
| **45-64** | 1.00 | | |
| **65-74** | 9.242 | <0.001 | [9.001 – 9.490] |
| **>75** | 5.679 | <0.001 | [5.533 – 5.828] |
| **Alcohol** | | | |
| **Non-drinker** | 1.00 | | |
| **Light (<15 units per week)** | 0.983 | 0.215 | [0.955 – 1.011] |
| **Moderate (14-42 units per week)** | 0.803 | <0.001 | [0.768 – 0.839] |
| **Heavy (>42 units per week)** | 0.752 | <0.001 | [0.687 – 0.824] |
| **Ethnicity** | | | |
| **White** | 1.00 | | |
| **Mixed** | 1.427 | <0.001 | [1.281 – 1.590] |
| **Asian** | 1.440 | <0.001 | [1.339 – 1.549] |
| **Black** | 0.815 | <0.001 | [0.731 – 0.909] |
| **Other** | 0.996 | 0.948 | [0.884 – 1.122] |
| **BMI** | | | |
| **Underweight (<18.5)** | 1.266 | <0.001 | [1.187 – 1.351] |
| **Normal (18.5-25)** | 1.00 | <0.001 | |
| **Overweight (25-30)** | 1.117 | <0.001 | [1.092 – 1.143] |
| **Obese (>30)** | 1.358 | <0.001 | [1.327 – 1.391] |
| **Smoking** | | | |
| **Non-smoker** | 1.00 | | |
| **Ex-smoker** | 1.584 | <0.001 | [1.546 – 1.622] |
| **Smoker** | 1.144 | <0.001 | [1.112 - 1.178] |
| **Deprivation** | | | |
| **1 (least deprived)** | | | |
| **2** | 1.118 | 0.017 | [1.028 - 1.215] |
| **3** | 1.127 | 0.021 | [1.027 - 1.236] |
| **4** | 1.274 | <0.001 | [1.192 - 1.361] |
| **5 (most deprived)** | 1.341 | <0.001 | [1.268 - 1.418] |
| **_cons** | 0.031 | <0.001 | [0.030 - 0.033] |

# 5 Production of Scottish local estimates

## 5.1 Methods

We used the English population model developed from ELSA to produce the prevalence of overall and severe hip OA, and overall and severe knee OA for Scotland. However local risk factor data availability affected the risk factor variables included in the prediction model. Therefore some variables had to be dropped from the English model because no local Scottish data was available. To validate the performance of the subsequent models internally they were compared with the English models by c-statistics (ROC curves). The external validity of this approach was evaluated subsequently using Scottish Biobank data.

## 5.2 Results

Gym membership and social economic status data were not available at any local levels in Scotland. We removed these two variables from the final OA models and fitted logistic regression models based on other available variables. The performance of the overall hip OA, overall knee OA, severe hip OA and severe knee OA models were listed below. The discrimination and prediction of the models between Scotland and England were still quite similar.

**Figure 15: Overall hip OA English model**

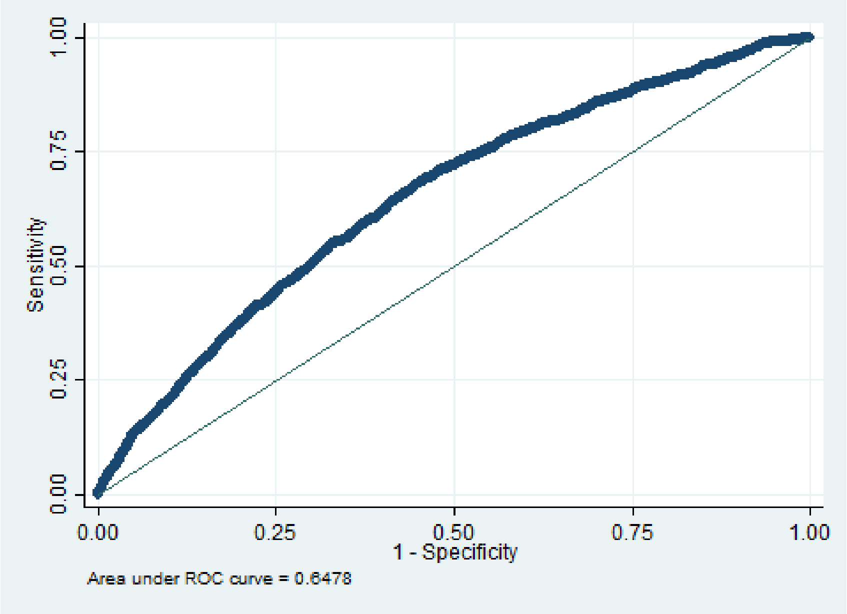


Figure 16: Overall hip OA Scottish model

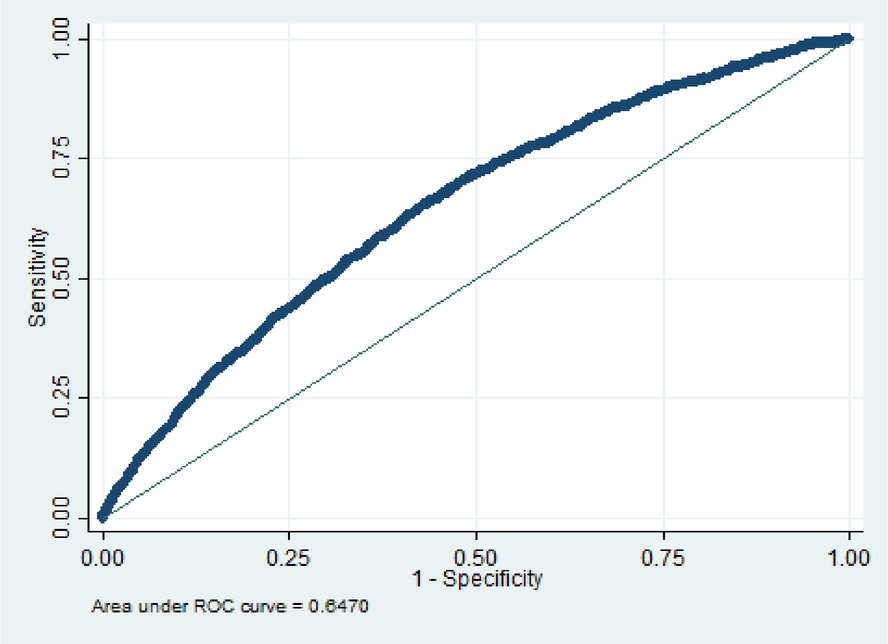


Figure 17: Severe hip OA English model

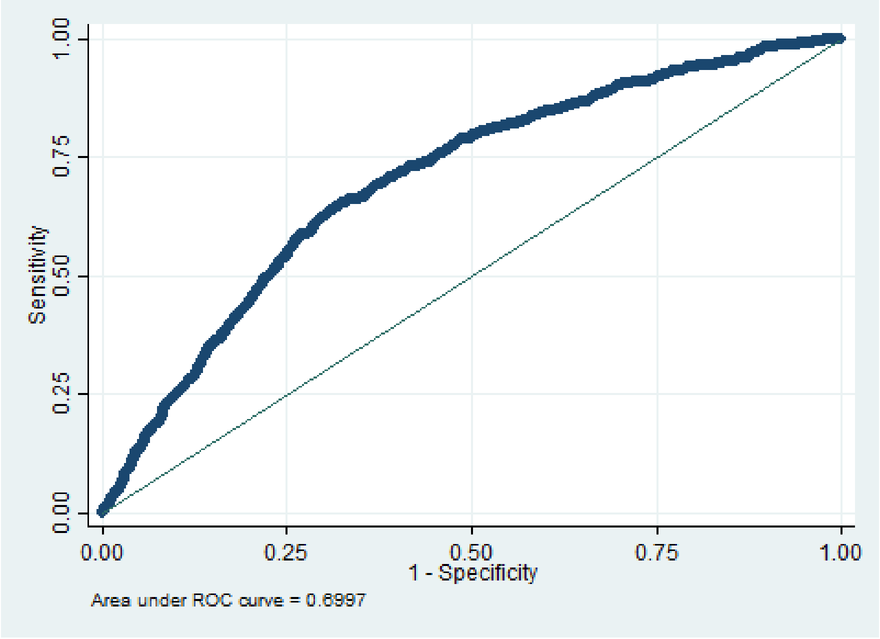


**Figure 18: Severe hip OA Scottish model**

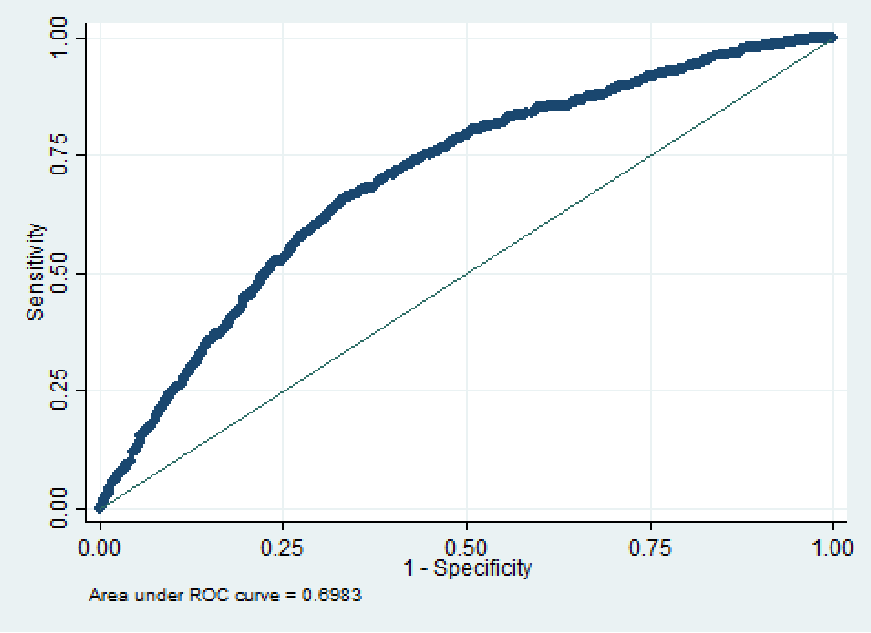


**Figure 19: General knee OA English model**

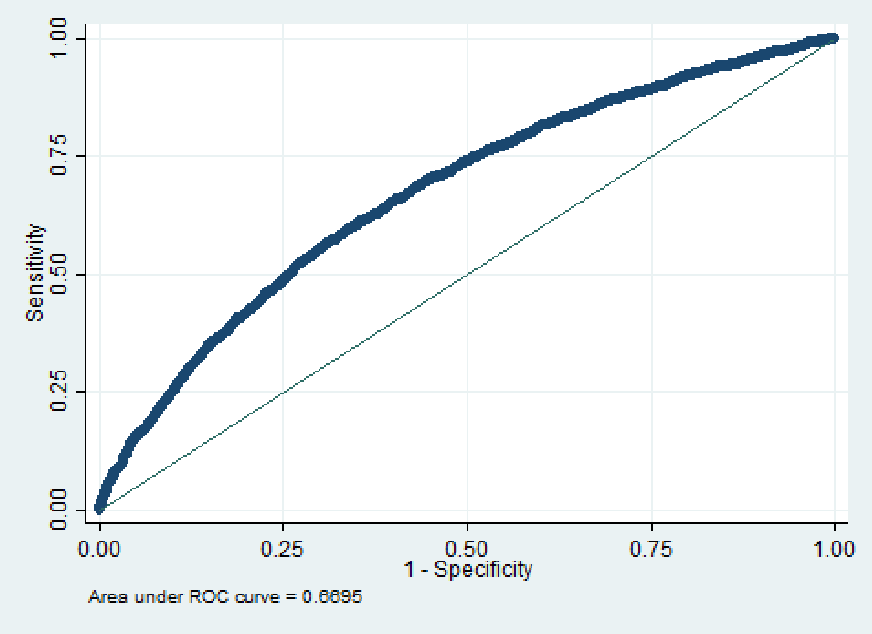


**Figure 20: General knee OA Scottish model**

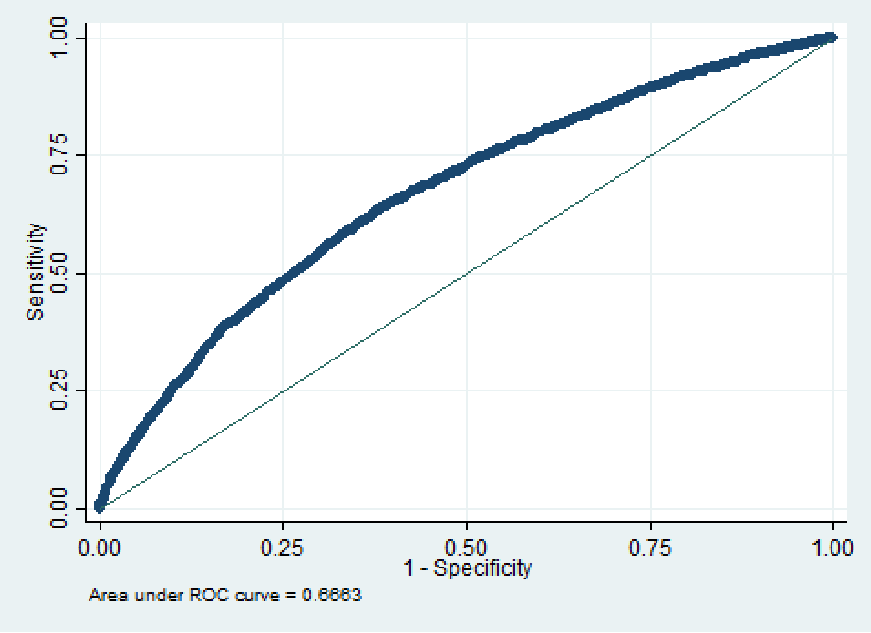


**Figure 21: Severe knee OA English model**

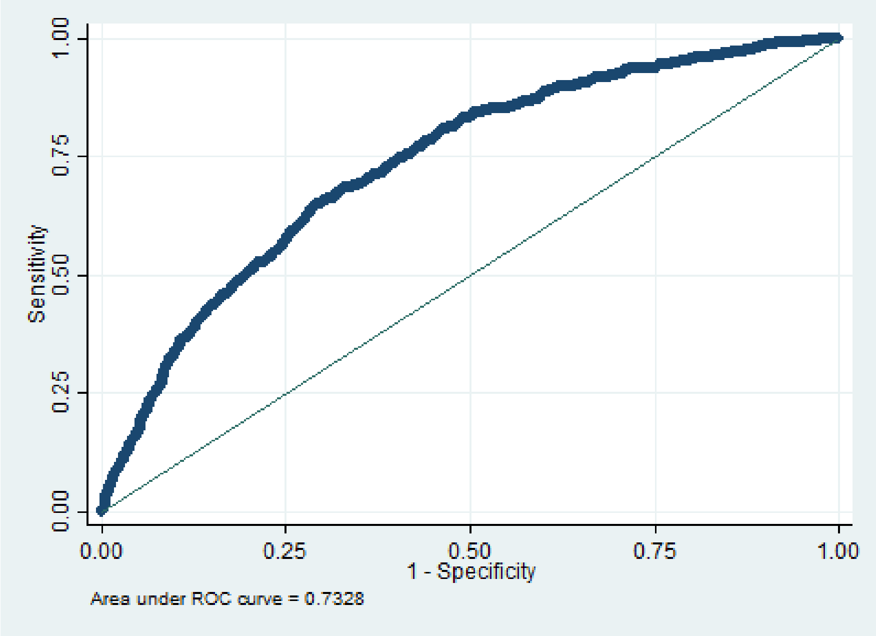


**Figure 22: Severe knee OA Scottish model**

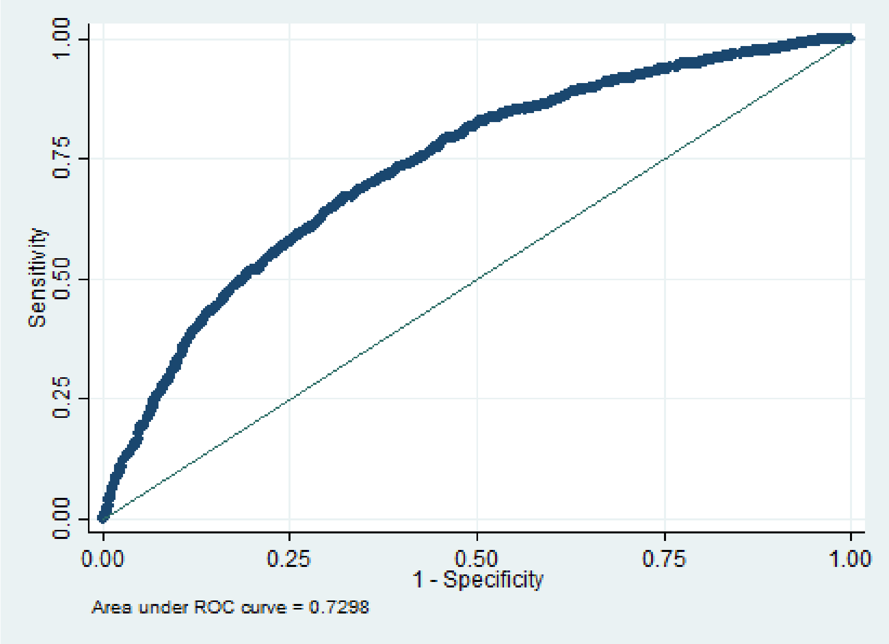


**Figure 23: Histogram of the prevalence of overall hip OA for Scotland at LA level**

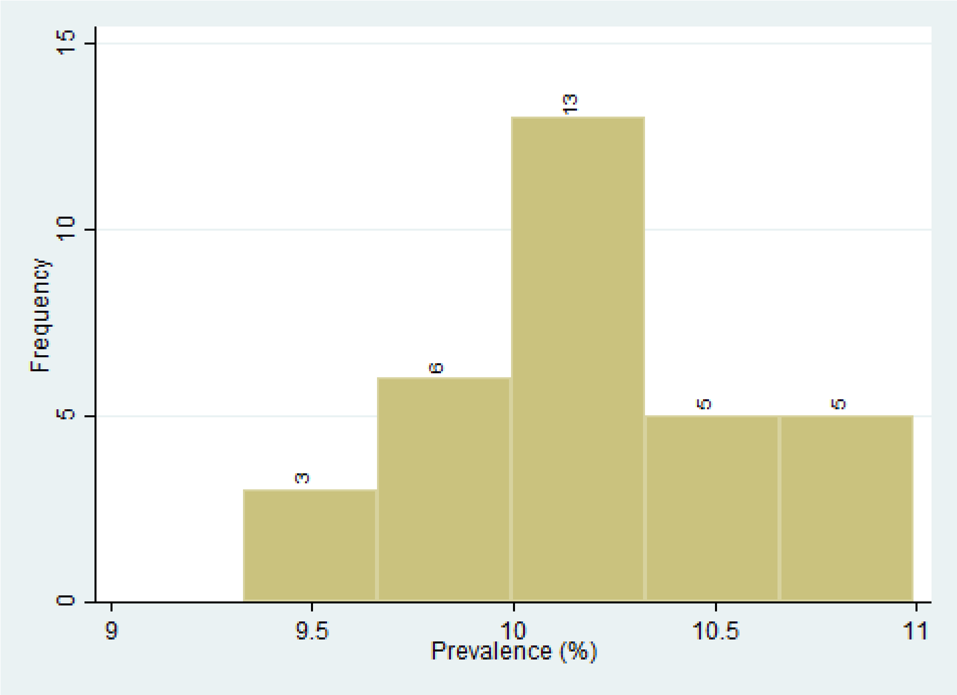


**Figure 24: Histogram of the prevalence of overall hip OA for Scotland at practice level**

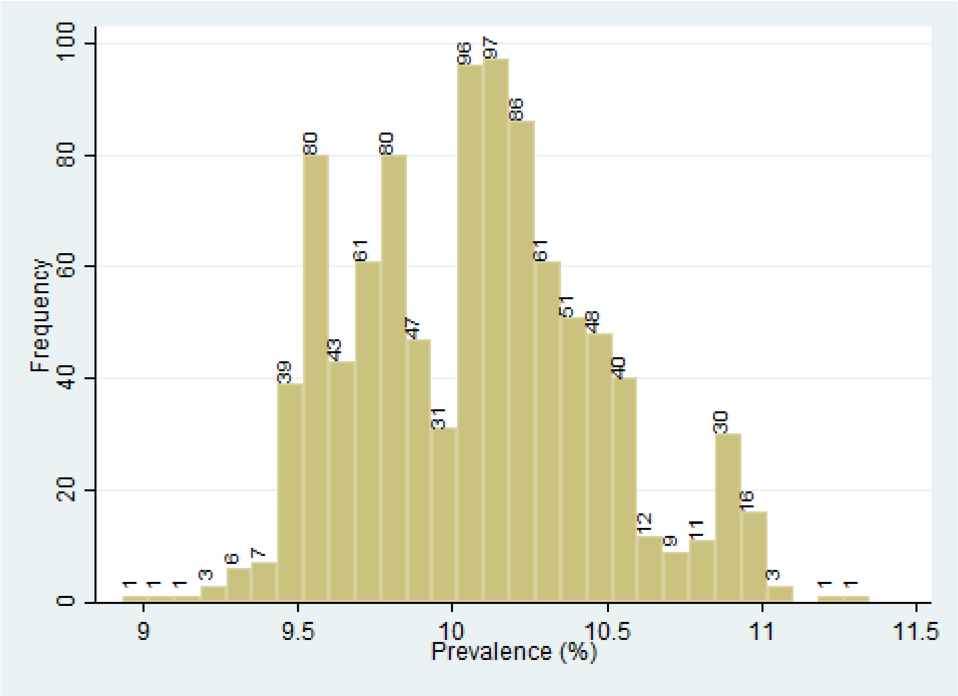


**Figure 25: Histogram of the prevalence of overall knee OA for Scotland at LA level**

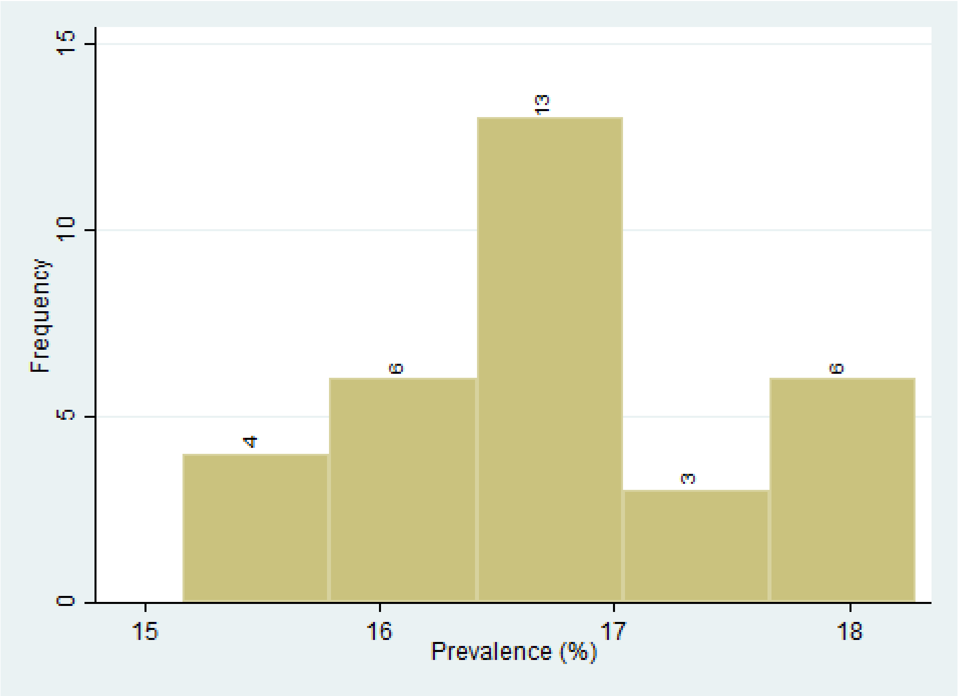


Figure 26: Histogram of the prevalence of overall knee OA for Scotland at practice level

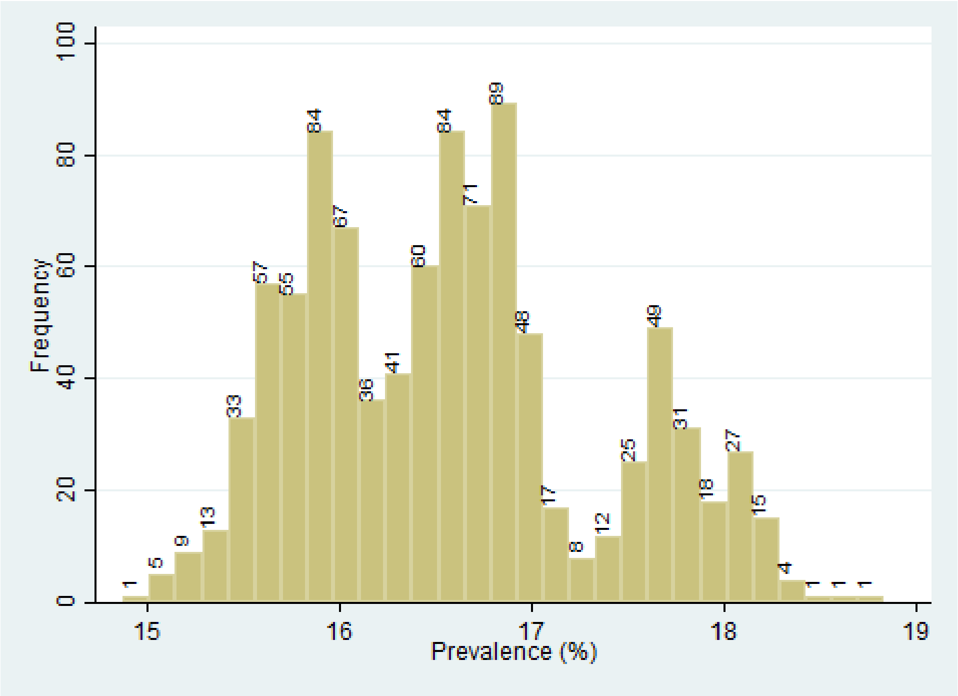


Figure 27: Histogram of the prevalence of severe hip OA for Scotland at practice level

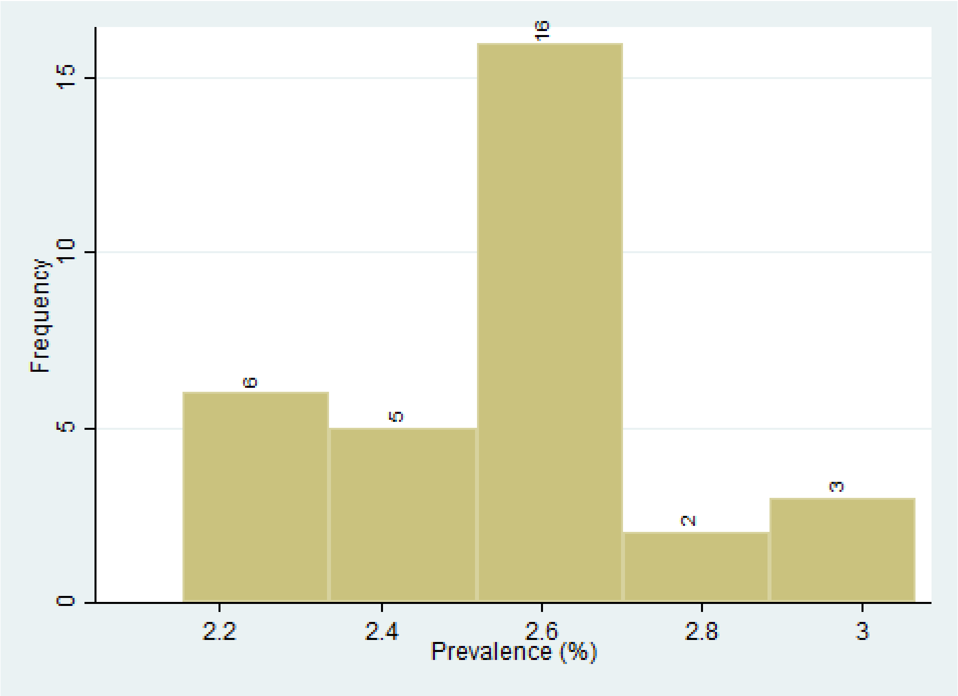


Figure 28: Histogram of the prevalence of severe hip OA for Scotland at practice level

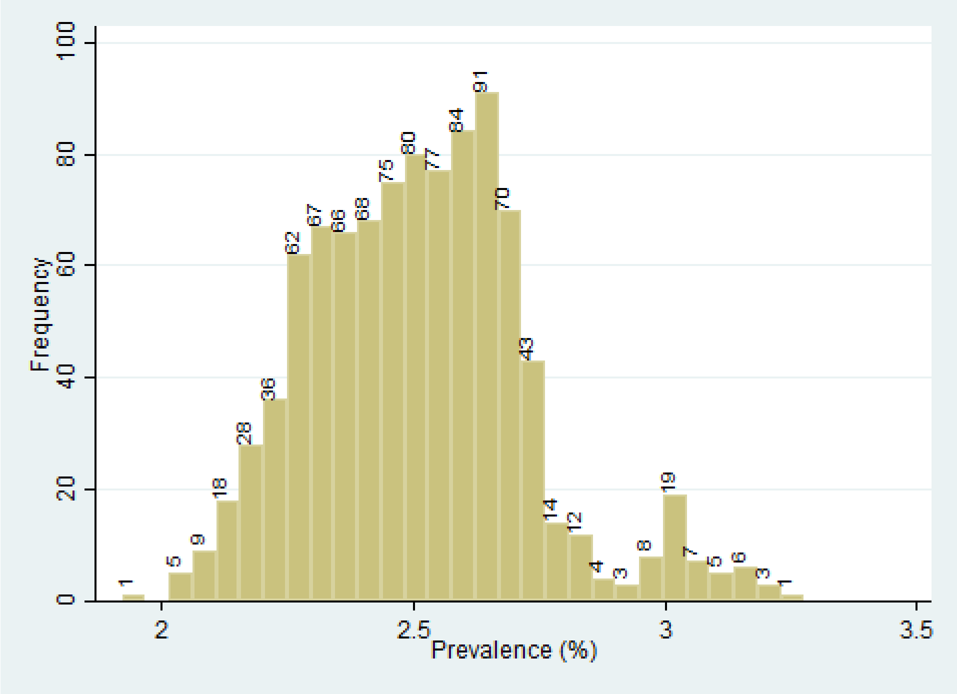


**Figure 29: Histogram of the prevalence of severe knee OA for Scotland at LA level**

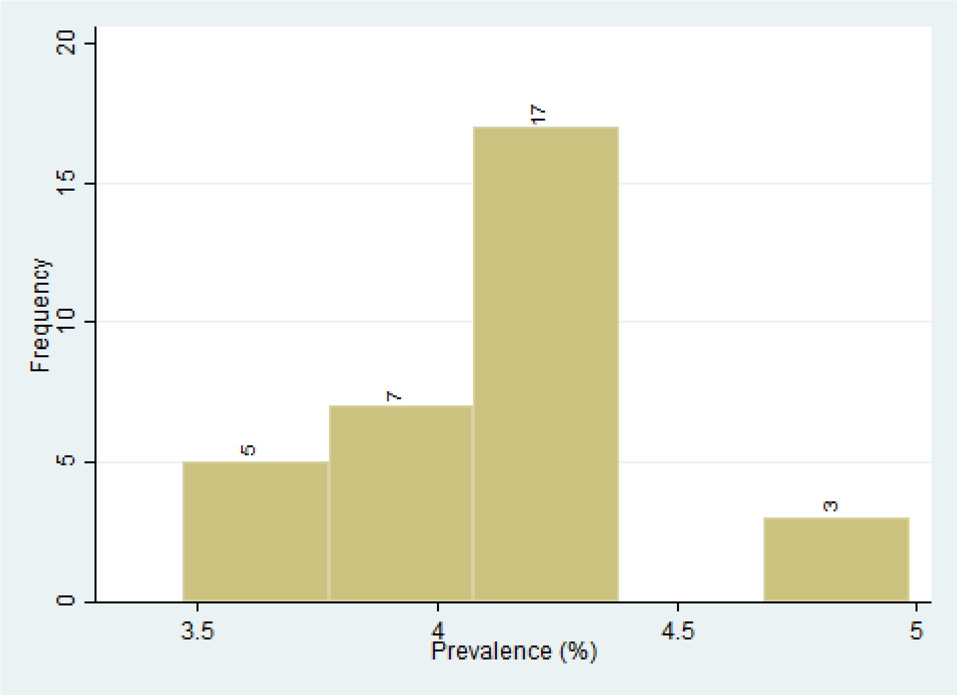
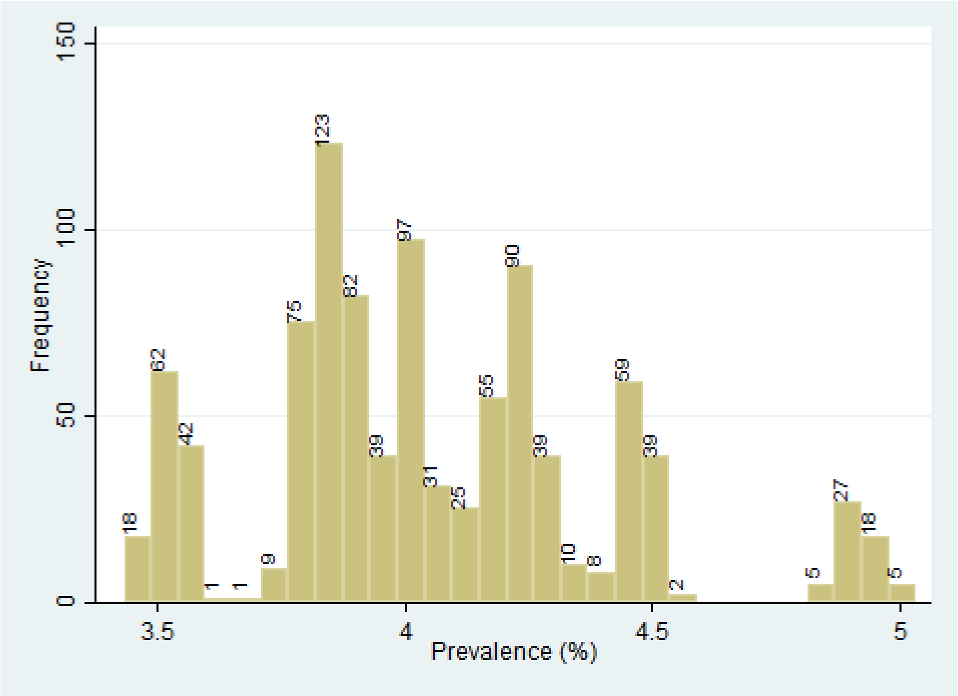


**Figure 30: Histogram of the prevalence of severe knee OA for Scotland at practice level**



# 6 Production of Wales local estimates

## 6.1 Methods

We were unable to produce MSK estimates for Wales during the main contract because lookup tables from GP practice populations to SOAs were unavailable. This meant we were unable to map key Census variables to GP practices and similarly unable to map practice variables to resident populations. For example, smoking data was available from practice QOF data, but because of the lack of a lookup table it could not be mapped to SOAs. However lookup tables became available in late 2017, so in 2018 we produced Wales. We used the model developed from ELSA to produce the prevalence of overall and severe hip OA, and overall and severe knee OA for Wales. However, risk factor data availability affected the risk factor variables included in the prediction model. Therefore, some variables were dropped from the English model because no local data was available. The performance of the models were compared by c-statistics (ROC curves).

## 6.2 Results

Gym membership and social economic status data were not available at any local levels in Wales. We removed these two variables from the final OA models and fitted on logistic regression model based on other available variables. The performance of the overall hip OA, overall knee OA, severe hip OA and severe knee OA models were listed below. However, the performances of the models between Wales and England tend to be similar.  Discrimination of the national/English and Welsh/local models is shown below.

**Figure 31: Overall hip OA English model**

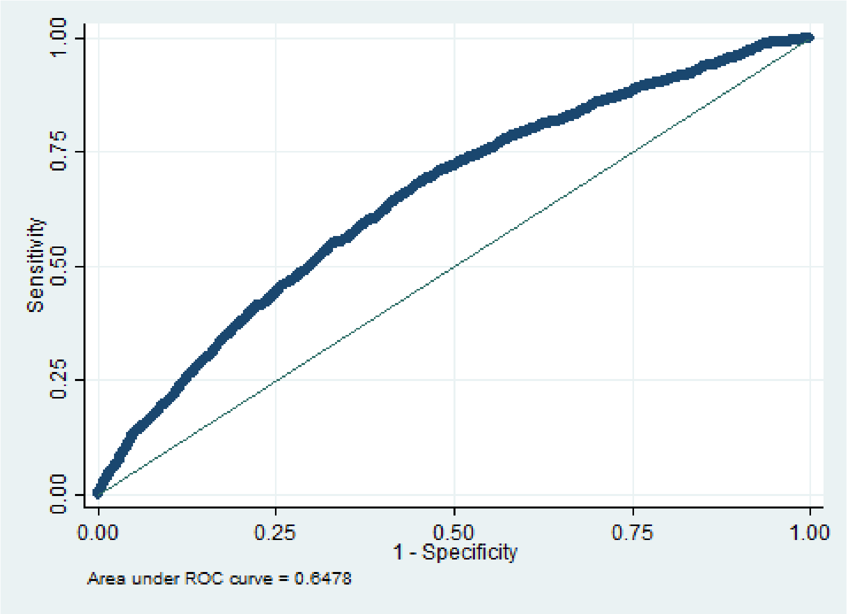


**Figure 32: Overall hip OA Welsh model**

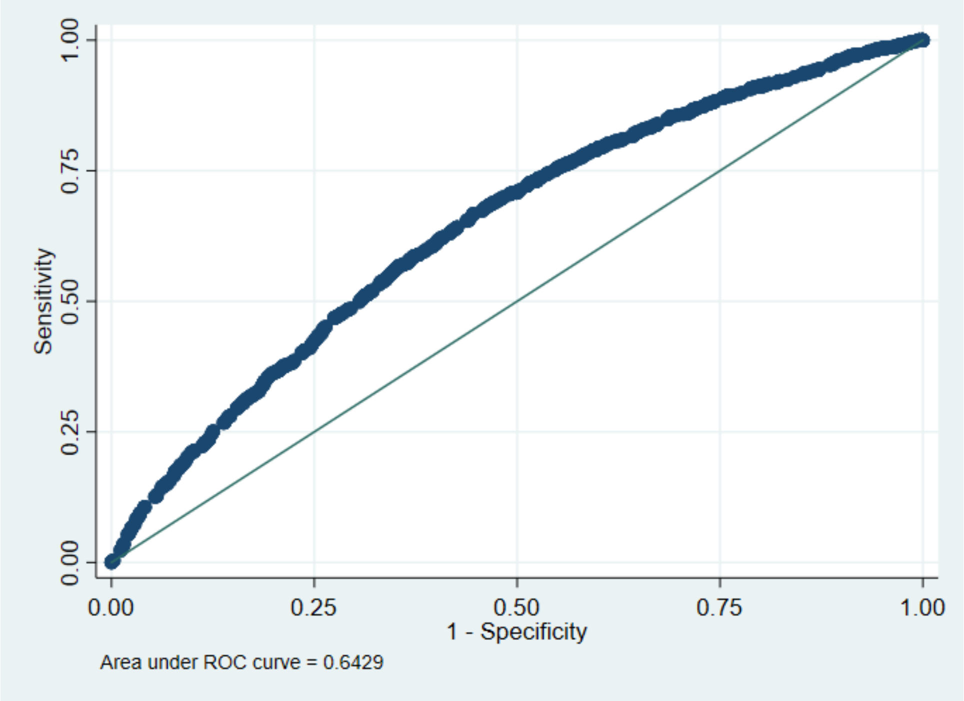


**Figure 33: Severe hip OA English model**

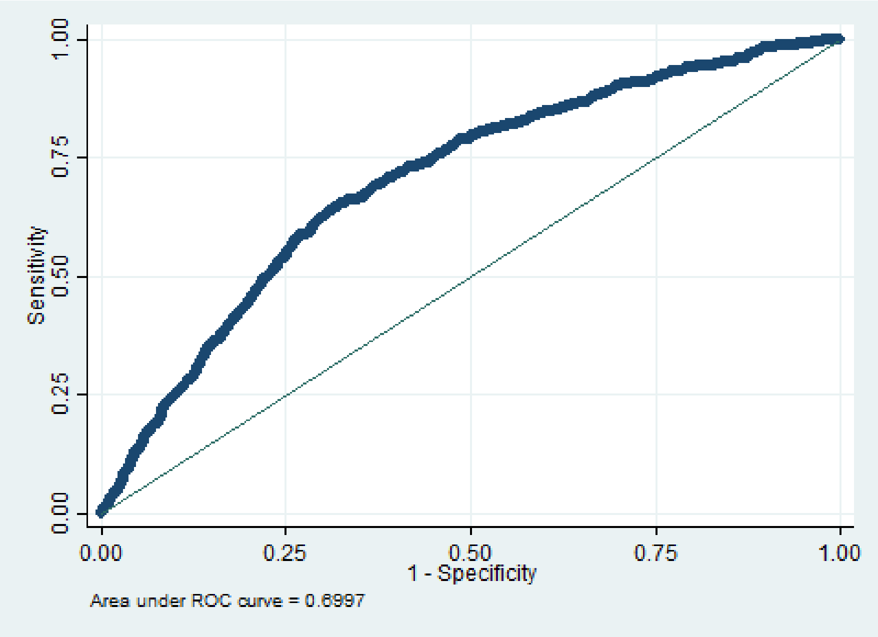


**Figure 34: Severe hip OA Welsh model**

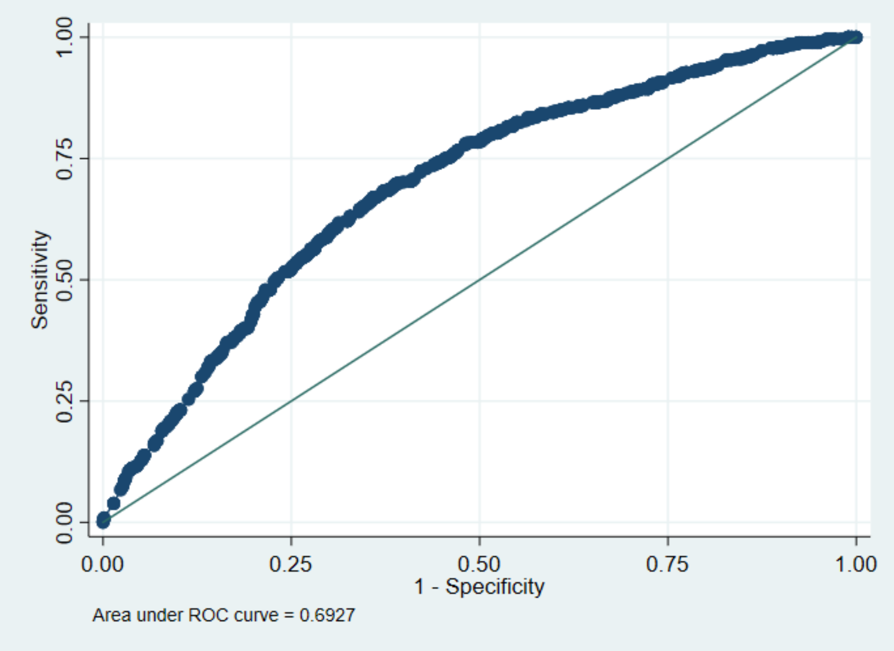


**Figure 35: General knee OA English model**

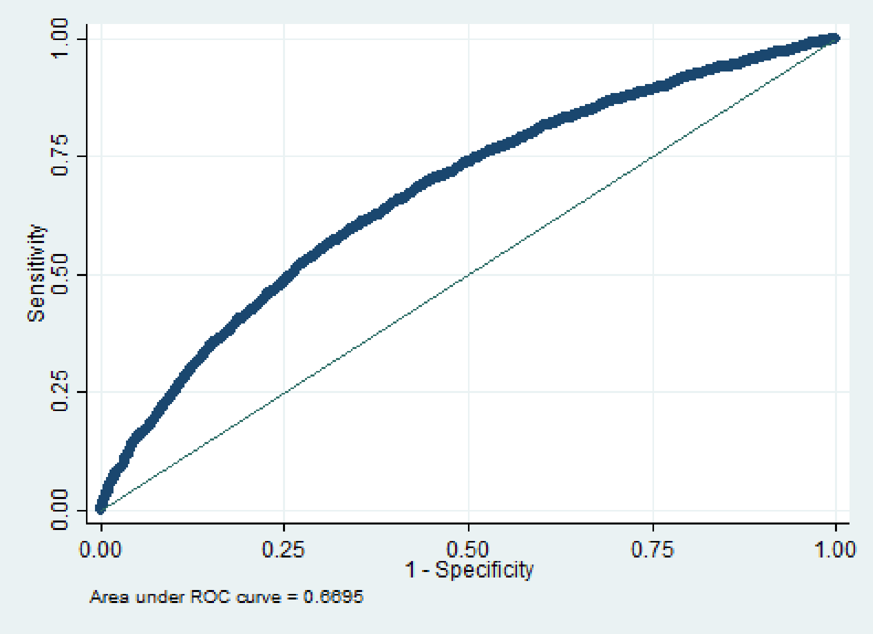


Figure 36: General knee OA Welsh model

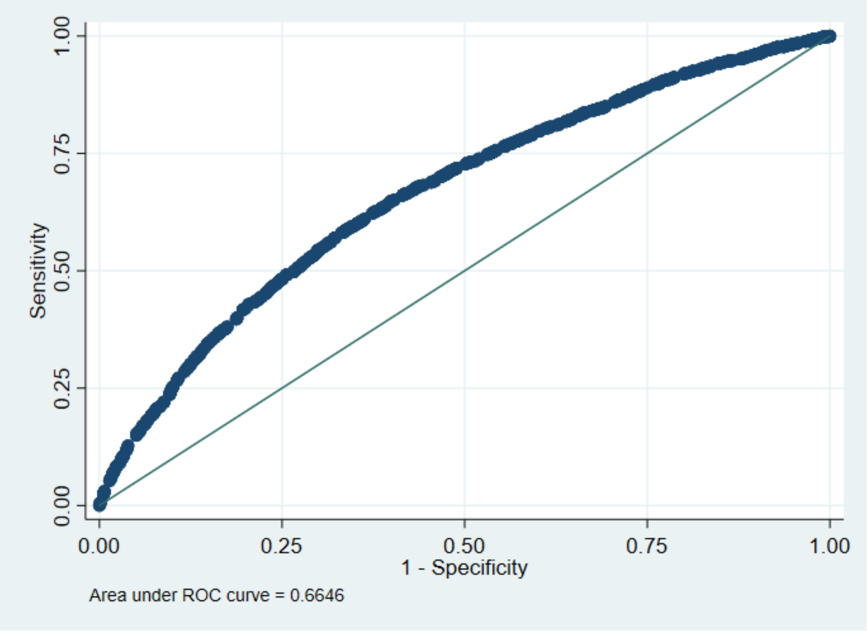
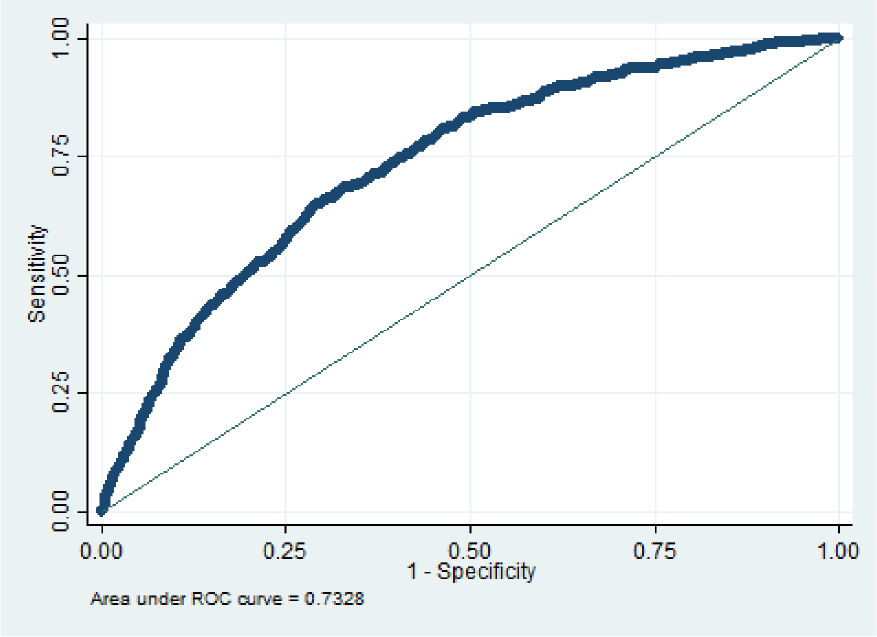
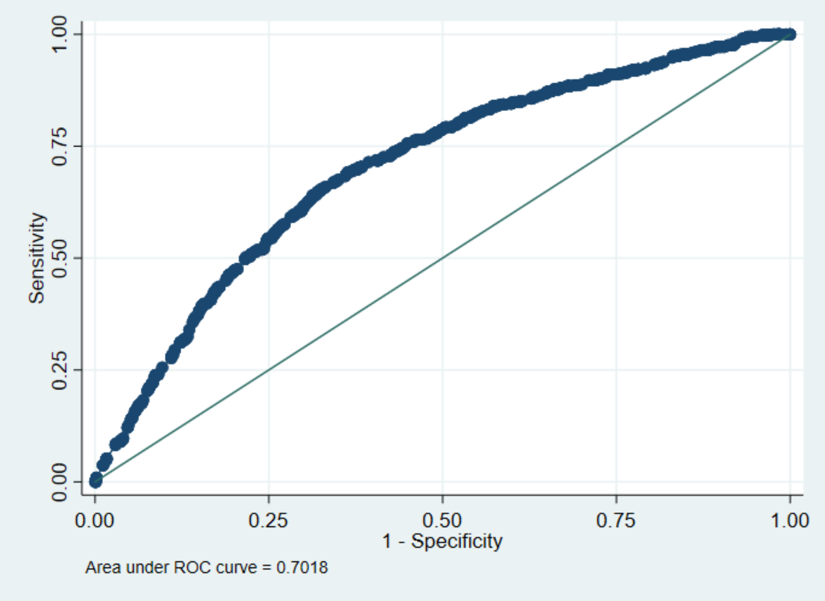


Figure 37: Severe knee OA English model



### Figure 38: Severe knee OA Welsh model



# 7 References

1. National Collaborating Centre for Chronic Conditions. Osteoarthritis: national clinical guideline for care and management in adults. NICE Clinical Guidelines. Updated: 13 August 2013 ed. London, 2008 Link: http://www.nice.org.uk/CG59.

2. Smith S. OA Nation 2012. The most comprehensive UK report of people with osteoarthritis. London: Arthritis Care, 2012 Link: http://www.arthritiscare.org.uk/LivingwithArthritis/oanation-2012.

3. Arthritis Research UK Primary Care Centre. Osteoarthritis in General Practice: Data & Perspectives. Keele University: Keele University, 2013 Link: http://www.arthritisresearchuk.org/policy-and-public-affairs/reports-and-resources/~/media/1C04D81E4C6048048E0246F588A0CD64.ashx.

4. Oliveria SA, Felson DT, Reed JI, Cirillo PA, Walker AM. Incidence of symptomatic hand, hip, and knee osteoarthritis among patients in a health maintenance organization. Arthritis & Rheumatism 2005;38:1134-41. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.1780380817/abstractfiles/750/abstract.html.

5. Oliveria SA, Felson DT, Reed JI, Cirillo PA, Walker AM. Incidence of symptomatic hand, hip, and knee osteoarthritis among patients in a health maintenance organization. Arthritis & Rheumatism 1995;38(8):1134-41. doi: 10.1002/art.1780380817 Link: http://dx.doi.org/10.1002/art.1780380817.

6. Wallace IJ, Worthington S, Felson DT, Jurmain RD, Wren KT, et al. Knee osteoarthritis has doubled in prevalence since the mid-20th century. Proceedings of the National Academy of Sciences 2017;Online publication doi: 10.1073/pnas.1703856114 [published Online First: 14/8/2017] Link: http://www.pnas.org/content/early/2017/08/08/1703856114.abstract.

7. Department of Health. National health expenditure data (2003-04 to 2010-11), 2012 Link: http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Managingyourorganisation/Financeandplanning/Programmebudgeting/DH_075743#_3.

8. McAlindon TE, Wilson PWF, Aliabadi P, Weissman B, Felson DT. Level of physical activity and the risk of radiographic and symptomatic knee osteoarthritis in the elderly: the Framingham study. The American journal of medicine 1999;106:151-57. Link: http://www.sciencedirect.com/science/article/pii/S0002934398004136files/804/S0002934398004136.html.

9. National Institute for Clinical & Public Health Excellence. Osteoarthritis: Care and management in adults: National Institute for Clinical & Public Health Excellence, 2014 Link: http://www.nice.org.uk/guidance/CG177.

10. Bannuru RR, Schmid CH, Kent DM, Vaysbrot EE, Wong JB, et al. Comparative Effectiveness of Pharmacologic Interventions for Knee OsteoarthritisA Systematic Review and Network Meta-analysisPharmacologic Interventions for Knee OA. Annals of Internal Medicine 2015;162(1):46-54. doi: 10.7326/M14-1231 Link: http://dx.doi.org/10.7326/M14-1231.

11. Blagojevic M, Jinks C, Jeffery A, Jordan KP. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. Osteoarthritis and Cartilage 2010;18(1):24-33. doi: http://dx.doi.org/10.1016/j.joca.2009.08.010 Link: http://www.sciencedirect.com/science/article/pii/S1063458409002258.

12. Hart DJ, Doyle DV, Spector TD. Incidence and risk factors for radiographic knee osteoarthritis in middle-aged women: the Chingford Study. Arthritis & Rheumatism 2001;42(1):17-24. Link: http://onlinelibrary.wiley.com/doi/10.1002/1529-0131(199901)42:1%3C17::AID-ANR2%3E3.0.CO;2-E/abstract
files/680/abstract.html.

13. Oliveria SA, Felson DT, Reed JI, Cirillo PA, Walker AM. Incidence of symptomatic hand, hip, and knee osteoarthritis among patients in a health maintenance organization. Arthritis & Rheumatism 2005;38(8):1134-41. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.1780380817/abstract
files/750/abstract.html.

14. Srikanth VK, Fryer JL, Zhai G, Winzenberg TM, Hosmer D, et al. A meta-analysis of sex differences prevalence, incidence and severity of osteoarthritis. Osteoarthritis and Cartilage 2005;13(9):769-81. doi: 10.1016/j.joca.2005.04.014 Link: http://www.sciencedirect.com/science/article/pii/S1063458405001123.

15. Holliday KL, McWilliams DF, Maciewicz RA, Muir KR, Zhang W, et al. Lifetime body mass index, other anthropometric measures of obesity and risk of knee or hip osteoarthritis in the GOAL case-control study. Osteoarthritis and Cartilage 2011;19(1):37-43. Link: http://www.sciencedirect.com/science/article/pii/S1063458410003559
files/760/S1063458410003559.html.

16. Coggon D, Reading I, Croft P, McLaren M, Barrett D, et al. Knee osteoarthritis and obesity. International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity 2001;25(5) Link: http://ukpmc.ac.uk/abstract/MED/11360143
files/754/11360143.html.

17. Cooper C, Snow S, McAlindon TE, Kellingray S, Stuart B, et al. Risk factors for the incidence and progression of radiographic knee osteoarthritis. Arthritis & Rheumatism 2000;43(5):995-1000. Link: http://onlinelibrary.wiley.com/doi/10.1002/1529-0131(200005)43:5%3C995::AID-ANR6%3E3.0.CO;2-1/abstract
files/756/abstract.html.

18. Sayer AA, Poole J, Cox V, Kuh D, Hardy R, et al. Weight from birth to 53 years: a longitudinal study of the influence on clinical hand osteoarthritis. Arthritis & Rheumatism 2003;48(4):1030-33. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.10862/full
files/758/full.html.

19. Wills AK, Black S, Cooper R, Coppack RJ, Hardy R, et al. Life course body mass index and risk of knee osteoarthritis at the age of 53 years: evidence from the 1946 British birth cohort study. Annals of the Rheumatic Diseases 2012;71(5):655-60. Link: http://ard.bmj.com/content/71/5/655.short
files/763/PMC3329229.html
files/762/655.html.

20. Lievense AM, Bierma-Zeinstra SMA, Verhagen AP, Van Baar ME, Verhaar JAN, et al. Influence of obesity on the development of osteoarthritis of the hip: a systematic review. Rheumatology 2002;41(10):1155-62. Link: http://rheumatology.oxfordjournals.org/content/41/10/1155.short
files/685/1155.html
files/684/1155.html.

21. Yusuf E, Nelissen RG, Ioan-Facsinay A, Stojanovic-Susulic V, DeGroot J, et al. Association between weight or body mass index and hand osteoarthritis: a systematic review. Annals of the Rheumatic Diseases 2010;69(4):761-65. Link: http://ard.bmj.com/content/69/4/761.short

files/769/761.html.

22. Stürmer T, Günther KP, Brenner H. Obesity, overweight and patterns of osteoarthritis: the Ulm Osteoarthritis Study. Journal of Clinical Epidemiology 2000;53(3) Link: http://ukpmc.ac.uk/abstract/MED/10760642
files/771/10760642.html.

23. Grotle M, Hagen KB, Natvig B, Dahl FA, Kvien TK. Obesity and osteoarthritis in knee, hip and/or hand: an epidemiological study in the general population with 10 years follow-up. BMC Musculoskeletal Disorders 2008;9(1) Link: http://www.biomedcentral.com/1471-2474/9/132/
files/773/132.html.

24. Jarvholm B, Lewold S, Malchau H, Vingard E. Age, bodyweight, smoking habits and the risk of severe osteoarthritis in the hip and knee in men. Eur J Epidemiol 2005;20(6):537-42. Link: <Go to ISI>://MEDLINE:16121763.

25. Tanamas S, Hanna FS, Cicuttini FM, Wluka AE, Berry P, et al. Does knee malalignment increase the risk of development and progression of knee osteoarthritis? A systematic review. Arthritis Care & Research 2009;61(4):459-67. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.24336/full
files/775/full.html.

26. Muthuri SG, McWilliams DF, Doherty M, Zhang W. History of knee injuries and knee osteoarthritis: a meta-analysis of observational studies. Osteoarthritis and Cartilage 2011;19(11):1286-93. Link: http://www.sciencedirect.com/science/article/pii/S1063458411002299
files/777/S1063458411002299.html.

27. Cooper C, Inskip H, Croft P, Campbell L, Smith G, et al. Individual risk factors for hip osteoarthritis: obesity, hip injury and physical activity. American Journal of Epidemiology 1998;147(6):516-22. Link: http://aje.oxfordjournals.org/content/147/6/516.short
files/780/516.html.

28. Toivanen AT, Heliövaara M, Impivaara O, Arokoski JPA, Knekt P, et al. Obesity, physically demanding work and traumatic knee injury are major risk factors for knee osteoarthritis—a population-based study with a follow-up of 22 years. Rheumatology 2010;49(2):308-14. doi: 10.1093/rheumatology/kep388 Link: http://rheumatology.oxfordjournals.org/content/49/2/308.abstract.

29. Øiestad BE, Engebretsen L, Storheim K, Risberg MA. Knee osteoarthritis after anterior cruciate ligament injury a systematic review. The American journal of sports medicine 2009;37(7):1434-43. Link: http://ajs.sagepub.com/content/37/7/1434.short
files/786/1434.html.

30. Lohmander LS, Östenberg A, Englund M, Roos H. High prevalence of knee osteoarthritis, pain, and functional limitations in female soccer players twelve years after anterior cruciate ligament injury. Arthritis & Rheumatism 2004;50(10):3145-52. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.20589/full
files/789/full.html.

31. Gelber AC, Hochberg MC, Mead LA, Wang NY, Wigley FM, et al. Joint injury in young adults and risk for subsequent knee and hip osteoarthritis. Annals of Internal Medicine 2000;133(5) Link: http://ukpmc.ac.uk/abstract/MED/10979876
files/791/10979876.html.

32. Hui M, Doherty M, Zhang W. Does smoking protect against osteoarthritis? Meta-analysis of observational studies. Annals of the Rheumatic Diseases 2011;70(7):1231-37. Link: http://ard.bmj.com/content/70/7/1231.short
files/793/1231.html.

33. Coggon D, Croft P, Kellingray S, Barrett D, McLaren M, et al. Occupational physical activities and osteoarthritis of the knee. Arthritis & Rheumatism 2001;43(7):1443-49. Link: http://onlinelibrary.wiley.com/doi/10.1002/1529-0131(200007)43:7%3C1443::AID-ANR5%3E3.0.CO;2-1/abstract
files/795/abstract.html.

34. Maetzel A, Mäkelä M, Hawker G, Bombardier C. Osteoarthritis of the hip and knee and mechanical occupational exposure–a systematic overview of the evidence. The Journal of Rheumatology 1997;24(8) Link: http://ukpmc.ac.uk/abstract/MED/9263158
files/797/9263158.html.

35. Lievense A, Bierma-Zeinstra S, Verhagen A, Verhaar JAN, Koes B. Influence of work on the development of osteoarthritis of the hip: a systematic review. The Journal of Rheumatology 2001;28(11):2520-28. Link: http://www.jrheum.org/content/28/11/2520.short
files/708/2520.html.

36. McWilliams DF, Leeb BF, Muthuri SG, Doherty M, Zhang W. Occupational risk factors for osteoarthritis of the knee: a meta-analysis. Osteoarthritis and Cartilage 2011;19(7):829-39. doi: http://dx.doi.org/10.1016/j.joca.2011.02.016 Link: http://www.sciencedirect.com/science/article/pii/S1063458411000677.

37. Spector TD, Harris PA, Hart DJ, Cicuttini FM, Nandra D, et al. Risk of osteoarthritis associated with long-term weight-bearing sports: a radiologic survey of the hips and knees in female ex-athletes and population controls. Arthritis & Rheumatism 2005;39(6):988-95. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.1780390616/abstract
files/802/abstract.html.

38. McAlindon TE, Wilson PWF, Aliabadi P, Weissman B, Felson DT. Level of physical activity and the risk of radiographic and symptomatic knee osteoarthritis in the elderly: the Framingham study. The American journal of medicine 1999;106(2):151-57. Link: http://www.sciencedirect.com/science/article/pii/S0002934398004136
files/804/S0002934398004136.html.

39. Felson DT, Niu J, Clancy M, Sack B, Aliabadi P, et al. Effect of recreational physical activities on the development of knee osteoarthritis in older adults of different weights: the Framingham Study. Arthritis Care & Research 2007;57(1):6-12. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.22464/full
files/806/full.html.

40. Kuijt MTK, Inklaar H, Gouttebarge V, Frings-Dresen MHW. Knee and ankle osteoarthritis in former elite soccer players: A systematic review of the recent literature. Journal of Science and Medicine in Sport 2012 Link: http://www.sciencedirect.com/science/article/pii/S1440244012000631files/808/S1440244012000631.html.

41. Dalstra JAA, Kunst AE, Borrell C, Breeze E, Cambois E, et al. Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries. International

journal of epidemiology 2005;34(2):316-26. Link: http://ije.oxfordjournals.org/content/34/2/316.short.

42. Theis KA, Murphy L, Hootman JM, Helmick CG, Yelin E. Prevalence and correlates of arthritis-attributable work limitation in the US population among persons ages 18–64: 2002 National Health Interview Survey Data. Arthritis Care & Research 2007;57(3):355-63. doi: 10.1002/art.22622 Link: http://dx.doi.org/10.1002/art.22622.

43. Spector TD, Nandra D, Hart DJ, Doyle DV. Is hormone replacement therapy protective for hand and knee osteoarthritis in women? The Chingford study. Annals of the Rheumatic Diseases 1997;56(7):432-34. doi: Doi 10.1136/Ard.56.7.432 Link: <Go to ISI>://A1997XP59700009.

44. Blagojevic M, Jinks C, Jeffery A, Jordan KP. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. Osteoarthritis and Cartilage 2010;18:24-33. doi: http://dx.doi.org/10.1016/j.joca.2009.08.010 Link: http://www.sciencedirect.com/science/article/pii/S1063458409002258.

45. Hart DJ, Doyle DV, Spector TD. Incidence and risk factors for radiographic knee osteoarthritis in middle-aged women: the Chingford Study. Arthritis & Rheumatism 2001;42:17-24. Link: http://onlinelibrary.wiley.com/doi/10.1002/1529-0131(199901)42:1%3C17::AID-ANR2%3E3.0.CO;2-E/abstractfiles/680/abstract.html.

46. Srikanth VK, Fryer JL, Zhai G, Winzenberg TM, Hosmer D, et al. A meta-analysis of sex differences prevalence, incidence and severity of osteoarthritis. Osteoarthritis and Cartilage 2005;13:769-81. doi: 10.1016/j.joca.2005.04.014 Link: http://www.sciencedirect.com/science/article/pii/S1063458405001123.

47. Felson DT, Zhang Y, Anthony JM, Naimark A, Anderson JJ. Weight loss reduces the risk for symptomatic knee osteoarthritis in women. The Framingham Study. Annals of Internal Medicine 1992;116(7):535-9. Link.

48. Toivanen AT, Heliövaara M, Impivaara O, Arokoski JPA, Knekt P, et al. Obesity, physically demanding work and traumatic knee injury are major risk factors for knee osteoarthritis—a population-based study with a follow-up of 22 years. Rheumatology 2010;49:308-14. doi: 10.1093/rheumatology/kep388 Link: http://rheumatology.oxfordjournals.org/content/49/2/308.abstract.

49. Hui M, Doherty M, Zhang W. Does smoking protect against osteoarthritis? Meta-analysis of observational studies. Annals of the Rheumatic Diseases 2011;70:1231-37. Link: http://ard.bmj.com/content/70/7/1231.shortfiles/793/1231.html.

50. Coggon D, Croft P, Kellingray S, Barrett D, McLaren M, et al. Occupational physical activities and osteoarthritis of the knee. Arthritis & Rheumatism 2001;43:1443-49. Link: http://onlinelibrary.wiley.com/doi/10.1002/1529-0131(200007)43:7%3C1443::AID-ANR5%3E3.0.CO;2-1/abstractfiles/795/abstract.html.

51. Felson DT, Niu J, Clancy M, Sack B, Aliabadi P, et al. Effect of recreational physical activities on the development of knee osteoarthritis in older adults of different weights: the Framingham Study. Arthritis Care & Research 2007;57:6-12. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.22464/fullfiles/806/full.html.

52. Spector TD, Harris PA, Hart DJ, Cicuttini FM, Nandra D, et al. Risk of osteoarthritis associated with long-term weight-bearing sports: a radiologic survey of the hips and knees in female ex-athletes

and population controls. Arthritis & Rheumatism 2005;39:988-95. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.1780390616/abstractfiles/802/abstract.html.

53. Roos EM, Dahlberg L. Positive effects of moderate exercise on glycosaminoglycan content in knee cartilage: a four-month, randomized, controlled trial in patients at risk of osteoarthritis. Arthritis and rheumatism 2005;52(11):3507-14. doi: 10.1002/art.21415 [published Online First: 2005/11/01] Link: http://www.ncbi.nlm.nih.gov/pubmed/16258919.

54. Department of Health. Long Term Conditions Compendium of Information: Third Edition. Long Term Conditions Compendium of Information, 2012 Link: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216528/dh_134486.pdf.

55. Health & Social Care Information Centre. Health Survey for England 2011 Volume2 - Methods and documentation, 2012 Link.

56. Dalstra JAA, Kunst AE, Borrell C, Breeze E, Cambois E, et al. Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries. International journal of epidemiology 2005;34:316-26. Link: http://ije.oxfordjournals.org/content/34/2/316.short.

57. UCL Research Department of Epidemiology and Public Health, Institute for Fiscal Studies, NatCen Social Research, The University of Manchester SoSS. English Longitudinal Study of Ageing: About ELSA, 2012 Link: http://www.ifs.org.uk/ELSA/about.

58. English Longitudinal Study of Ageing. Documentation 2014 [Available from: http://www.elsa-project.ac.uk/documentation accessed 06 February 2014 2014.

59. Hensor EMA, Dube B, Kingsbury SR, Tennant A, Conaghan PG. Towards a clinical definition of early osteoarthritis: Onset of patient-reported knee pain begins on stairs – data from the osteoarthritis initiative. Arthritis Care & Research 2014:n/a-n/a. doi: 10.1002/acr.22418 Link: http://dx.doi.org/10.1002/acr.22418.

60. Kirkwood B R SJAC. Regression modelling. In: K M, ed. Medical Statistics. USA: Blackwell Publishing company 2003:339-42.

61. Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, et al. Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System. American Journal of Epidemiology 2014;179(8):1025-33. doi: 10.1093/aje/kwu018 Link: http://aje.oxfordjournals.org/content/179/8/1025.abstract.

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

63. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36. Link: <Go to ISI>://MEDLINE:7063747.

64. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 2007;115(5):654-7. Link: http://circ.ahajournals.org/content/115/5/654.

65. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal 2014;35(29):1925-31. doi:

10.1093/eurheartj/ehu207 Link: http://eurheartj.oxfordjournals.org/content/35/29/1925.abstract.

66. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839-43. Link: <Go to ISI>://MEDLINE:6878708.

67. Arthritis Research UK. Osteoarthritis in general practice, 2013 Link: http://www.arthritisresearchuk.org/arthritis-information/data-and-statistics/osteoarthritis.aspx.

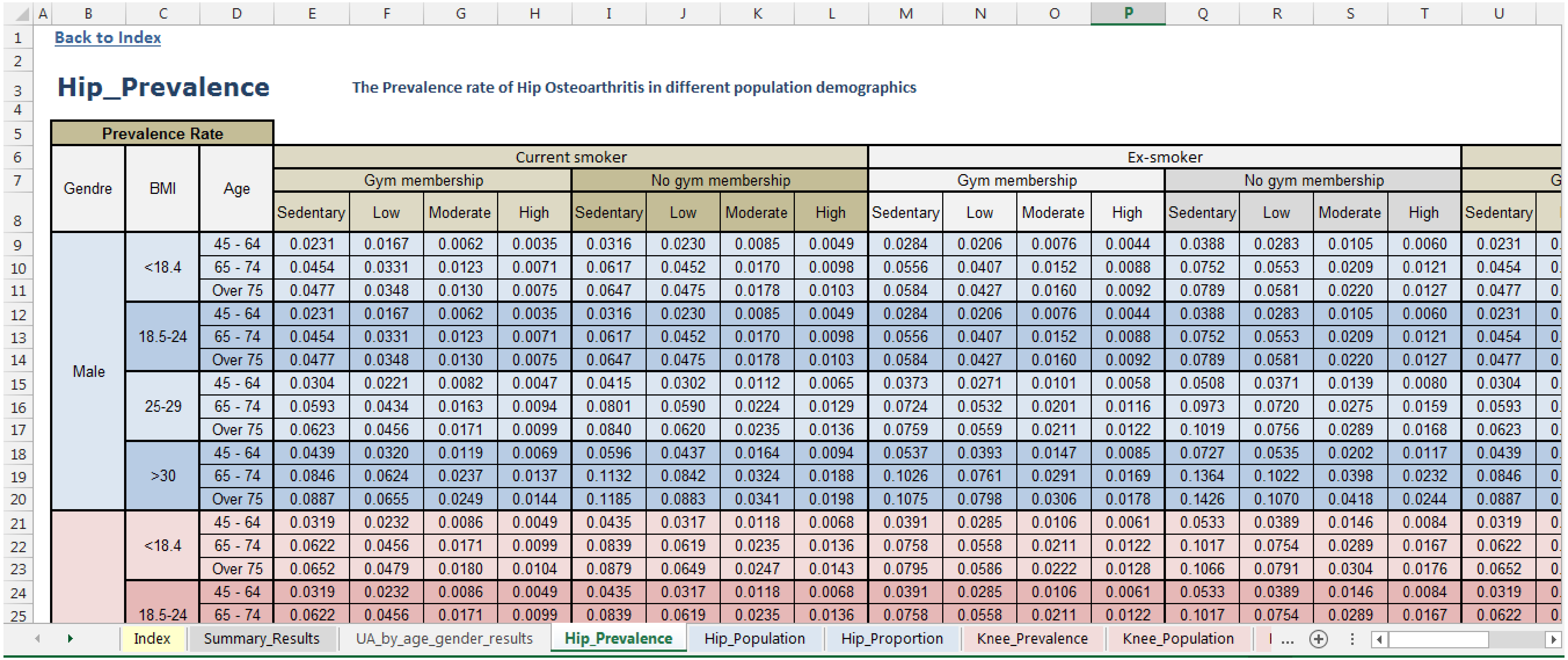
68. Newson R. Confidence intervals for rank statistics: Somers' D and extensions. Stata Journal 2006;6(3):309-34. Link: http://www.stata-journal.com/article.html?article=snp15_6.

69. Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Statistics in Medicine 2007;26:1802-11. doi: 10.1002/sim.2801 Link: http://dx.doi.org/10.1002/sim.2801.

70. Sánchez BN, Raghunathan TE, Diez Roux AV, Zhu Y, Lee O. Combining data from primary and ancillary surveys to assess the association between neighborhood-level characteristics and health outcomes: the Multi-Ethnic Study of Artherosclerosis. Statistics in Medicine 2008;27:5745-63. doi: 10.1002/sim.3384 Link: http://dx.doi.org/10.1002/sim.3384.

# 8 Appendix 1: spreadsheets of prevalence rates for population subcategories

Figure 39: prevalence estimates for population subcategories from national data



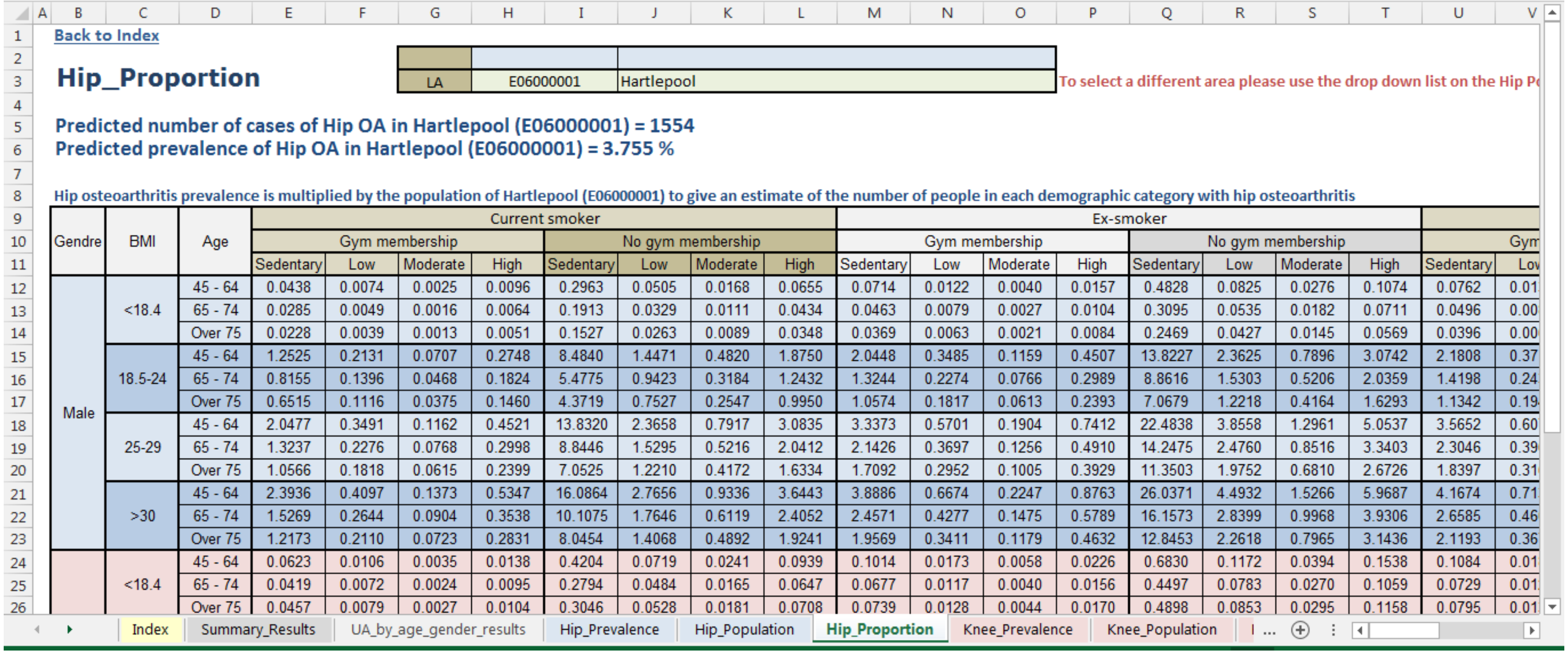
Back to Index

Hip_Prevalence

The Prevalence rate of Hip Osteoarthritis in different population demographics

| Prevalence Rate | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Current smoker | | | | | | | | Ex-smoker | | | | | | | | | |
| | | | Gym membership | | | | No gym membership | | | | Gym membership | | | | No gym membership | | | | G | |
| Gendre | BMI | Age | Sedentary | Low | Moderate | High | Sedentary | Low | Moderate | High | Sedentary | Low | Moderate | High | Sedentary | Low | Moderate | High | Sedentary | |
| Male | <18.4 | 45 - 64 | 0.0231 | 0.0167 | 0.0062 | 0.0035 | 0.0316 | 0.0230 | 0.0085 | 0.0049 | 0.0284 | 0.0206 | 0.0076 | 0.0044 | 0.0388 | 0.0283 | 0.0105 | 0.0060 | 0.0231 | 0. |
| | | 65 - 74 | 0.0454 | 0.0331 | 0.0123 | 0.0071 | 0.0617 | 0.0452 | 0.0170 | 0.0098 | 0.0556 | 0.0407 | 0.0152 | 0.0088 | 0.0752 | 0.0553 | 0.0209 | 0.0121 | 0.0454 | 0. |
| | | Over 75 | 0.0477 | 0.0348 | 0.0130 | 0.0075 | 0.0647 | 0.0475 | 0.0178 | 0.0103 | 0.0584 | 0.0427 | 0.0160 | 0.0092 | 0.0789 | 0.0581 | 0.0220 | 0.0127 | 0.0477 | 0. |
| | 18.5-24 | 45 - 64 | 0.0231 | 0.0167 | 0.0062 | 0.0035 | 0.0316 | 0.0230 | 0.0085 | 0.0049 | 0.0284 | 0.0206 | 0.0076 | 0.0044 | 0.0388 | 0.0283 | 0.0105 | 0.0060 | 0.0231 | 0. |
| | | 65 - 74 | 0.0454 | 0.0331 | 0.0123 | 0.0071 | 0.0617 | 0.0452 | 0.0170 | 0.0098 | 0.0556 | 0.0407 | 0.0152 | 0.0088 | 0.0752 | 0.0553 | 0.0209 | 0.0121 | 0.0454 | 0. |
| | | Over 75 | 0.0477 | 0.0348 | 0.0130 | 0.0075 | 0.0647 | 0.0475 | 0.0178 | 0.0103 | 0.0584 | 0.0427 | 0.0160 | 0.0092 | 0.0789 | 0.0581 | 0.0220 | 0.0127 | 0.0477 | 0. |
| | 25-29 | 45 - 64 | 0.0304 | 0.0221 | 0.0082 | 0.0047 | 0.0415 | 0.0302 | 0.0112 | 0.0065 | 0.0373 | 0.0271 | 0.0101 | 0.0058 | 0.0508 | 0.0371 | 0.0139 | 0.0080 | 0.0304 | 0. |
| | | 65 - 74 | 0.0593 | 0.0434 | 0.0163 | 0.0094 | 0.0801 | 0.0590 | 0.0224 | 0.0129 | 0.0724 | 0.0532 | 0.0201 | 0.0116 | 0.0973 | 0.0720 | 0.0275 | 0.0159 | 0.0593 | 0. |
| | | Over 75 | 0.0623 | 0.0456 | 0.0171 | 0.0099 | 0.0840 | 0.0620 | 0.0235 | 0.0136 | 0.0759 | 0.0559 | 0.0211 | 0.0122 | 0.1019 | 0.0756 | 0.0289 | 0.0168 | 0.0623 | 0. |
| | >30 | 45 - 64 | 0.0439 | 0.0320 | 0.0119 | 0.0069 | 0.0596 | 0.0437 | 0.0164 | 0.0094 | 0.0537 | 0.0393 | 0.0147 | 0.0085 | 0.0727 | 0.0535 | 0.0202 | 0.0117 | 0.0439 | 0. |
| | | 65 - 74 | 0.0846 | 0.0624 | 0.0237 | 0.0137 | 0.1132 | 0.0842 | 0.0324 | 0.0188 | 0.1026 | 0.0761 | 0.0291 | 0.0169 | 0.1364 | 0.1022 | 0.0398 | 0.0232 | 0.0846 | 0. |
| | | Over 75 | 0.0887 | 0.0655 | 0.0249 | 0.0144 | 0.1185 | 0.0883 | 0.0341 | 0.0198 | 0.1075 | 0.0798 | 0.0306 | 0.0178 | 0.1426 | 0.1070 | 0.0418 | 0.0244 | 0.0887 | 0. |
| | <18.4 | 45 - 64 | 0.0319 | 0.0232 | 0.0086 | 0.0049 | 0.0435 | 0.0317 | 0.0118 | 0.0068 | 0.0391 | 0.0285 | 0.0106 | 0.0061 | 0.0533 | 0.0389 | 0.0146 | 0.0084 | 0.0319 | 0. |
| | | 65 - 74 | 0.0622 | 0.0456 | 0.0171 | 0.0099 | 0.0839 | 0.0619 | 0.0235 | 0.0136 | 0.0758 | 0.0558 | 0.0211 | 0.0122 | 0.1017 | 0.0754 | 0.0289 | 0.0167 | 0.0622 | 0. |
| | | Over 75 | 0.0652 | 0.0479 | 0.0180 | 0.0104 | 0.0879 | 0.0649 | 0.0247 | 0.0143 | 0.0795 | 0.0586 | 0.0222 | 0.0128 | 0.1066 | 0.0791 | 0.0304 | 0.0176 | 0.0652 | 0. |
| | 18.5-24 | 45 - 64 | 0.0319 | 0.0232 | 0.0086 | 0.0049 | 0.0435 | 0.0317 | 0.0118 | 0.0068 | 0.0391 | 0.0285 | 0.0106 | 0.0061 | 0.0533 | 0.0389 | 0.0146 | 0.0084 | 0.0319 | 0. |
| | | 65 - 74 | 0.0622 | 0.0456 | 0.0171 | 0.0099 | 0.0839 | 0.0619 | 0.0235 | 0.0136 | 0.0758 | 0.0558 | 0.0211 | 0.0122 | 0.1017 | 0.0754 | 0.0289 | 0.0167 | 0.0622 | 0. |

## Figure 40: local estimates for a local population (MLSOA in Hartlepool )



Back to Index

### Hip_Proportion

| | | |
|---|---|---|
| LA | E06000001 | Hartlepool |

To select a different area please use the drop down list on the Hip P

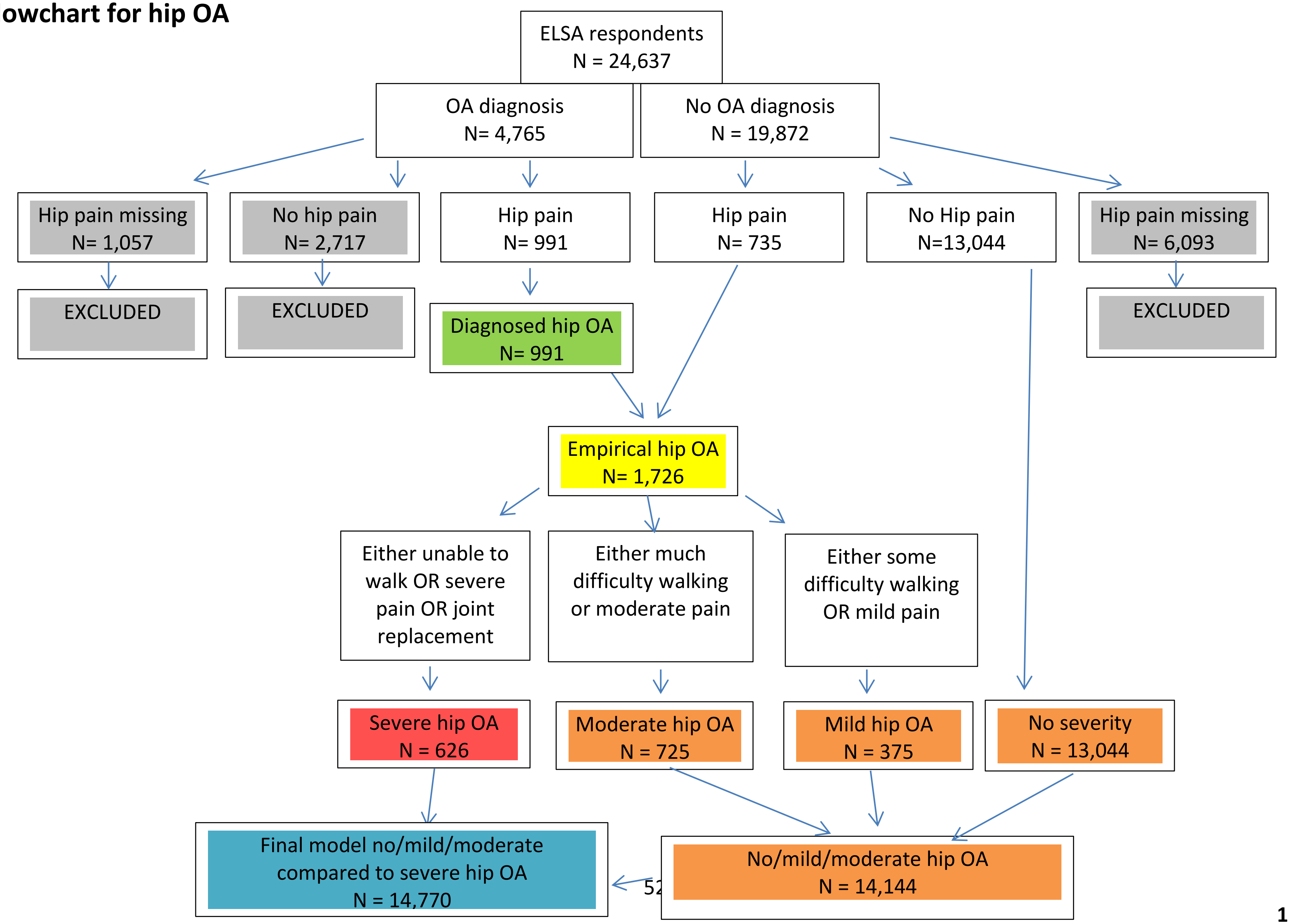
**Predicted number of cases of Hip OA in Hartlepool (E06000001) = 1554**
**Predicted prevalence of Hip OA in Hartlepool (E06000001) = 3.755 %**

Hip osteoarthritis prevalence is multiplied by the population of Hartlepool (E06000001) to give an estimate of the number of people in each demographic category with hip osteoarthritis

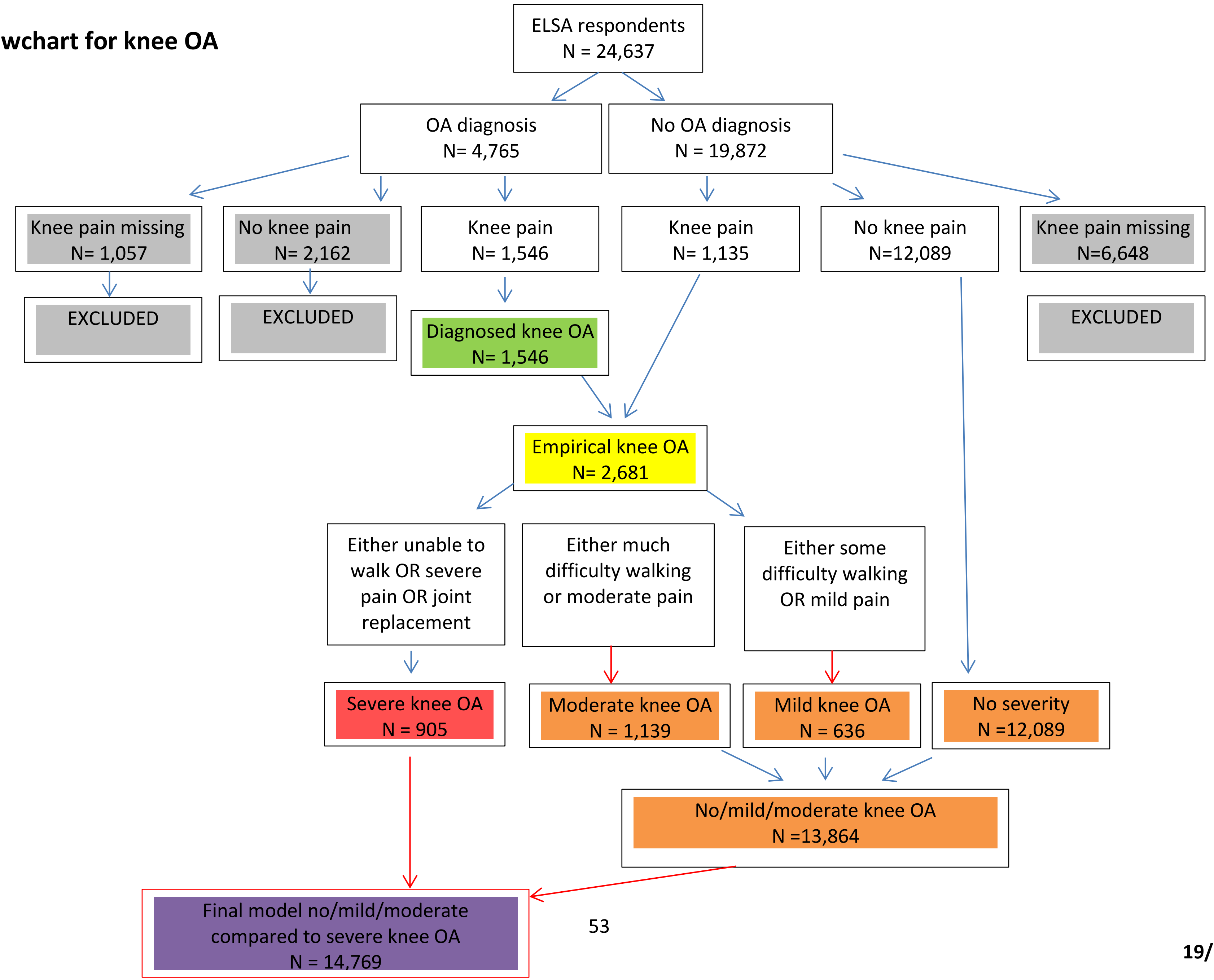
| Gendre | BMI | Age | Current smoker | | | | | | | | Ex-smoker | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Gym membership | | | | No gym membership | | | | Gym membership | | | | No gym membership | | | | Gym | |
| | | | Sedentary | Low | Moderate | High | Sedentary | Low | Moderate | High | Sedentary | Low | Moderate | High | Sedentary | Low | Moderate | High | Sedentary | Lo |
| Male | <18.4 | 45 - 64 | 0.0438 | 0.0074 | 0.0025 | 0.0096 | 0.2963 | 0.0505 | 0.0168 | 0.0655 | 0.0714 | 0.0122 | 0.0040 | 0.0157 | 0.4828 | 0.0825 | 0.0276 | 0.1074 | 0.0762 | 0.01 |
| | | 65 - 74 | 0.0285 | 0.0049 | 0.0016 | 0.0064 | 0.1913 | 0.0329 | 0.0111 | 0.0434 | 0.0463 | 0.0079 | 0.0027 | 0.0104 | 0.3095 | 0.0535 | 0.0182 | 0.0711 | 0.0496 | 0.00 |
| | | Over 75 | 0.0228 | 0.0039 | 0.0013 | 0.0051 | 0.1527 | 0.0263 | 0.0089 | 0.0348 | 0.0369 | 0.0063 | 0.0021 | 0.0084 | 0.2469 | 0.0427 | 0.0145 | 0.0569 | 0.0396 | 0.00 |
| | 18.5-24 | 45 - 64 | 1.2525 | 0.2131 | 0.0707 | 0.2748 | 8.4840 | 1.4471 | 0.4820 | 1.8750 | 2.0448 | 0.3485 | 0.1159 | 0.4507 | 13.8227 | 2.3625 | 0.7896 | 3.0742 | 2.1808 | 0.37 |
| | | 65 - 74 | 0.8155 | 0.1396 | 0.0468 | 0.1824 | 5.4775 | 0.9423 | 0.3184 | 1.2432 | 1.3244 | 0.2274 | 0.0766 | 0.2989 | 8.8616 | 1.5303 | 0.5206 | 2.0359 | 1.4198 | 0.24 |
| | | Over 75 | 0.6515 | 0.1116 | 0.0375 | 0.1460 | 4.3719 | 0.7527 | 0.2547 | 0.9950 | 1.0574 | 0.1817 | 0.0613 | 0.2393 | 7.0679 | 1.2218 | 0.4164 | 1.6293 | 1.1342 | 0.19 |
| | 25-29 | 45 - 64 | 2.0477 | 0.3491 | 0.1162 | 0.4521 | 13.8320 | 2.3658 | 0.7917 | 3.0835 | 3.3373 | 0.5701 | 0.1904 | 0.7412 | 22.4838 | 3.8558 | 1.2961 | 5.0537 | 3.5652 | 0.60 |
| | | 65 - 74 | 1.3237 | 0.2276 | 0.0768 | 0.2998 | 8.8446 | 1.5295 | 0.5216 | 2.0412 | 2.1426 | 0.3697 | 0.1256 | 0.4910 | 14.2475 | 2.4760 | 0.8516 | 3.3403 | 2.3046 | 0.39 |
| | | Over 75 | 1.0566 | 0.1818 | 0.0615 | 0.2399 | 7.0525 | 1.2210 | 0.4172 | 1.6334 | 1.7092 | 0.2952 | 0.1005 | 0.3929 | 11.3503 | 1.9752 | 0.6810 | 2.6726 | 1.8397 | 0.31 |
| | >30 | 45 - 64 | 2.3936 | 0.4097 | 0.1373 | 0.5347 | 16.0864 | 2.7656 | 0.9336 | 3.6443 | 3.8886 | 0.6674 | 0.2247 | 0.8763 | 26.0371 | 4.4932 | 1.5266 | 5.9687 | 4.1674 | 0.71 |
| | | 65 - 74 | 1.5269 | 0.2644 | 0.0904 | 0.3538 | 10.1075 | 1.7646 | 0.6119 | 2.4052 | 2.4571 | 0.4277 | 0.1475 | 0.5789 | 16.1573 | 2.8399 | 0.9968 | 3.9306 | 2.6585 | 0.46 |
| | | Over 75 | 1.2173 | 0.2110 | 0.0723 | 0.2831 | 8.0454 | 1.4068 | 0.4892 | 1.9241 | 1.9569 | 0.3411 | 0.1179 | 0.4632 | 12.8453 | 2.2618 | 0.7965 | 3.1436 | 2.1193 | 0.36 |
| | <18.4 | 45 - 64 | 0.0623 | 0.0106 | 0.0035 | 0.0138 | 0.4204 | 0.0719 | 0.0241 | 0.0939 | 0.1014 | 0.0173 | 0.0058 | 0.0226 | 0.6830 | 0.1172 | 0.0394 | 0.1538 | 0.1084 | 0.01 |
| | | 65 - 74 | 0.0419 | 0.0072 | 0.0024 | 0.0095 | 0.2794 | 0.0484 | 0.0165 | 0.0647 | 0.0677 | 0.0117 | 0.0040 | 0.0156 | 0.4497 | 0.0783 | 0.0270 | 0.1059 | 0.0729 | 0.01 |
| | | Over 75 | 0.0457 | 0.0079 | 0.0027 | 0.0104 | 0.3046 | 0.0528 | 0.0181 | 0.0708 | 0.0739 | 0.0128 | 0.0044 | 0.0170 | 0.4898 | 0.0853 | 0.0295 | 0.1158 | 0.0795 | 0.01 |

# 9 Appendix 2: ELSA data flowcharts

## 9.1 Flowchart for hip OA



## 9.2 Flowchart for knee OA



# 10 Appendix 3: synthetic estimation using Stata software

For the purposes of this example using Stata, we shall use the chosen logistic regression predictive model for severe knee OA.

## 10.1 Synthetic estimates

The proportion of our population according to age and sex are known. The proportion by educational status can be applied to these numbers, taking account of the fact that the distribution by educational status differs by age group. This gives estimated proportion by age, sex and educational status. This information is reflected in the variables below (variable names starting m_noed_, m_othed, m_nqv_, f_noed, f_othed, f_nqv).

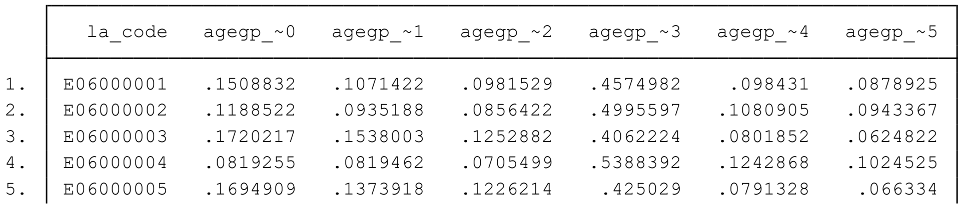
Within stata, a new set of variables is created, one for each combination of these risk factors pertinent to the logistic regression model for the chosen disease. For instance, if there are two binary variables for age group included in the regression model, then there are three relevant age groups (those with the first variable=1, those with the second variable=1, and those where both variables=0 – it is not possible to have both variables =1 since this would imply being in two separate age groups at the same time). With a binary variable for gender included, we would need groups for each gender – but some models don't include gender, like the one here we are using in this illustration. With one binary variable for educational status included in our predictive model, there are 2 categories for education (those with and those without this specified educational status, which here is no qualifications). The total number of combinations of age/ sex/ education groups then becomes 3x1x2=6. Corresponding to these 6 categories we have 6 variables as follows, which are created by summing sub categories (the categories that have equivalent risk within the model in question):

```
gen agegp_23_edu7_1_0=m_noed_4564+ f_noed_4564
gen agegp_23_edu7_1_1=m_noed_6574+ f_noed_6574
gen agegp_23_edu7_1_2=m_noed_75p+ f_noed_75p
gen agegp_23_edu7_1_3=f_othed_4564 +f_nvq_4564+ m_othed_4564+ m_nvq_4564
gen agegp_23_edu7_1_4=f_othed_6574 +f_nvq_6574+ m_othed_6574+ m_nvq_6574
gen agegp_23_edu7_1_5=f_othed_75p +f_nvq_75p+ m_othed_75p+ m_nvq_75p
```

These are calculated based on 3 initial education groups – those with no education ( _noed_ variables), those with NQVs only ( _nvq_ variables) and those with other education ( _othed_ ) although in this model there is no distinction between those with NVQs and other education (since only the binary variable for no education is included in the model). There is also no distinction between males (variables names starting m_) and females (starting f_). There is distinction between each of the three age groups (45-64, 65-74 and 75 plus), since both binary variables for age categories are included in this model.

Of course, they could be calculated in any way convenient, provided the result is the anticipated proportion in each age/ sex/ educational group, pertinent to the model in hand. They can be named in any convenient way, providing each has the same name apart from having a different number at the end. This allows use of the reshape command in stata.

In practice, we do not want to find a synthetic estimate on just one population, but rather on many populations, for instance on each local authority separately. We have a data set containing information on the risk factors in all the different local authorities (LAs) and also other regions, with one line of data per region. The above variables give the proportions for each specified combination of age/ sex/ education categories. There are other variables giving the proportions by each additional risk factor separately (e.g. the proportion of non-smokers, current smokers and ex-smokers).



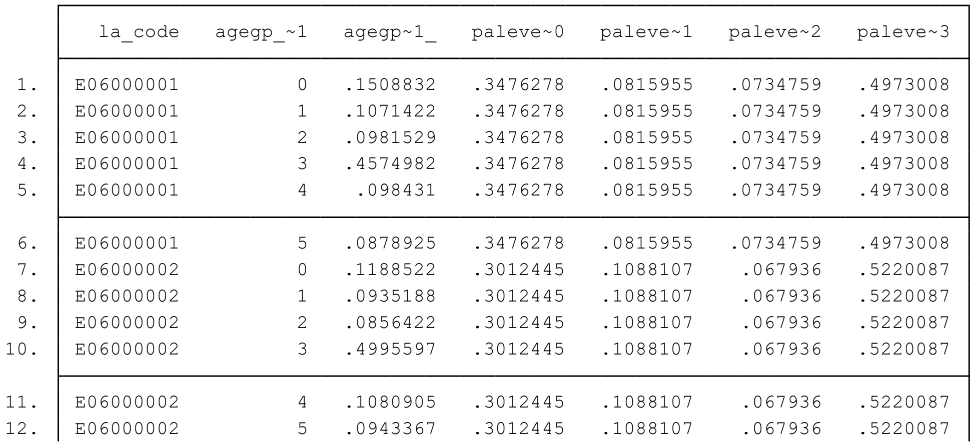
```
      la_code    agegp_~0    agegp_~1    agegp_~2    agegp_~3    agegp_~4    agegp_~5

 1.   E06000001   .1508832    .1071422    .0981529    .4574982     .098431    .0878925
 2.   E06000002   .1188522    .0935188    .0856422    .4995597    .1080905    .0943367
 3.   E06000003   .1720217    .1538003    .1252882    .4062224    .0801852    .0624822
 4.   E06000004   .0819255    .0819462    .0705499    .5388392    .1242868    .1024525
 5.   E06000005   .1694909    .1373918    .1226214     .425029    .0791328     .066334
```

A reshape long command on the set of 6 agegp_23_edu7_1_ variables (as defined above) is used as follows:

reshape long agegp_23_edu7_1_, i(ccg_code) j(agegp_23_edu7_1)

this gives 6 lines of data per region (since in this example there are 6 categories of age/ sex/ educational status, and 6 corresponding variables) from the starting place of one line of data per region. As well as a variable defining the categories (agegp_23_edu7_1 as named in the j() part of the above command), we now have a variable giving the proportion in each row of data (called variable agegp_23_edu7_1_ note that this name ends in _). These proportions were originally 6 variables on each line, and now we have 6 separate lines for each region. (If you look at the data listing above, the row of proportions turns into a column of 6 proportions, then the second row becomes a column of another 6 proportions below the first six, against the second LA code). The i() part of the command gives a unique identifier for each line of data.



```
      la_code   agegp_~1   agegp~1_   paleve~0   paleve~1   paleve~2   paleve~3

  1.  E06000001         0   .1508832   .3476278   .0815955   .0734759   .4973008
  2.  E06000001         1   .1071422   .3476278   .0815955   .0734759   .4973008
  3.  E06000001         2   .0981529   .3476278   .0815955   .0734759   .4973008
  4.  E06000001         3   .4574982   .3476278   .0815955   .0734759   .4973008
  5.  E06000001         4    .098431   .3476278   .0815955   .0734759   .4973008

  6.  E06000001         5   .0878925   .3476278   .0815955   .0734759   .4973008
  7.  E06000002         0   .1188522   .3012445   .1088107    .067936   .5220087
  8.  E06000002         1   .0935188   .3012445   .1088107    .067936   .5220087
  9.  E06000002         2   .0856422   .3012445   .1088107    .067936   .5220087
 10.  E06000002         3   .4995597   .3012445   .1088107    .067936   .5220087

 11.  E06000002         4   .1080905   .3012445   .1088107    .067936   .5220087
 12.  E06000002         5   .0943367   .3012445   .1088107    .067936   .5220087
```

For a risk factor, such as smoking status, where the number by age, sex and other risk factors is not known, the proportion of smokers and of ex-smokers in the population is applied to each age/ sex/ educational status group. Another such variable is physical activity (PA) level (palevel), which is in 4 categories, so has 3 corresponding binary variables, all of which are included in this predictive logistic regression model. This is the next one dealt with in practice.

Four relevant variables are created as follows for PA level, with the requirement that they all have the same name, except for the different numbers at the end, as follows:

gen palevelf_0=1-pa_low-pa_mod - pa_high
gen palevelf_1=pa_low
gen palevelf_2=pa_mod
gen palevelf_3=pa_high

(derived from pa_ variables for low, moderate and high physical activity levels).

With those 4 variables, a further reshape long command can be applied. Note that we already have 6 lines of data per region. This gives 4 lines of data (one for each PA level) from each line, which gives 6x4=24 lines of data per region now. The i() part of the command that gives the unique identifier now needs to include the age/sex/ education categories variable (agegp_23_edu7_1) as well as the region coding variable (ccg_code). The j() part tells stata to name the newly created categorical variable palevelf, which represents the different PA level categories. The palevelf_ variable (note _ at end of this name) gives the proportion within each PA level category (these add to one for each la_code/ agegp_23_edu7_1 combination, i.e. for each set of 4 lines – again the 4 values that are listed horizontally above are now listed vertically into this palevelf_ column).

reshape long palevelf_, i(ccg_code agegp_23_edu7_1) j(palevelf)

```
     |   la_code   agegp_~1   palevelf   agegp~1_   paleve~_   bmicat~0   bmicat~2   bmicat~3 |
     |-------------------------------------------------------------------------------------------|
  1. | E06000001          0          0   .1508832   .3476278    .315303   .3785597   .3061373 |
  2. | E06000001          0          1   .1508832   .0815955    .315303   .3785597   .3061373 |
  3. | E06000001          0          2   .1508832   .0734759    .315303   .3785597   .3061373 |
  4. | E06000001          0          3   .1508832   .4973008    .315303   .3785597   .3061373 |
  5. | E06000001          1          0   .1071422   .3476278    .315303   .3785597   .3061373 |
     |-------------------------------------------------------------------------------------------|
  6. | E06000001          1          1   .1071422   .0815955    .315303   .3785597   .3061373 |
  7. | E06000001          1          2   .1071422   .0734759    .315303   .3785597   .3061373 |
  8. | E06000001          1          3   .1071422   .4973008    .315303   .3785597   .3061373 |
  9. | E06000001          2          0   .0981529   .3476278    .315303   .3785597   .3061373 |
 10. | E06000001          2          1   .0981529   .0815955    .315303   .3785597   .3061373 |
     |-------------------------------------------------------------------------------------------|
 11. | E06000001          2          2   .0981529   .0734759    .315303   .3785597   .3061373 |
 12. | E06000001          2          3   .0981529   .4973008    .315303   .3785597   .3061373 |
 13. | E06000001          3          0   .4574982   .3476278    .315303   .3785597   .3061373 |
 14. | E06000001          3          1   .4574982   .0815955    .315303   .3785597   .3061373 |
 15. | E06000001          3          2   .4574982   .0734759    .315303   .3785597   .3061373 |
     |-------------------------------------------------------------------------------------------|
 16. | E06000001          3          3   .4574982   .4973008    .315303   .3785597   .3061373 |
 17. | E06000001          4          0    .098431   .3476278    .315303   .3785597   .3061373 |
 18. | E06000001          4          1    .098431   .0815955    .315303   .3785597   .3061373 |
 19. | E06000001          4          2    .098431   .0734759    .315303   .3785597   .3061373 |
 20. | E06000001          4          3    .098431   .4973008    .315303   .3785597   .3061373 |
     |-------------------------------------------------------------------------------------------|
 21. | E06000001          5          0   .0878925   .3476278    .315303   .3785597   .3061373 |
 22. | E06000001          5          1   .0878925   .0815955    .315303   .3785597   .3061373 |
 23. | E06000001          5          2   .0878925   .0734759    .315303   .3785597   .3061373 |
 24. | E06000001          5          3   .0878925   .4973008    .315303   .3785597   .3061373 |
```

Similarly for other risk factors. For this model, the other risk factors are BMI (obese, overweight and not overweight categories), smoking (where only ex-smoking is relevant, smokers and non-smokers are combined), gym membership and socio-economic status (with 3 relevant binary variables, giving 4 categories). Therefore for this model, there are
6 x 4 x 3 x 2 x 2 x 4 = 1152 different combinations of predictor variables. With 6 different “reshape long” commands in total, we end up with 1152 lines of data per region.

The weights for each region can be obtained by multiplying the relevant proportions together.
Weight = (proportion in specified age/ sex/ education category ) x (proportion by PA level) x (proportion by BMI group) x (proportion by smoking status) x (proportion by gym membership) x (proportion by relevant socio-economic status group).

gen xyz= agegp_23_edu7_1_ * palevelf_* bmicatf2_* smokef2_* hobby1_* ssec8_

These weights ("xyz") will sum to one for each region. It is a good idea to check that they do so in practice.

For practical purposes, so that we can use Stata efficiently, it is also necessary to create all the binary variables used in the original logistic regression modelling, and used to derive our preferred local predictive model with associated regression coefficients. The names and coding of these variables must be identical to those used in the original data set.

The most complex is recreating age, education and sex variables, since they are combined above for the purposes of the reshape command. For the model in our example, we do not need a sex variable, but we do need the following variables (check with the above commands which define them initially to make sure the appropriate codings are used – the tab2 command below also allows for some checking):

```
gen agegp2=agegp_23_edu7_1==1 | agegp_23_edu7_1==4
gen agegp3=agegp_23_edu7_1==2 | agegp_23_edu7_1==5
gen educ7=agegp_23_edu7_1==0 | agegp_23_edu7_1==1 | agegp_23_edu7_1==2
tab2 agegp_23_edu7_1 agegp2 agegp3 educ7, missing
```

For other variables, such as PA level, it is straight forward to create the required binary variables (the tab2 command again allows for some checking):

```
gen palevelf1=palevelf==1
gen palevelf2=palevelf==2
gen palevelf3=palevelf==3
tab2 palevelf palevelf1 palevelf2 palevelf3, missing
```

Note on creation of above variables: the right hand side are expressions, such as palevelf==1 – the variable is coded as =1 when this is true and =0 when this is false and including for missing values of palevelf (here we excluded any data with missing values earlier so this does not apply).

With our dataset set up in this way, we can now use stata's "predict" command to give us the predicted log odds. For this to work, the last regression that we have undertaken in stata must be the definitive predictive logistic regression equation for the chosen disease, which requires the dataset used to derive that to be in stata's memory at the time. When we use the "predict" command we need the dataset described above (after all the above described transformations), to be in stata's memory, since that gives the characteristics of the regions on which we want the synthetic estimates. It would also be possible to programme in the linear equation from the logistic regression manually, but I have not done that, since there is then more scope for errors.

The predict commands gives predicted log odds, and we then find the prevalence as follows: exp(log odds) / [1+ exp(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). Using stata, the weighted average can be found using the "collapse" command as follows, which results in one line of data per region (using a region identifier as the by() variable).

```
predict pred_values, xb
gen pred_OR=exp(pred_values)
gen pred_prev=pred_OR/(1+pred_OR)
gen wt_pred_prev=pred_prev*xyz
collapse (sum) wt_pred_prev xyz, by(ccg_code)
```

Thus the region is a data set with one line of data per region, with an estimate of prevalence against each region, based on the definitive logistic regression equation.

## 10.2 Calculating confidence intervals for prevalence estimates using bootstrap procedures

There is uncertainty in these synthetic estimates of prevalence based on the imprecision in the estimated coefficients from the logistic regression equations. A boot strap procedure can be used to construct confidence intervals on these synthetic estimates of prevalence, based on the imprecision in these logistic regression coefficients.

**Boot-strap procedures**

The philosophy underlying the boot-strap 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.

Boot strap samples are taken from our initial populations (the subsets of the ELSA population that has complete data on appropriate risk factors). The first person to be included in our new boot strap data set is chosen at random from our starting (ELSA) dataset, with each person being equally likely to be chosen. Then the second person to be included in this boot strap data set is chosen at random in the same way, again with each person being equally likely to be chosen. It is noteworthy that the second person to be chosen could be the same person as the person selected first (with probability 1/n where n=sample size, the same probability that any individual will be selected). We then select a 3rd person for our boot strap sample, then a 4th, 5th, 6th, and so on up to an nth person (where n is the size of our starting dataset). We are effectively selecting at random “with replacement”, which means that the same person can be selected twice, or indeed many times. (This is why I say that the population is effectively considered to have many copies of each person in it).

Therefore the boot strap data is the same size (same number of people in it) as the original dataset used to derive the logistic regression model. It is theoretically possible (though extremely unlikely) that a boot strap data set could be identical to that original dataset. However, it is far more likely that there will be differences, since some people will be included in the boot strap data set twice or more, and many are not included at all, although many would also be included just once.

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 region, 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 region). From these 1,000 synthetic estimates of prevalence of each region, 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).

The following commands describe how this is done in stata:

```
forvalues j=1/1000 {
use bootstrap, clear
                    (NB line above reads in original version of the data use used for logistic regrn eqn)
gen howmany=0
forvalues i=1/11516 {
local nn=floor(uniform()*11516)+1
quietly replace howmany=howmany+1 if nnn==`nn'
}
```

nn is a random variable, derived from a uniform random variable which takes values between 0 and 1, to give a random variable between 1 and the total sample size.
The variable "howmany" records how many times each individual has been selected (for the specific bootstrap sample)

```
drop if howmany==0
expand howmany
```

The above 2 lines drop any people that have not been selected in our sample, and then repeat lines (twice or more) of any that have been selected twice or more.

```
quietly logit kneecategory2 agegp2 agegp3 palevelf1 palevelf2 palevelf3 smokef2 bmicatf22 bmicatf23 educ7 ssec8_5 ssec8_6 ssec8_7 hobby1 [pweight=10*probwtks]
```

The above lines run the chosen logistic regression on this boot strap sample of data, to get new estimates of regression coefficients.

```
use temp0, clear
```

The above reads in data set of all possible combination of risk factors, for purposes of calculating confidence intervals

```
*** the saved data set has 1 extra variable, so storing the extra bootstrapped estimate
predict est`j', xb
save temp0, replace

}
```

To get boot strap confidence intervals on specific regions, we need to firstly find predicted prevalences from these predicted log odds (by taking exp(log odds)/ [1+exp(log odds)] for each bootstrap estimate.

```
forvalues j=0/1000 {
gen prev`j'=exp(est`j')/(1+exp(est`j'))
}
```

Remember we are working on a data set with one line of data for each combination of risk factors. We then need to merge this data set, with the data set which gives appropriate weighted for each combination of risk factors for each region (which has many lines of data per region, 1152 for severe knee OA model).

```
merge agegp2 agegp3 palevelf1 palevelf2 palevelf3 smokef2 bmicatf22 bmicatf23 educ7 ssec8_5 ssec8_6 ssec8_7 hobby1 using prevalences0
```

(This above commands lists each risk factor binary variable in the model as a variable that we are merging on).

For each boot strap sample, the synthetic prevalence estimate in any population is found by applying the same weights as above, according to the expected proportion of that population with any specified combination of risk factors (as follows – use of collapse command means that we conveniently end up with one line of data per patient).

```
forvalues j=0/1000 {
gen wt_prev`j'=xyz*prev`j'
}
collapse (sum) xyz wt_prev* (mean) c_pt45p c_tot_mf_ages, by(ccg_code)
```

This gives 1,000 different synthetic estimates of prevalence for each population, one for each of the boot strapped samples of data. The confidence interval is found on these by taking the 2.5th and 97.5th centiles. Alternatively, the confidence interval can be found by taking the mean and SD of these prevalence estimates, and taking the mean +/- 1.96 SDs. [In practice, for estimates of severe knee OA, both these sets of estimates agreed very well, suggesting that the distribution of these estimates approximates very closely to the normal distribution – therefore the second method, using mean +/- 1.96 SD, is a bit more precise]

```
egen meanr=rowmean(wt_prev1-wt_prev1000)
egen p2_5r=rowpctile(wt_prev1-wt_prev1000), p(2.5)
egen p97_5r=rowpctile(wt_prev1-wt_prev1000), p(97.5)
egen medianr=rowpctile(wt_prev1-wt_prev1000), p(50)
egen sdr=rowsd(wt_prev1-wt_prev1000)
```

Why is it not possible to put confidence intervals separately on each combination of risk factors? It is possible, but then averaging these would not agree to finding confidence intervals directly on appropriately weighted average prevalences of these, appropriate to specific populations. So that would not be a possible way forward with our objectives here.

Why is it necessary to divide the data into different groups by each combination of risk factors, rather simply taking account of the overall distribution of risk factors in the population? The weighted average of prevalences for a person with “average” risk factors is not the same as the weighted average prevalence, across all combinations of risk factors (appropriately weighted). The latter is what we want, and what we calculate directly.

This approach would work to find the appropriately weighted averaged log odds, since this a linear combination of risk factors. However, there is then a change of scale, taking the exponential to get the odds ratio, and then transforming again to get the prevalences.